Cape Peninsula
University of Technology

HYBRIDISED INDEXING FOR RESEARCH BASED INFORMATION RETRIEVAL

by

KYLE ANDREW FITZGERALD

VOLUME II

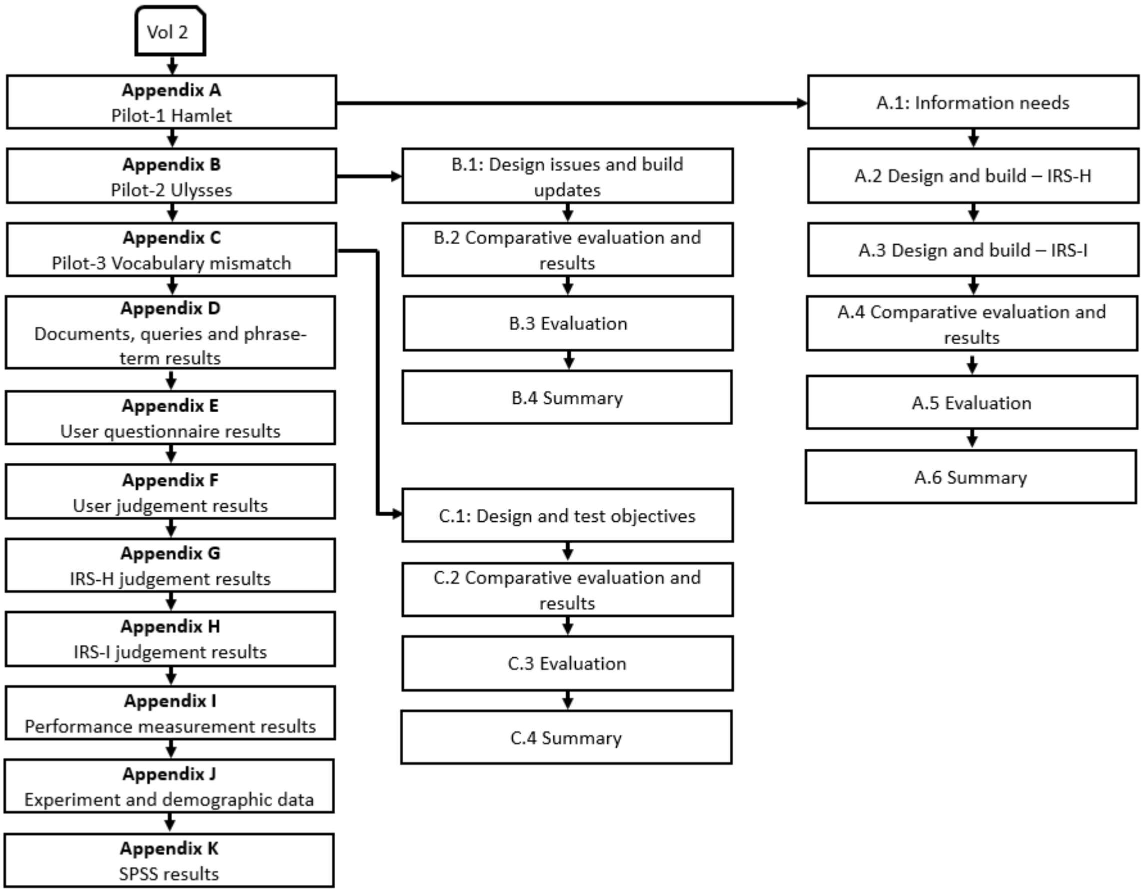
This volume contains the Appendices supporting the thesis, submitted separately

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

APPENDICES


Vol 2
Appendix A
Pilot-1 Hamlet
Appendix B
Pilot-2 Ulysses
Appendix C
Pilot-3 Vocabulary mismatch
Appendix D
Documents, queries and phrase-term results
Appendix E
User questionnaire results
Appendix F
User judgement results
Appendix G
IRS-H judgement results
Appendix H
IRS-I judgement results
Appendix I
Performance measurement results
Appendix J
Experiment and demographic data
Appendix K
SPSS results
B.1: Design issues and build updates
B.2 Comparative evaluation and results
B.3 Evaluation
B.4 Summary
C.1: Design and test objectives
C.2 Comparative evaluation and results
C.3 Evaluation
C.4 Summary
A.1: Information needs
A.2 Design and build – IRS-H
A.3 Design and build – IRS-I
A.4 Comparative evaluation and results
A.5 Evaluation
A.6 Summary

In Volume II of the thesis, there are eleven appendices. The first three appendices describe and present the design, build and test results for the three pilot tests based on design science research:

i) Pilot 1 uses the two pages from Hamlet Act 1 Scene 3 written by William Shakespeare. This book was specifically selected for its Elizabethan English / Early modern English and catchy phrases.

ii) Pilot 2 uses the 666 pages from the book Ulysses written by James Joyce. This book was selected for the author's use of unimaginable phrases, length of words, morphemes[1], and phonemes[2].

iii) Pilot 3 uses 20 journal articles, a few relevant to vocabulary mismatch and a few not. This topic was specifically selected to present how vocabulary mismatch itself has challenges in mismatching vocabulary and how it has multiple phrase-term synonyms.

The remaining eight appendices contain expansive data relevant to the results of this research (Volume I, Chapter Four and Chapter Five).

[1] A meaningful morphological (the study of words) unit of a language that cannot be further divided
[2] A distinct unit of sound in a specified language distinguishing one word from another

APPENDIX A: PILOT 1 HAMLET

A.1 Information needs

The information needs required to evaluate an IRS is a two-part process. Firstly, there are the information needs that the user defines, and secondly, once defined, the user must judge each of the documents to determine which are relevant to each of the information needs.

A.1.1 User information needs

For Pilot 1, four user information needs were compiled, covering one popular quotation *'to be or not to be that is the question'* from the script of Hamlet. The information needs listed in Table A.1 express a user's need to search for and retrieve those documents, within the document collection, relevant to each of the four information needs.

Table A.1: Pilot 1 – User information needs

In No	Information Need
in01	I want to find all documents relevant to any of these phrases
in02	I want to find all documents relevant to the phrase "to be"
in03	I want to find all documents relevant to the phrase "to be or not to be"
in04	I want to find all documents relevant to the phrase "that is the question"

A.1.2 User relevant document judgement

One activity of the user is to judge each document to determine whether a document is relevant to an information need. To accommodate this activity for this single document d_{01}, the text was manually searched for the phrases 'to be', 'to be or not to be', and 'that is the question' to ensure they actually existed within the document. As they did all exist, all four information needs were judged relevant by the user. The results are listed as a questionnaire in Table A.2.

Table A.2: Pilot 1 – User relevant document judgement

Document number - d01 Please indicate whether this document is relevant to any of the following information needs (please tick)		
In No	**Information Need**	**Relevant**
in01	I want to find all documents relevant to any of these phrases	√
in02	I want to find all documents relevant to the phrase "to be"	√
in03	I want to find all documents relevant to the phrase "to be or not to be"	√
in04	I want to find all documents relevant to the phrase "that is the question"	√

To accommodate this data, an information-need-by-document matrix was designed as a table within the evaluation system. Boolean data were converted to binary quantitative data. To indicate relevant, *'true'* was converted to 1 and to indicate non-relevant, *'false'* was converted to 0. The results for document number d_{01} judged relevant for information needs in_{01} through to in_{04} are listed in Table A.3. As all the cells within the matrix contain the value '1' all four information needs were judged relevant by the user for document d_{01}.

Table A.3: Pilot 1 – User information need-by-document matrix

doc	in01	in02	in03	in04
d01	1	1	1	1

A.2 Design and build – IRS-H

The design and build of the first IRS using one of the two indexing methods is now discussed. The IRS using the hybrid indexing method referred to as IRS-H is comprised of two processes: Process 1 to gather the information and Process 2 to trigger the search engine.

A.2.1 Process 1: Information gathering

The information gathering process illustrated in Figure A.1 consists of four stages: text acquisition, text transformation, the data store, and the hybrid token index.

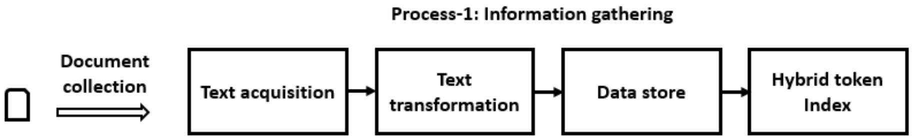


Figure A.1: Pilot 1 – IRS-H: The information gathering process

A.2.1.1 Text acquisition

Text acquisition is the first stage of designing the information gathering process. In Chapter Two, theories for text acquisition were discussed acquiring documents either manually, from the Web, social media or test collections, and thereafter converting these documents into text, and storing the information from the text to data stores. Social media and existing test collections were omitted as these were not within the scope of this study.

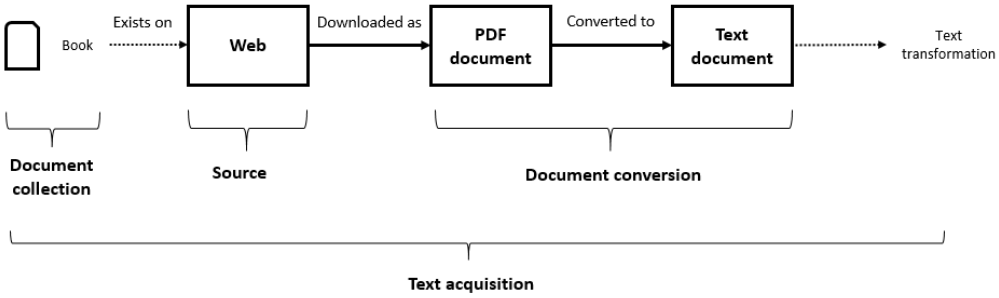


Figure A.2: Pilot 1 – IRS-H: Text acquisition

Thus, to pilot Process 1 the document collection was a single document, the book Hamlet written by William Shakespeare in the late 1590s. Only Act 3 Scene 1, consisting of two pages, was sourced from the Web and downloaded as an electronic pdf document to the research computer. Thereafter, the pdf document was converted to text format using a software application Adobe Reader and presented as a single text document ready for text transformation. The contents of this saved text file used as the input file, is presented in the text transformation build section in Figure A.2, illustrating the text acquisition for Pilot 1.

A.2.1.2 Text transformation

Text transformation is the second stage of designing the information gathering process. It is a process of transforming document text into word tokens. Numerous theoretical methods and techniques used in text transformation were discussed as options in Chapter Two. A few of these options not adopted for this IRS-H design were:

i) Classifiers – this is a method of identifying class related metadata for specific sections of documents, for example, the subject category, title, keywords, summary, and others (Croft et al., 2015). Classifiers played no role in this study, as the aim of this study was to use the complete text document in its whole form, therefore this theoretical method was not adopted.

ii) Stemming – this is a method of grouping words derived from a common stem (Croft et al., 2015). As this study is about efficiency, looking for words in their whole form, this theoretical method was not adopted.

iii) Stopping – this is a method for removing common or short frequently occurring words such as*: 'of', 'and', 'the'* from the text (Manning et al., 2008; Croft et al., 2015). As one aim of this study is to provide a method of returning documents judged exactly relevant, making use of phrase-terms that include stop words, then all these stop words were maintained, and therefore this theoretical method was not adopted.

iv) Suffix stripping – this is a method where similar terms are reduced to a single term through the removal of suffixes. The advantages suggested by Porter (1980) are increased IRSs performance and reduced database size and complexity. However, similar to stemming, all these words had to be maintained in this study in their full forms, and therefore this method was not adopted.

v) Web page links – this is a method of gathering information pertaining to links to Web pages that can be extracted and analysed using various algorithms. This forms the basis of the PageRank method used by Brin and Page (1998) in their search engine. As this study is not related to Web pages as documents, but only to document files downloaded from the Web and other sources in pdf format, this theoretical method was not adopted.

vi) The methods and techniques adopted for text transformation for this IRS-H design at this stage were:

vii) Levenshtein distance – although this is a measurement between two strings, it is the method performed that is of significance. The method encompasses edit operations for deletions, insertions and/or replacements of characters to transform one string into another string (Levenshtein, 1965). In this method the following special characters, and others, are replaced with the pipe delimiter:

" ", ",", ".", ";", ":", ")", "(", """", "%", "/", "\", "=", ">", "<", "“", "?", "+", "`", "[", "]", "{", "}", "&", "@", "*", "”", "!", "#", "…", "‡", "‡", "€", "†", "®", "§", "¥", "£", "¢", "»", "«", "±", "˜", "~", "$", "—", "-",”□”,””.

viii) De-hyphenation – although this forms part of the previous method, it warrants its own discussion. The use of the hyphen (‘-‘), a punctuation mark used to join words, complicates information retrieval because of the numerous ways words can be presented in the text. In the English language the main purpose of a hyphen is to split up vowels in words, join nouns as names and copyediting (the process of improving text formatting, style, and accuracy) (Manning et al*.,* 2008). To avoid additional phrase-terms in the queries and to compensate for hyphenation, hyphens are replaced by the pipe delimiter providing words in their pure form.

ix) Delimiting – the delimiter is a character that may be used to separate individual words. In the English language, the character traditionally used is a whitespace (a space between words). In this study the pipe[3] ‘|’ character (or vertical bar as it is sometimes referred to) used (Harris, 2002) to separate words and also as the replacement of special characters in the Levenshtein distance and de-hyphenation methods above.

x) Case folding – in order to treat all words equally and to match word tokens within phrase-terms, query terms and indices, all text is case folded to lower case (Manning et al*.,* 2008).

xi) Tokenisation – This is the method of acquiring the various chunks of text as individual words. After delimiting the text, these words are surrounded by pipe delimiters. This method extracts the words between the delimiters and provides these words, referred to as tokens, to the token index (Lang, 1995; Manning et al*.,* 2008).

Figure A.3 is an example of the text transformation design stage. The example below makes use of the first line of Hamlet for the only document (*d_{01}*) in the collection and is based on the ideas and concepts from Gray (1947), Levenshtein (1965), Lang (1995), Harris (2002), Tordai (2006), Manning et al. (2008) and Croft et al. (2015).

Reading from left-to-right: (1) the downloaded document from the Web is in pdf format; (2) the document is converted to text; (3) the first line of Hamlet Act 3 Scene 1 is used for exemplification; (4) de-hyphenation is applied to replace hyphens with the delimiters; (5) the pipe delimiter is used; (6) ordinal positions are noted and the ordinal positions are clearly indicated as they would be read in word sequential order by a user; (7) commas,

[3] The pipe delimiter is the preferred delimiter in information systems data retrieval processes where data is extracted from tables of a legacy information system and converted to files that contain text. A delimiter is used to separate the data in textual format emanating from the table columns. Software manufacturers traditionally use a comma as a delimiter in their comma separated values file (csv) formats but a comma often exists within data causing data misalignment in the textual output.

whitespaces are replaced with delimiters; (8) special characters are replaced with delimiters; (9) all text is case folded to lowercase and (10) text strings between delimiters are created as tokens.

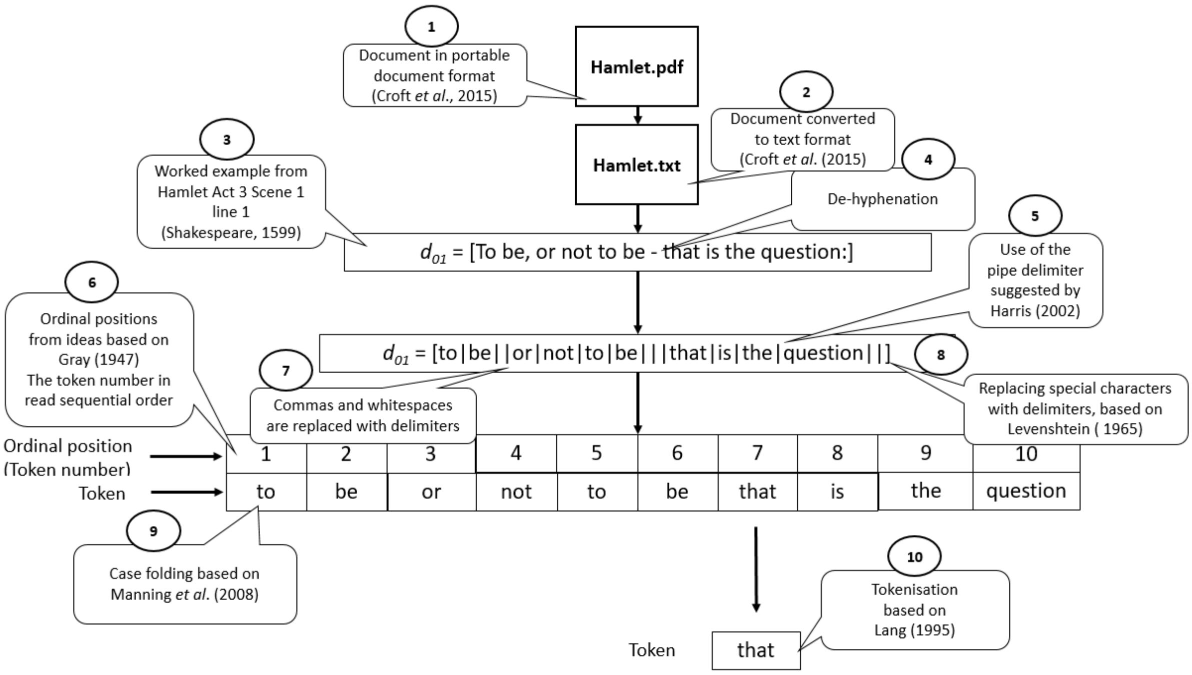

1
Document in portable document format (Croft *et al*., 2015)
Hamlet.pdf
2
Document converted to text format (Croft *et al.* (2015)
Hamlet.txt
3
Worked example from Hamlet Act 3 Scene 1 line 1 (Shakespeare, 1599)
4
De-hyphenation
d_{01} = [To be, or not to be - that is the question:]
5
Use of the pipe delimiter suggested by Harris (2002)
6
Ordinal positions from ideas based on Gray (1947) The token number in read sequential order
7
Commas and whitespaces are replaced with delimiters
d_{01} = [to|be||or|not|to|be|||that|is|the|question||]
8
Replacing special characters with delimiters, based on Levenshtein (1965)
Ordinal position (Token number)
Token
1 2 3 4 5 6 7 8 9 10
to be or not to be that is the question
9
Case folding based on Manning *et al.* (2008)
10
Tokenisation based on Lang (1995)
Token
that

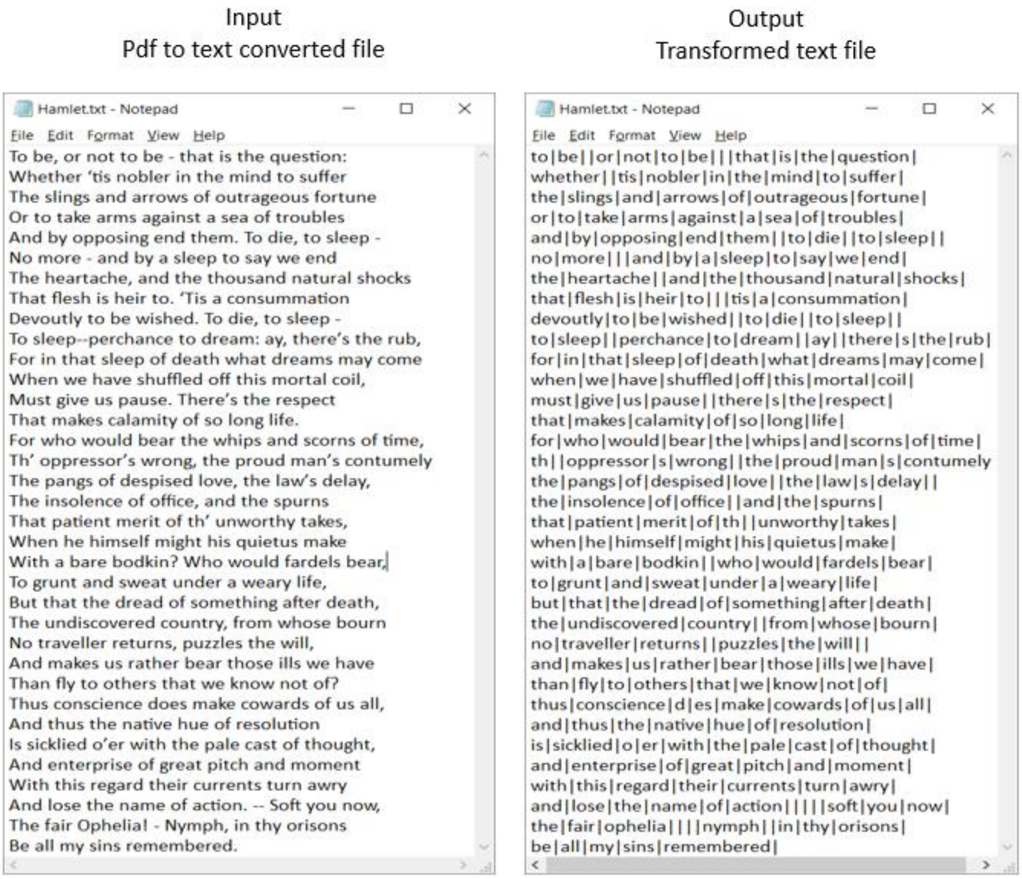
Figure A.3: Pilot 1 – IRS-H: Text transformation

To develop and build the text transformation stage a computer, a database, and a programming language were used. The development tools included a laptop with an I7 central processing unit, a solid-state disc drive, and 16 gigabytes of random access memory. The database software was Microsoft Access (MS Access) and the programming language was Visual Basic (VB). The basic functionality of transforming the text was to:

i) Read and store the details of the files in the directory. The metadata for these files were then placed in a table that forms part of the data store to be discussed later in this chapter.

ii) Run a script that reads the input text files, perform the transformation routines, and then write the data to an output text file. The input text file *'Hamlet.txt'* resides in the *'Txt'* folder, the output text file *'Hamlet.txt'* resides in the *'TxtOut'* folder and the original *'Hamlet.pdf'* file in the root folder.

Figure A.4 illustrates the input and output files for text transformation. On the left is the input file converted from the pdf file downloaded from the Web, and on the right is the output file transformed from the input file that made use of the adopted methods and techniques discussed earlier.



Input
Pdf to text converted file

```
To be, or not to be - that is the question:
Whether 'tis nobler in the mind to suffer
The slings and arrows of outrageous fortune
Or to take arms against a sea of troubles
And by opposing end them. To die, to sleep -
No more - and by a sleep to say we end
The heartache, and the thousand natural shocks
That flesh is heir to. 'Tis a consummation
Devoutly to be wished. To die, to sleep -
To sleep--perchance to dream: ay, there's the rub,
For in that sleep of death what dreams may come
When we have shuffled off this mortal coil,
Must give us pause. There's the respect
That makes calamity of so long life.
For who would bear the whips and scorns of time,
Th' oppressor's wrong, the proud man's contumely
The pangs of despised love, the law's delay,
The insolence of office, and the spurns
That patient merit of th' unworthy takes,
When he himself might his quietus make
With a bare bodkin? Who would fardels bear,
To grunt and sweat under a weary life,
But that the dread of something after death,
The undiscovered country, from whose bourn
No traveller returns, puzzles the will,
And makes us rather bear those ills we have
Than fly to others that we know not of?
Thus conscience does make cowards of us all,
And thus the native hue of resolution
Is sicklied o'er with the pale cast of thought,
And enterprise of great pitch and moment
With this regard their currents turn awry
And lose the name of action. -- Soft you now,
The fair Ophelia! - Nymph, in thy orisons
Be all my sins remembered.
```

Output
Transformed text file

```
to|be||or|not|to|be|||that|is|the|question|
whether||tis|nobler|in|the|mind|to|suffer|
the|slings|and|arrows|of|outrageous|fortune|
or|to|take|arms|against|a|sea|of|troubles|
and|by|opposing|end|them||to|die||to|sleep||
no|more|||and|by|a|sleep|to|say|we|end|
the|heartache||and|the|thousand|natural|shocks|
that|flesh|is|heir|to|||tis|a|consummation|
devoutly|to|be|wished||to|die||to|sleep||
to|sleep||perchance|to|dream||ay||there|s|the|rub|
for|in|that|sleep|of|death|what|dreams|may|come|
when|we|have|shuffled|off|this|mortal|coil|
must|give|us|pause||there|s|the|respect|
that|makes|calamity|of|so|long|life|
for|who|would|bear|the|whips|and|scorns|of|time|
th||oppressor|s|wrong||the|proud|man|s|contumely
the|pangs|of|despised|love||the|law|s|delay||
the|insolence|of|office||and|the|spurns|
that|patient|merit|of|th||unworthy|takes|
when|he|himself|might|his|quietus|make|
with|a|bare|bodkin||who|would|fardels|bear|
to|grunt|and|sweat|under|a|weary|life|
but|that|the|dread|of|something|after|death|
the|undiscovered|country||from|whose|bourn|
no|traveller|returns||puzzles|the|will||
and|makes|us|rather|bear|those|ills|we|have|
than|fly|to|others|that|we|know|not|of|
thus|conscience|d|es|make|cowards|of|us|all|
and|thus|the|native|hue|of|resolution|
is|sicklied|o|er|with|the|pale|cast|of|thought|
and|enterprise|of|great|pitch|and|moment|
with|this|regard|their|currents|turn|awry|
and|lose|the|name|of|action|||||soft|you|now|
the|fair|ophelia||||nymph||in|thy|orisons|
be|all|my|sins|remembered|
```

Figure A.4: Pilot 1 – IRS-H: Converted text file

A.2.1.3 Data store

Any IRS needs to perform mathematical computations using formulae to determine certain criteria. The data store must store these data emanating from the numerous computations that are performed, during information gathering and query processing, in addition to the document tokens and query terms discussed later in this chapter. The data store is the third stage in the design of the information gathering process. The document data store is a database that manages large volumes of documents and the structured data associated with them. Typically, a relational database contains the metadata from the documents collected (Croft et al., 2015). At this point in the process, the metadata pertains to data about the documents within the document collection and data about the word tokens acquired from the text. These data are then used to create the hybrid token index.

To develop and build the data store, the same development tools are used but additional algorithms and tables are created to populate the data store. In the build of the data store, at this stage of the information gathering process, one database table, *'File Names',* was created to store the following:

i) the unique record identity number,

ii) the document number,

iii) the file name of the document, and

iv) the name of the folder including its path.

For Process 1 using the hybrid indexing method, the following attributes were held within the database:

i) the original file name of each document in the collection,

ii) the converted to text file name of each document in the collection,

iii) the transformed text file name of each document in the collection,

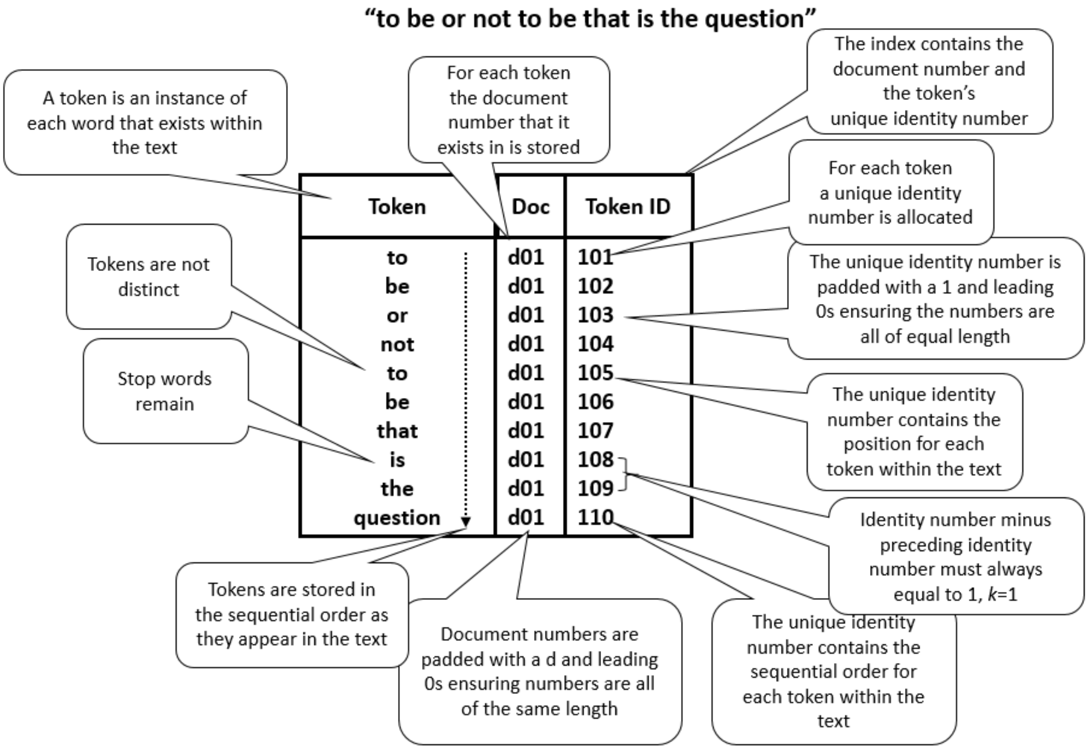
iv) the path in which each physical document resided,

v) a unique sequentially allocated document number for each document, and

vi) the hybrid token index.

A.2.1.4 Hybrid token index

The hybrid token index is the fourth stage of designing the information gathering process. To exemplify the design of the hybrid token index, the first line of text from Hamlet is utilised: *'to be or not to be that is the question'.* The phrase *'to be or not to be that is the question'* consists of ten words (*w*) denoted by w_1, w_2, w_3, w_4, w_5, w_6, w_7, w_8, w_9 and w_{10}. These words, defined as tokens, are acquired from the text through the process of text acquisition and text transformation. The hybrid token index is then created by using each of the ten tokens. Referring to Figure A.5, the design features of the hybrid token index are presented.



"to be or not to be that is the question"
A token is an instance of each word that exists within the text
For each token the document number that it exists in is stored
The index contains the document number and the token's unique identity number
For each token a unique identity number is allocated
Tokens are not distinct
The unique identity number is padded with a 1 and leading 0s ensuring the numbers are all of equal length
Stop words remain
The unique identity number contains the position for each token within the text
Identity number minus preceding identity number must always equal to 1, k=1
Tokens are stored in the sequential order as they appear in the text
Document numbers are padded with a d and leading 0s ensuring numbers are all of the same length
The unique identity number contains the sequential order for each token within the text

Token	Doc	Token ID
to	d01	101
be	d01	102
or	d01	103
not	d01	104
to	d01	105
be	d01	106
that	d01	107
is	d01	108
the	d01	109
question	d01	110

Figure A.5: Pilot 1 – IRS-H: Hybrid token index features

The major design features for the hybrid token index, using this example, are thus:

i) The text consists of the phrase: *'to be or not to be that is the question'.*

ii) The text contains ten tokens: *'to', 'be', 'or', 'not', 'to', 'be', 'that', 'is', 'the',* and *'question'*.

iii) The index contains three parts: the tokens, the document numbers and the unique token identity numbers, referred to in this research as the Token IDs.

iv) There are ten tokens constituting the dictionary and each token is an instance of each word, as it exists within the text.

v) The tokens are ‘non-distinct’ contrary to the tradition of inverted index design theory, with two tokens repeated because they appear twice in the text and the remaining six appearing once.

vi) Reading from top-to-bottom the tokens appear in word order as they appear in the text from which they were acquired.

vii) In the last two columns, each token has a document number followed by its unique Token ID. This is the key design feature of the hybrid token index.

viii) The document number points back, as it does in the traditional inverted index, to the document in which the token exits. The document number is first padded with the letter *d* and thereafter padded with leading 0s. In this example, the document number is *1* and is denoted by d_{01}. The length of padding can vary and the range of numbers selected must accommodate the number of documents in the collection.

ix) Similarly, within the postings list, each token points back to the text from where the token was acquired, and is allocated a unique Token ID. The Token ID is first padded with the number 1 and thereafter padded with leading 0s. In this example, for the first token *‘to’,* the Token ID is *101*. Again, the length of padding can vary and the range of numbers selected must accommodate the total number of all non-distinct tokens within the texts in the document collection.

x) Referring to the token *‘to’,* it is repeated as it appears twice in the text. For the first instance, the index refers to document d_{01} and Token ID *101* and in the second instance, to document d_{01} and Token ID *105*. By using these Token IDs, positioning and ordinality of words within the text are preserved, and the *k*-word proximity rule applied in this study where k = 1 always, is enforced (Gupta, 2008; Manning et al., 2008).

The functionality in populating the hybrid token index is as follows:

i) read the transformed text file names from the data store,

ii) read the lines of text from the file,

iii) for each line, extract the characters between each delimiter defined as a token (here the concept of a Token ID is introduced in this research; this is performed by allocating a sequential unique Token ID consisting of three numbers for each token extracted beginning at *‘101’*), and

iv) populate the hybrid token index with the token, the padded document number for the text file and the unique Token ID.

The hybrid token index therefore stores:

i) the token,

ii) the token's document number, and

iii) the token's unique Token ID.

This concludes the design and build of the information gathering process for IRS-H.

A.2.2 Process 2: Search engine

Designing the search engine process is the second of the two processes for IRS-H, as illustrated in Figure A.6. The process consists of four stages: the query design, the phrase-terms, the data store, and the hybrid query index.

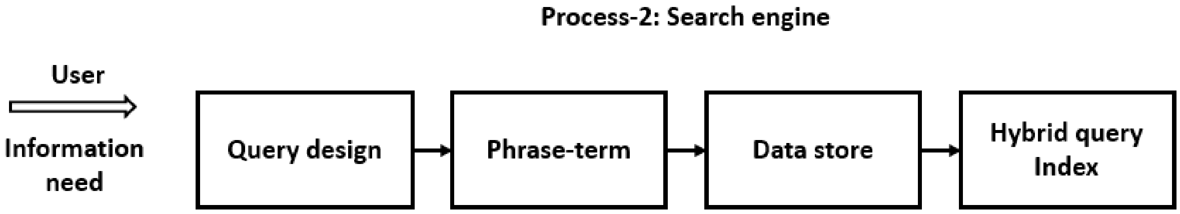


Figure A.6: Pilot 1 – IRS-H: The search engine process

A.2.2.1 Query design

Query design is the first stage of the search engine process. Query design makes use of multi-word phrase-terms in lieu of traditional single-word terms. The phrases are presented as strings surrounded by inverted commas and separated by the Boolean OR indicator. To satisfy the information need of the user, multiple queries can be applied either using a single phrase or expanded multiple phrases. An example of an expanded query is the query (q_{01}) below containing three phrase-terms:

$$q_{01} = [\text{ "to be" OR "to be or not to be" OR "that is the question" }]$$

And represented as words:

$$L_{01} = [\text{ "w1 w2" } OR \text{ "w1 w2 w3 w4 w5 w6" } OR \text{ "}w7\ w8\ w9\ w10\text{" }]$$

In addition, this example can be presented as three single phrase queries:

$$q_{02} = [\text{“to be”}]$$

$$q_{03} = [\text{"to be or not to be"}]$$

$$q_{04} = [\text{"that is the question"}]$$

Figure A.7 illustrates the design of the relationships between the four information needs and the four queries. Each information need has a one-to-one relationship with a query that expresses the phrase-terms used in the attempt to satisfy that information need.

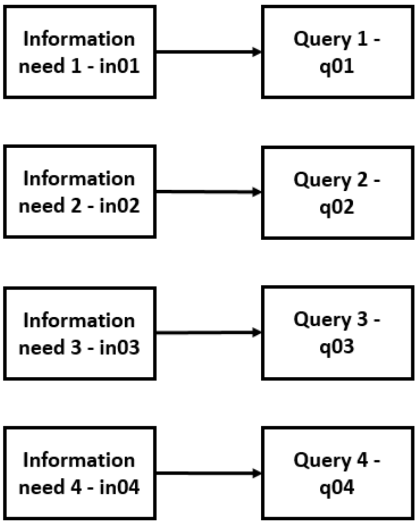

Information need 1 - in01
Query 1 - q01
Information need 2 - in02
Query 2 - q02
Information need 3 - in03
Query 3 - q03
Information need 4 - in04
Query 4 - q04

Figure A.7: Pilot 1 – Information needs and query relationships

The tricky part is to simulate search engine functionality. To develop and build a simulation of a query in IRS-H the same development tools are used as before, plus one additional table is created to present the query to the search engine. The basic functionality of this table (Table A.4) was to store the sequence number (Seq), the information need number (In no), the phrase-term's unique identity number (pt), and the phrase-term itself (Phrase-term).

Table A.4: Pilot 1 – IRS-H: building the search query

Seq	In No	pt	Phrase-term
1	in01	pt01	to be
2	in01	pt02	to be or not to be
3	in01	pt03	that is the question
1	in02	pt01	to be
1	in03	pt02	to be or not to be
1	in04	pt03	that is the question

The design of the search query is illustrated in Figure A.8.

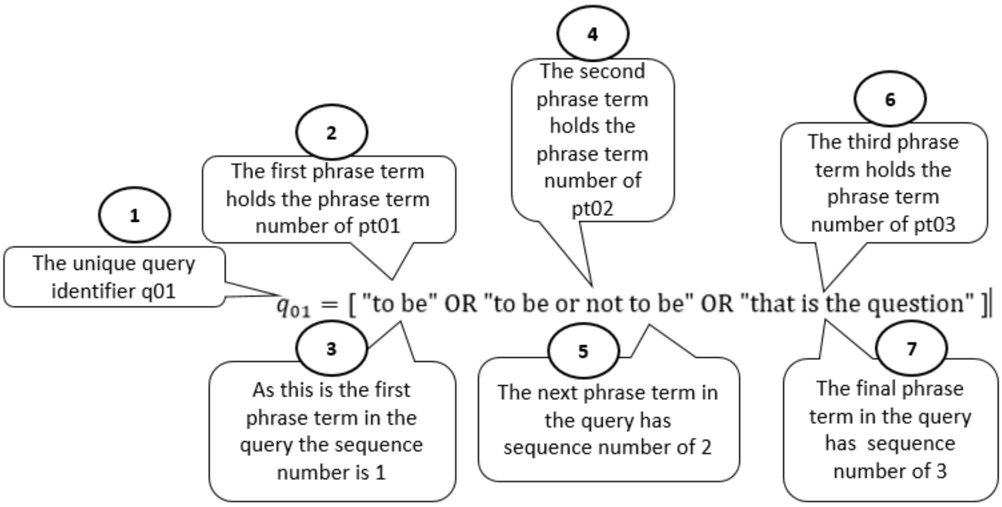

4
The second phrase term holds the phrase term number of pt02
2
The first phrase term holds the phrase term number of pt01
6
The third phrase term holds the phrase term number of pt03
1
The unique query identifier q01
q_{01} = ["to be" OR "to be or not to be" OR "that is the question"]
3
As this is the first phrase term in the query the sequence number is 1
5
The next phrase term in the query has sequence number of 2
7
The final phrase term in the query has sequence number of 3

Figure A.8: Pilot 1 – IRS-H: Query explanation

Referring to the search query and reading from left-to-right (1) the unique query identifier is *q01*; (2) the first phrase-term *'to be'* is allocated a phrase-term number of *pt01*; (3) as *'to be'* is the first phrase-term within the query it is allocated a sequence number of 1; (4) the second phrase-term *'to be or not to be'* is allocated a phrase-term number of *pt02*; (5) as *'to be or not to be'* is the second phrase-term within the query it is allocated a sequence number of 2; (6) the third phrase-term *'that is the question'* is allocated a phrase-term number of *pt03*; (7) as *'that is the question'* is the third phrase-term within the query it is allocated a sequence number of 3.

A.2.2.2 Phrase-terms

Phrase-term design is the second stage of the search engine process. The phrase-terms used in the queries are represented as follows:

$$pt_{01} = \text{"to be"}$$

$$pt_{02} = \text{"to be or not to be"}$$

$$pt_{03} = \text{"that is the question"}$$

The first phrase-term consists of two words, the second, six words and the last, four words, expressed as:

$$pt_{01} = "w1\ w2"$$

$$pt_{02} = "w1\ w2\ w3\ w4\ w5\ w6"$$

$$pt_{03} = "w7\ w8\ w9\ w10"$$

Each of these words must ultimately be matched to tokens acquired from the transformed text file during the information gathering process. Each of the phrase-terms can be used individually and/or simultaneously within numerous queries as described above.

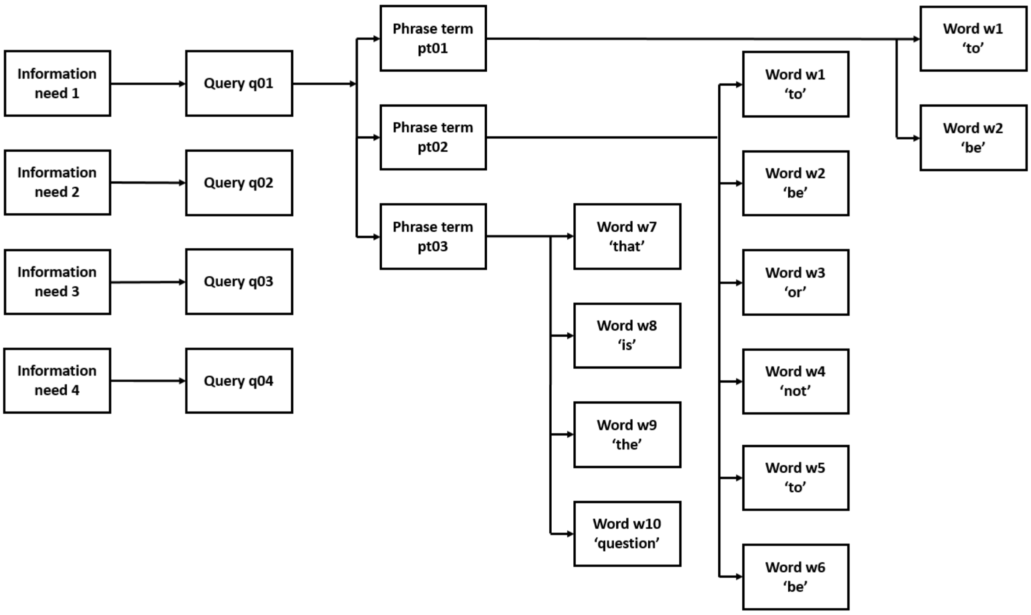


Figure A.9: Pilot 1 – IRS-H: Information needs, queries and phrase-term relationships

Figure A.9 illustrates the design of the relationships between the four information needs, the four queries, the phrase-terms, and the words that exist within the phrase-terms. For this pilot, each query may have one or more phrase-terms and each phrase-term may have one or more words. In the example above in Figure A.9, query q_{01} contains the three phrase-terms pt_{01}, pt_{02} and pt_{03} where pt_{01} consists of two words w_1 *and* w_2, *'to'* and *'be',* with corresponding unique Token IDs of *101* and *102* respectively. Similarly, pt_{02} consists of the six words w_1, w_2, w_3, w_4, w_5 *and* w_6, *'to', 'be', 'or', 'not', 'to,* and *'be'* with corresponding unique identity numbers of *101* through to *106* and pt_{03} consists of the four words w_7, w_8, w_9, *and* w_{10}, *'that', 'is', 'the'* and *'question'* with corresponding unique identity numbers of *107* through to *110*. The basic functionality during the phrase-term stage was to read the phrase-terms within the phrase-term table, and to read and determine which phrase-terms exist within the queries in the query design table. To evaluate the document collection using the hybrid indexing method, the three multi-word phrase-terms were used to describe the four information needs where each information need had one or more phrase-terms allocated to it. The final phrase-terms are presented in Table A.5, the information need to phrase-term relationships in Table A.6, and the information need to query relationships in Table A.7.

Table A.5: Pilot 1 – IRS-H: Phrase-terms

pt	Phrase-term
pt01	to be
pt02	to be or not to be
pt03	that is the question

Table A.6: Pilot 1 – IRS-H: Phrase-term / information need relationships

In No	Information Need	pt	Phrase-term
in01	I want to find all documents relevant to any of these phrases	pt01	to be
in01	I want to find all documents relevant to any of these phrases	pt02	to be or not to be
in01	I want to find all documents relevant to any of these phrases	pt03	that is the question
in02	I want to find all documents relevant to the phrase "to be"	pt01	to be
in03	I want to find all documents relevant to the phrase "to be or not to be"	pt02	to be or not to be
in04	I want to find all documents relevant to the phrase "that is the question"	pt03	that is the question

Table A.7: Pilot 1 – IRS-H: Query / information need relationships

In No	q	Query
in01	q01	"to be" OR "to be or not to be" OR "that is the question"
in02	q02	"to be"
in03	q03	"to be or not to be"
in04	q04	"that is the question"

A.2.2.3 Data store

The data store is the third stage of the search engine process. The query data store is the same database that manages the large volumes of documents and the structured data associated with them, but includes the management of the queries and the data associated with them. It effectively contains the metadata from the queries processed by the user (Croft et al., 2015). The metadata pertains to data about the queries expressed as information needs of the user and the phrase-terms structuring the queries. These data are then stored to enable the hybrid query index to be created and these are:

i) the unique number allocated to the query,

ii) each phrase-term acquired from the query, and

iii) the relationship of each phrase-term to each document.

In the design of the information gathering process two database tables were created, File Names to store the unique record identity number, the document number, the file name of the document and the name of the folder including the path, and the other, the Hybrid Token Index to store the token, the token's document number and the token's ID. Now, at this stage of the search engine process the data store contains those preceding two tables together with a phrase-terms table to store the unique record identity number, the phrase-term number and the phrase-term itself; and query search to store the unique record identity number, the query number, the phrase-term's unique identity number and the sequence order the phrase-term appears in the query. Next is the hybrid query index, which is added to the data store. To evaluate generated data a number of statistical tables are created. For Process 2 using the hybrid indexing method, the attributes held within the database are:

i) the information needs,

ii) the queries and their relationships to the information needs,

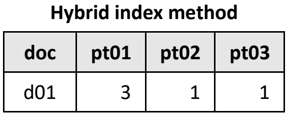
iii) the phrase-terms and their relationships to the queries,

iv) the phrase-term-by-document matrix containing phrase-term frequencies, and

v) the hybrid query index.

One additional table held within the data store is the phrase-term-by-document matrix. This table is populated with the number of times each phrase-term within a query occurs in each document. Table A.8 illustrates the phrase-term-by-document matrix for Pilot 1 using the hybrid indexing method. The rows represent the documents, the columns represent the phrase-terms, and the values in each cell represent the phrase-term frequency *ptf*. Phrase-term pt_{01} therefore occurs three times in document d_{01} and is represented as $ptf_{pt01,\ d01} = 3$, both phrase-terms pt_{02} and pt_{03} occur once in document d_{01} represented by $ptf_{pt02,\ d01} = 1$ and $ptf_{pt03,\ d01} = 1$ respectively. No term weighting or inverted term weighting is applied to enhance or suppress phrase-term frequency. Note that the phrase-term-by-document matrix forms the basis of the information-need-by-document matrix discussed in the performance measurements section later in this chapter.

Table A.8: Pilot 1 – IRS-H: Phrase-term-by-document matrix



Hybrid index method

doc	pt01	pt02	pt03
d01	3	1	1

A.2.2.4 Hybrid query index

The hybrid query index is the fourth stage of designing the search engine process. The data required to be stored to enable the index to be created are thus:

i) the unique number allocated to the query,
ii) each phrase-term acquired from the query,
iii) the relationship of each phrase-term to each document,
iv) the begin Token ID for the first word appearing in the phrase-term,
v) the end Token ID for the last word appearing in the phrase-term, and
vi) the relationship of the phrase-term to the document and to its begin and end Token IDs.

The first line of text consisting of ten words (L_{01}) and the three phrase-terms pt_{01}, pt_{02} and pt_{03} that exist within the query q_{01} are used to explain the hybrid query design.

$$L_{01} = [\ \text{"to be to be or not to be that is the question"}\ ]$$

$$pt_{01} = \text{"to be"},\quad pt_{02} = \text{"to be or not to be"},\quad pt_{03} = \text{"that is the question"}$$

$$q_{01} = [\ \text{"to be" OR "to be or not to be" OR "that is the question"}\ ]$$

The hybrid query index is created by using each of the phrase-terms within the query. The search engine firstly searches for the first phrase-term and checks to see whether it exists within the text (from the first page until the last page) by using the token index. A match between the phrase-term and the token index is defined when all the words within the phrase-term exist in the hybrid token index, the Token IDs for the words matched to the tokens are performed in sequential order and the values for *k* from *k*-word proximity indicator theory are always equal to one, therefore *k* = 1 at all times. The major features for the hybrid query index using the first ten words from the text as examples are thus:

i) The text consists of the phrase: *'to be or not to be that is the question'.*
ii) From the text used in this example this phrase consists of ten words (*w*) denoted by $w_1,\ w_2,\ w_3,\ w_4,\ w_5,\ w_6,\ w_7,\ w_8,\ w_9\ and\ w_{10}$.
iii) The query contains three phrase-terms: *'to be'', 'to be or not to be'* and *'that is the question'* each surrounded by inverted commas and separated by the Boolean operator OR.
iv) The index contains three parts: the phrase-terms, the document number, and begin and end Token IDs.
v) There are three phrase-terms that constitute the vocabulary with one repeated as it appears twice in this text with the remaining two appearing once.
vi) Referring to the phrase-term *'to be'* that uses words w_1 and w_2 it is repeated, as it appears twice in the text. For the first instance, the index refers to document d_{01} and Token IDs *101* and *102* and in the second instance to document d_{01} and identity numbers *105* and *106*. By using these Token IDs, positions of words within the text

and word order are preserved, and *k*-word proximity rule applied, is enforced. In this study where *k* = 1 always, the $/k$ operator is used to determine the occurrences of word $w1$ within k words of $w2$ and therefore, if it is required that $w1$ is to be adjacent to $w2$ as in this case, and if $w1$ is in position p then $w2$ must be in position $p + 1$ (Gupta, 2008; Manning et al*.,* 2008). In this example for the second instance of the first phrase $w1 = 'to'$ and $w2 =' be'$ with the positions of *105* and *106* respectively, if $p = 105$ then $p + 1 = 106$. According to the theory for two adjacent tokens by Clarke et al. (2000), for this hybrid phrase index, cover length will always be equal to two. This holds true in this case where w = 106 - 105 + 1 = 2.

vii) By using the Token IDs in the index the begin position and end position for each phrase-term can be derived. For the second instance of the phrase-term *'to be'* the begin position of the phrase, as it exists within the text within the document, is *105* and the end position is *106*. Similarly for the second phrase-term *'to be or not to be'* the begin position is *101* and the end position is *106*.

The basic functionality in populating the hybrid query index can now be presented as:

i) read all the tokens and their corresponding data in the hybrid token index from the data store and store in-memory in the sequential order they were read from the original transformation text file,

ii) read the phrase-terms for each query in the query search table,

iii) for each phrase-term extract each word within the phrase-term preserving ordinality and proximity,

iv) for each document in the collection attempt to match words the within the phrase-term with the tokens in the hybrid token index: if a match occurs then store the word and its Token ID, if not then ignore, and

v) populate the hybrid query index with the phrase-term, the document number, and begin and end Token IDs.

The format of the hybrid query index for the first ten words in the text is now presented in Table A.9. The first column represents the phrase-term number, the second the phrase-term, the third the document number, the fourth the Begin Token ID and the fifth the End Token ID.

Table A.9: Pilot 1 – IRS-H: hybrid query index

pt	Phrase-term	doc	Begin Token ID	End Token ID
pt01	to be	d01	101	102
pt02	to be or not to be	d01	101	106
pt01	to be	d01	105	106
pt03	that is the question	d01	107	110

This concludes the design and build of the search engine process for IRS-H.

A.3 Design and build – IRS-I

The design and build of the IRS using the inverted indexing method referred to as IRS-I, is comprised of the same two processes as the hybrid indexing method but with differing theoretical design concepts in the two main processes: Process 1 for the information gathering and Process 2 for the search engine. The design and build of these processes are now discussed in detail.

A.3.1 Information gathering

The information gathering stage mirrors that of the hybrid indexing method except for the final stage where the inverted token index replaces the hybrid token index illustrated in Figure A.10.

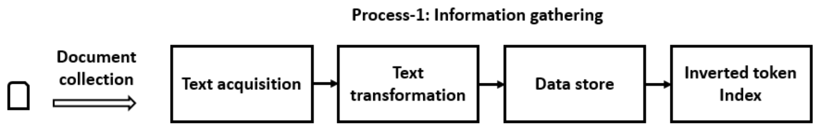
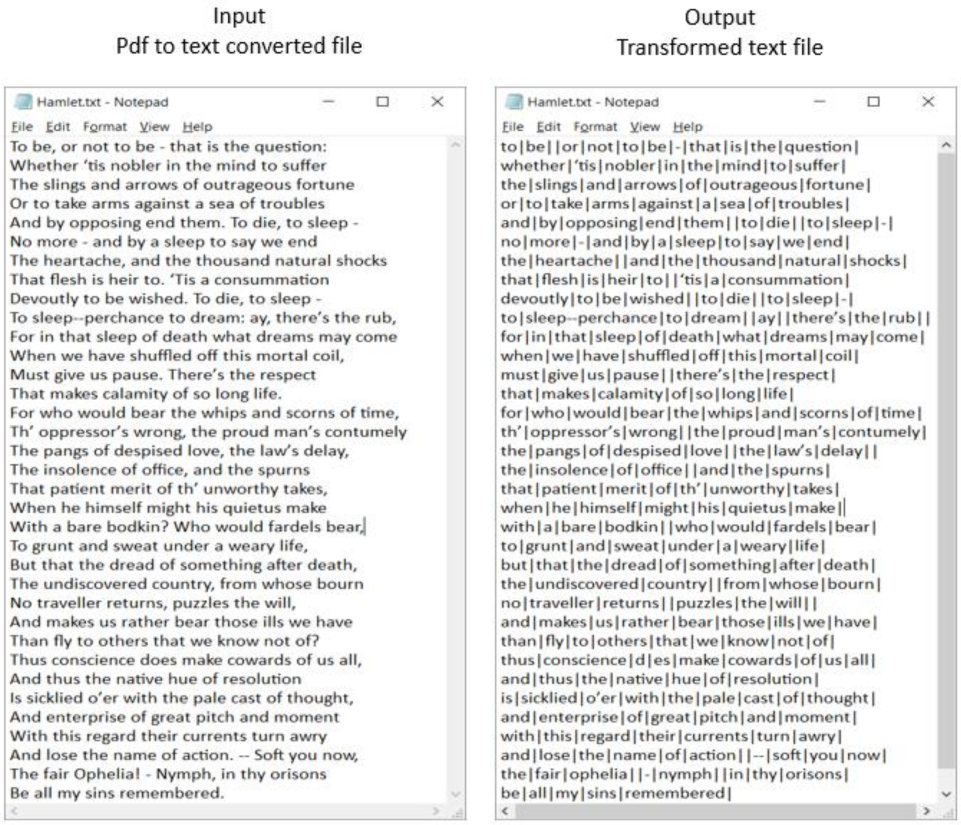


Figure A.10: Pilot 1 – IRS-I: The information gathering process

A.3.1.1 Text acquisition

The text acquisition stage mirrors that of the hybrid indexing method and therefore remains the same.

A.3.1.2 Text transformation


Input
Pdf to text converted file
Output
Transformed text file
Hamlet.txt - Notepad
File Edit Format View Help
Hamlet.txt - Notepad
File Edit Format View Help

```
To be, or not to be - that is the question:
Whether 'tis nobler in the mind to suffer
The slings and arrows of outrageous fortune
Or to take arms against a sea of troubles
And by opposing end them. To die, to sleep -
No more - and by a sleep to say we end
The heartache, and the thousand natural shocks
That flesh is heir to. 'Tis a consummation
Devoutly to be wished. To die, to sleep -
To sleep--perchance to dream: ay, there's the rub,
For in that sleep of death what dreams may come
When we have shuffled off this mortal coil,
Must give us pause. There's the respect
That makes calamity of so long life.
For who would bear the whips and scorns of time,
Th' oppressor's wrong, the proud man's contumely
The pangs of despised love, the law's delay,
The insolence of office, and the spurns
That patient merit of th' unworthy takes,
When he himself might his quietus make
With a bare bodkin? Who would fardels bear,
To grunt and sweat under a weary life,
But that the dread of something after death,
The undiscovered country, from whose bourn
No traveller returns, puzzles the will,
And makes us rather bear those ills we have
Than fly to others that we know not of?
Thus conscience does make cowards of us all,
And thus the native hue of resolution
Is sicklied o'er with the pale cast of thought,
And enterprise of great pitch and moment
With this regard their currents turn awry
And lose the name of action. -- Soft you now,
The fair Ophelia! - Nymph, in thy orisons
Be all my sins remembered.
```

```
to|be||or|not|to|be|-|that|is|the|question|
whether|'tis|nobler|in|the|mind|to|suffer|
the|slings|and|arrows|of|outrageous|fortune|
or|to|take|arms|against|a|sea|of|troubles|
and|by|opposing|end|them||to|die||to|sleep|-|
no|more|-|and|by|a|sleep|to|say|we|end|
the|heartache||and|the|thousand|natural|shocks|
that|flesh|is|heir|to||'tis|a|consummation|
devoutly|to|be|wished||to|die||to|sleep|-|
to|sleep--perchance|to|dream||ay||there's|the|rub||
for|in|that|sleep|of|death|what|dreams|may|come|
when|we|have|shuffled|off|this|mortal|coil|
must|give|us|pause||there's|the|respect|
that|makes|calamity|of|so|long|life|
for|who|would|bear|the|whips|and|scorns|of|time|
th'|oppressor's|wrong||the|proud|man's|contumely|
the|pangs|of|despised|love||the|law's|delay||
the|insolence|of|office||and|the|spurns|
that|patient|merit|of|th'|unworthy|takes|
when|he|himself|might|his|quietus|make||
with|a|bare|bodkin||who|would|fardels|bear|
to|grunt|and|sweat|under|a|weary|life|
but|that|the|dread|of|something|after|death|
the|undiscovered|country||from|whose|bourn|
no|traveller|returns||puzzles|the|will||
and|makes|us|rather|bear|those|ills|we|have|
than|fly|to|others|that|we|know|not|of|
thus|conscience|d|es|make|cowards|of|us|all|
and|thus|the|native|hue|of|resolution|
is|sicklied|o'er|with|the|pale|cast|of|thought|
and|enterprise|of|great|pitch|and|moment|
with|this|regard|their|currents|turn|awry|
and|lose|the|name|of|action||--|soft|you|now|
the|fair|ophelia||-|nymph||in|thy|orisons|
be|all|my|sins|remembered|
```

Figure A.11: Pilot 1 – IRS-I: Converted text file

Text transformation is similar to the hybrid indexing method. The differentiating factors are that hyphenation and apostrophes are retained within the text. Note the token ’tis with the preceding apostrophe and the single and double hyphens that become evident in the converted text file in Figure A.11.

A.3.1.3 Data store

For Process 1, using the inverted indexing method, the following attributes are held in the database:

i) the original file name of each document in the collection,

ii) the converted to text file name of each document in the collection,

iii) the transformed text file name of each document in the collection,

iv) the path in which each physical document resides,

v) a unique sequentially allocated document number for each document (these document numbers mirror those used in the hybrid indexing method allowing comparisons between the two methods to be performed), and

vi) the inverted token index.

A.3.1.4 Inverted token index

The traditional inverted token index uses a distinct list of words within the text and each distinct word is associated with the document numbers of the documents in which they exist, in the postings list, whereas the hybrid token index lists every occurrence of a word together with its set of Token IDs and the document number.

This concludes the design and build of the information gathering process for IRS-I.

A.3.2 Process 2: Search engine

The search engine stage is similar to the hybrid indexing method except for stages two and four. Stage two refers to terms rather than phrase-terms and stage four refers to the inverted query index rather than the hybrid query index, as illustrated in Figure A.12.

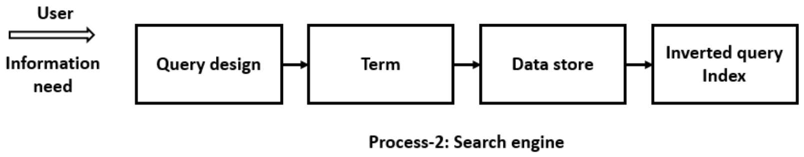


Figure A.12: Pilot 1 – IRS-I: The search engine process

A.3.2.1 Query design

The query design for the inverted index method makes use of single-word terms. The terms are presented as strings separated by the Boolean OR indicator. The inverted index method has no control over word proximity or ordinality, so the distinct set of words are presented using the bag of words concept where BoW = [be is not or question that the to] rather than *‘to be or not that is the question’.* To satisfy the information need of the user, multiple queries can be applied either using a single term or using multiple terms (to expand the query). Therefore, the expanded query (*q₀₁*) below contains eight terms:

$$q_{01} = [\text{ be OR is OR not OR or OR question OR that OR the OR to}]$$

In addition, this example can be presented as three single queries:

$$q_{02} = [\text{to OR be}]$$

$$q_{03} = [\text{to OR be OR or OR not}]$$

$$q_{04} = [\text{that OR is OR the OR question}]$$

Figure A.13 illustrates the design of the relationships between the four information needs and the four queries discussed earlier. Each information need has a one-to-one relationship with a query that expresses the terms used, in an attempt to satisfy the information need.

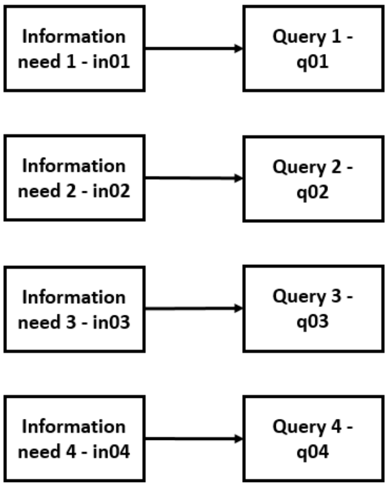

Information need 1 - in01
Query 1 - q01
Information need 2 - in02
Query 2 - q02
Information need 3 - in03
Query 3 - q03
Information need 4 - in04
Query 4 - q04

Figure A.13: Pilot 1 – Information needs and query relationships

A.3.2.2 Terms

Term design is the second stage of the search engine process. The terms used in the queries are represented as follows:

$$t_{01} = "be"$$

$$t_{02} = "is"$$

$$t_{03} = "not"$$

$$t_{04} = "or"$$

$$t_{05} = "question"$$

$$t_{06} = "that"$$

$$t_{07} = "the"$$

$$t_{08} = "to"$$

All terms consist of single words and can be expressed as:

$$t_{01} = "w1"$$

$$t_{02} = "w2"$$

$$t_{03} = "w3"$$

$$t_{04} = "w4"$$

$$t_{05} = "w5"$$

$$t_{06} = "w6"$$

$$t_{07} = "w7"$$

$$t_{08} = "w8"$$

Each of these words must ultimately be matched with the tokens acquired from the transformed text file during the information gathering process. Each of the terms can be used individually and/or simultaneously within numerous queries as indicated above. To evaluate the document collection using the inverted indexing method, eight single-word terms were used to describe the four information needs. Each information need had more than one term allocated to it. Each term is listed in Table A.10 and its associated information need in Table A.11.

Table A.10: Pilot 1 – IRS-I: The set of terms

t	Term
t01	be
t02	is
t03	not
t04	or
t05	question
t06	that
t07	the
t08	to

Table A.11: Pilot 1 – IRS-I: Term per information need

In No	Information Need	t	Word
in01	I want to find all documents relevant to any of these phrases	t01	be
in01	I want to find all documents relevant to any of these phrases	t02	is
in01	I want to find all documents relevant to any of these phrases	t03	not
in01	I want to find all documents relevant to any of these phrases	t04	or
in01	I want to find all documents relevant to any of these phrases	t05	question
in01	I want to find all documents relevant to any of these phrases	t06	that
in01	I want to find all documents relevant to any of these phrases	t07	the
in01	I want to find all documents relevant to any of these phrases	t08	to
in02	I want to find all documents relevant to the phrase "to be"	t01	be
in02	I want to find all documents relevant to the phrase "to be"	t08	to
in03	I want to find all documents relevant to the phrase "to be or not to be"	t01	be
in03	I want to find all documents relevant to the phrase "to be or not to be"	t03	not

In No	Information Need	t	Word
in03	I want to find all documents relevant to the phrase "to be or not to be"	t04	or
in03	I want to find all documents relevant to the phrase "to be or not to be"	t08	to
in04	I want to find all documents relevant to the phrase "that is the question"	t02	is
in04	I want to find all documents relevant to the phrase "that is the question"	t05	question
in04	I want to find all documents relevant to the phrase "that is the question"	t06	that
in04	I want to find all documents relevant to the phrase "that is the question"	t07	the

For the inverted index, each information need of the user again creates queries using the terms from the bag of words concept, and these queries are structured, where each distinct term is separated by the logical OR operator, as listed in Table A.12. Note that query q_{01} is associated with information need in_{01}, q_{02} with in_{02} and so on.

Table A.12: Pilot 1 – IRS-I: Expanded query per information need

In No	q	Query
in01	q01	to OR be Or or OR not OR that OR is OR the OR question
in02	q02	to OR be
in03	q03	to OR be Or or OR not
in04	q04	that OR is OR the OR question

A.3.2.3 Data store

For Process 2, using the inverted indexing method, the following attributes are held in the data store:

i) the information needs (the identical set of information needs as used in the hybrid indexing method, thus enabling comparisons between the two methods),

ii) the queries and their relationships to the information needs,

iii) the terms and their relationships to the queries,

iv) the term-by-document matrix containing term frequencies, and

v) the inverted query index.

The term-by-document matrix is a method for capturing the number of times each term in a query occurs in each document. Table A.13 illustrates the term-by-document matrix for Pilot 1 using the inverted indexing method. The rows represent the documents, the columns represent the terms, and the values in each cell represent the term frequency. Term t_{01} therefore occurs four times in document d_{01} and is represented as $tf_{t01,\ d01}$ = 4, term t_{02} occurs three times in document d_{01} represented by $tf_{t02,\ d01}$ = 3, etc. No term weighting or inverted term weighting is applied to enhance or supress *tf*. Note that the term-by-document matrix forms the basis of the information-need-by-document matrix discussed in the performance measurements section later in this chapter.

Table A.13: Pilot 1 – IRS-I: Term-by-document matrix

doc	t01	t02	t03	t04	t05	t06	t07	t08
d01	4	3	2	2	1	7	20	15

A.3.2.4 Inverted query index

The inverted query index is the fourth stage of designing the search engine process. The extended steps required to store the data to enable the index to be created are thus:

i) the unique number allocated to the query,

ii) each term acquired from the query, and

iii) the relationship of each term to each document.

The format of the inverted query index for the first ten words in the text is now presented in Table A.14. The first column represents the term and the second the document numbers each term relates to, representing the postings list.

Table A.14: Pilot 1 – IRS-I: Inverted query index

Term	doc
be	d01
is	d01
not	d01
or	d01
question	d01
that	d01
the	d01
to	d01

This concludes the design and build of the search engine process for IRS-I.

A.4 Comparative evaluation and results

In this empirically comparative evaluation and results section, the preparation of the test collection is presented, followed by the data analysis, and finally the performance measurements for Pilot 1. The final evaluation compares the results of IRS-I to IRS-H.

A.4.1 Test collection preparation

After the design and build of the IRSs, the text collection was prepared to evaluate Pilot 1 rigorously. The five activities in preparing the collection for Pilot 1 were:

i) to collate the document collection,

ii) to gather the information needs of the user,

iii) to gather the results of the user's judged relevancy for the documents,

iv) to select the terms and phrase-terms to be used for each of the two indexing methods, and

v) to present the formal queries to be used for each of the two indexing methods.

A.4.1.1 Collate the document collection

The first activity in preparing the test collection was to collate the document collection. For Pilot 1 this is a single two-page document: Hamlet Act 3 Scene 1 and therefore N = 1.

A.4.1.2 User information needs

The second activity was to gather the information needs from the user. For Pilot 1, four information needs were compiled covering one popular quotation ‘to be or not to be that is the question’ from the script of Hamlet. These are listed in Table A.15 which expresses a user’s need to search for and to retrieve those documents that are relevant to any of these four needs within document collection N.

Table A.15: Pilot 1 – Information needs

In No	Information Need
in01	I want to find all documents relevant to any of these phrases
in02	I want to find all documents relevant to the phrase "to be"
in03	I want to find all documents relevant to the phrase "to be or not to be"
in04	I want to find all documents relevant to the phrase "that is the question"

A.4.1.3 User relevant document judgement

The third activity was to prepare the test collection for the user to judge each document to determine whether a document is relevant to an information need or not. To accommodate this activity of single document d_{01}, the text was manually searched to ensure the phrases ‘to be’, ‘to be or not to be’ and ‘that is the question’ actually existed within the document, which they did. Therefore, for this test collection all four information needs were judged relevant by the user. Table A.16 represents the information provided by the user representing his/her judgment on the questionnaire stapled to the document.

Table A.16: Pilot 1 – User relevant document judgement

Document number - d0001 Please indicate whether this document is relevant to any of the following information needs (please tick)		
In No	**Information Need**	**Relevant**
in01	I want to find all documents relevant to any of these phrases	√
in02	I want to find all documents relevant to the phrase "to be"	√
in03	I want to find all documents relevant to the phrase "to be or not to be"	√
in04	I want to find all documents relevant to the phrase "that is the question"	√

To accommodate this data, an information-need-by-document matrix was designed as a table within the evaluation system. Boolean data were converted to binary quantitative data. To indicate relevant, *‘true’* was converted to 1 and to indicate non-relevant, *‘false’* was converted to 0. The results for document number d_{01} judged relevant to information needs in_{01} through to in_{04} are listed in Table A.17.

Table A.17: Pilot 1 – User information-need-by-document matrix

doc	in01	in02	in03	in04
d01	1	1	1	1

A.4.1.4 Selecting the terms and phrase-terms

The fourth activity selected phrase-terms for queries when using the hybrid indexing method, and similarly selected terms for queries when using the inverted indexing method.

Phrase-terms – hybrid index method

To evaluate the document collection for Pilot 1 using the hybrid indexing method, three multi-word phrase-terms were used to describe the four information needs. One or more phrase-term was allocated to each information need. Each phrase-term is listed in Table A.18 and its associated information need is listed in Table A.19.

Table A.18: Pilot 1 – IRS-H: Phrase-terms

pt	Phrase-term
pt01	to be
pt02	to be or not to be
pt03	that is the question

Table A.19: Pilot 1 – IRS-H: Phrase-terms per information need

In No	Information Need	pt	Phrase-term
in01	I want to find all documents relevant to any of these phrases	pt01	to be
in01	I want to find all documents relevant to any of these phrases	pt02	to be or not to be
in01	I want to find all documents relevant to any of these phrases	pt03	that is the question
in02	I want to find all documents relevant to the phrase "to be"	pt01	to be
in03	I want to find all documents relevant to the phrase "to be or not to be"	pt02	to be or not to be
in04	I want to find all documents relevant to the phrase "that is the question"	pt03	that is the question

Terms – inverted index method

To evaluate the document collection using the inverted indexing method, eight single-word terms were used to describe the four information needs. Each information need had more than one term allocated to it. The words within the three phrase-terms from the hybrid method were distinctly acquired per information need, resulting in eight single-word terms used to describe the four information needs. However, as the inverted index had no control over word proximity or ordinality the distinct set of words was presented using the bag of words concept where BoW = [be is not or question that the to]. Each term is listed in Table A.20 and its associated information need is listed in Table A.21.

Table A.20: Pilot 1 – IRS-I: The set of terms

t	Term
t01	be
t02	is
t03	not
t04	or
t05	question
t06	that
t07	the
t08	to

Table A.21: Pilot 1 – IRS-I: Term per information need

In No	Information Need	t	Word
in01	I want to find all documents relevant to any of these phrases	t01	be
		t02	is
		t03	not
		t04	or
		t05	question
		t06	that
		t07	the
		t08	to
in02	I want to find all documents relevant to the phrase "to be"	t01	be
		t08	to
in03	I want to find all documents relevant to the phrase "to be or not to be"	t01	be
		t03	not
		t04	or
		t08	to
in04	I want to find all documents relevant to the phrase "that is the question"	t02	is
		t05	question
		t06	that
		t07	the

A.4.1.5 Presenting the queries

The fifth activity was to present the queries that express each information need to the search engine using both indexing methods.

Phrase-term queries – hybrid index method

Each information need of the user creates a query that expresses what is to be searched for based upon the information need. For the hybrid index method, the queries are structured as phrase-terms and where more than one phrase-term exists, each phrase-term is separated by the logical OR operator as listed in Table A.22.

Note that query q_{01} is associated with information need in_{01}, q_{02} with in_{02} and so on.

Table A.22: Pilot 1 – IRS-H: Query per information need

In No	q	Query
in01	q01	"to be" OR "to be or not to be" OR "that is the question"
in02	q02	"to be"
in03	q03	"to be or not to be"
in04	q04	"that is the question"

Term queries – inverted index method

For the inverted index method, each information need of the user again creates queries using the terms from the bag of words concept and these queries are structured where each distinct term is separated by the logical OR operator as listed in Table A.23. Note that query q_{01} is associated with information need in_{01}, q_{02} with in_{02} and so on.

Table A.23: Pilot 1 – IRS-I: Query per information need

In No	q	Query
in01	q01	to OR be Or or OR not OR that OR is OR the OR question
in02	q02	to OR be
in03	q03	to OR be Or or OR not
in04	q04	that OR is OR the OR question

In summary, for Pilot 1 the test collection comprised a single two-page document, Hamlet Act 3 Scene 1, with four information needs, four queries, three phrase-terms using the hybrid index method, and eight single-word terms using the inverted index method.

A.4.2 Data analysis

In this section, the list of file names in the document collection is presented followed by the token and query indices for both the inverted and hybrid indexing methods. Thereafter, the three forms of collection frequencies: token, term and phrase-term, and the stop words for both indexing methods, are computed and presented. Similarly, the two forms of document frequencies: token, and term and phrase-term, for both indexing methods are computed and presented. Finally, the term-by-document matrix with the computed values of term frequency for the inverted index method and the phrase-term-by-document matrix with the computed values of phrase-term frequency for the hybrid index method are computed and presented.

A.4.2.1 Pilot 1 results – File names

The file names of all the documents in the collection were stored in a table containing the document number, the file name, and the path of the file name residing on the computer. The results for this single document collection are presented in Table A.24.

Table A.24: Pilot 1 – File names

doc	File Name	Path
d01	Hamlet.txt	C:\Thesis\Pilot 1\

A.4.2.2 Pilot 1 results – The token indices

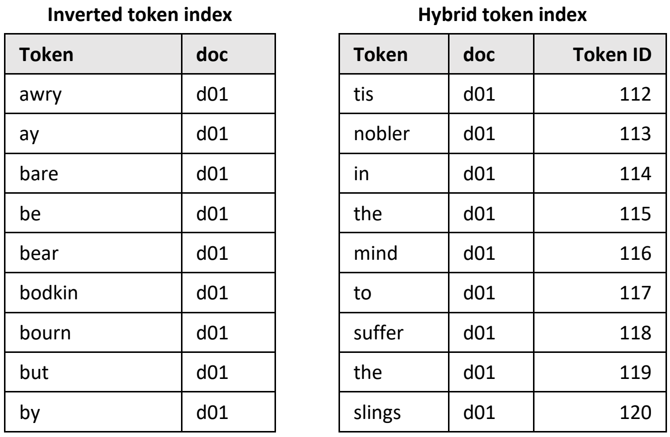
The inverted token index and the hybrid token index are now presented.

Inverted token index

Token	doc
-	d01
--	d01
‘tis	d01
a	d01
action	d01
after	d01
against	d01
all	d01
and	d01
arms	d01
arrows	d01

Hybrid token index

Token	doc	Token ID
to	d01	101
be	d01	102
or	d01	103
not	d01	104
to	d01	105
be	d01	106
that	d01	107
is	d01	108
the	d01	109
question	d01	110
whether	d01	111



Inverted token index

Token	doc
awry	d01
ay	d01
bare	d01
be	d01
bear	d01
bodkin	d01
bourn	d01
but	d01
by	d01

Hybrid token index

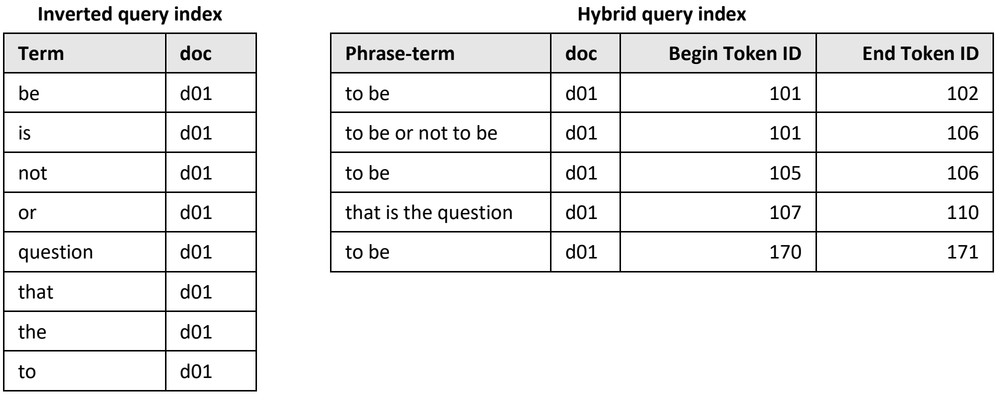
Token	doc	Token ID
tis	d01	112
nobler	d01	113
in	d01	114
the	d01	115
mind	d01	116
to	d01	117
suffer	d01	118
the	d01	119
slings	d01	120

Figure A.14: Pilot 1 – The token indexes

The inverted token index contains 170 distinct tokens, all having a relationship with a single document *d₀₁*. The hybrid token index contains 283 non-distinct tokens, all having a relationship with a single document *d₀₁*. In this pilot, the advantage of the inverted token index over the hybrid token index is fewer records and the advantage of the hybrid token index over the inverted token index is the addition of the unique Token ID preserving word ordinality and proximity. Figure A.14 presents the results of the first 20 tokens in sequential order for both token indexing methods. Note that the tokens in the inverted token index are in alphabetical order while the tokens in the hybrid token index are in the same order as they appear in the text.

A.4.2.3 Pilot 1 results – The query indices

The inverted query index and the hybrid query index are now presented. From the ten words used within the queries, the inverted query index contains eight distinct terms all having a relationship with a single document *d₀₁*. The hybrid query index contains five phrase-terms, one occurring three times in the text and two once, all having a relationship with a single document *d₀₁*. In this pilot, the advantages that the hybrid query index has over the inverted query index are fewer records and the addition of the begin Token ID and end Token ID (the positioning of these phrases in the text is now evident). Figure A.15 presents the results in sequential order for both query index methods.



Inverted query index

Term	doc
be	d01
is	d01
not	d01
or	d01
question	d01
that	d01
the	d01
to	d01

Hybrid query index

Phrase-term	doc	Begin Token ID	End Token ID
to be	d01	101	102
to be or not to be	d01	101	106
to be	d01	105	106
that is the question	d01	107	110
to be	d01	170	171

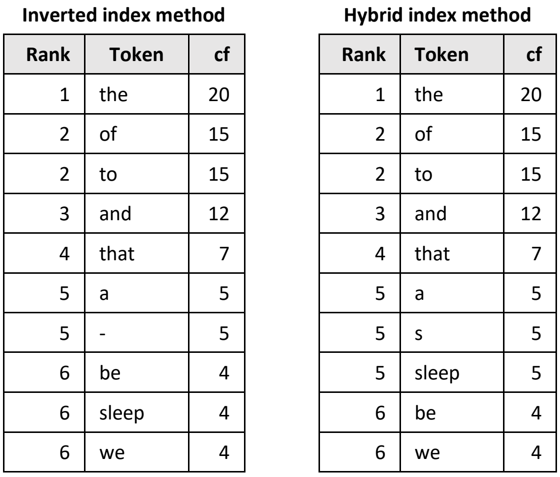
Figure A.15: Pilot 1 – The query indices

A.4.2.4 Pilot 1 results – Collection frequency

Using the collection frequency at the phrase-term level, the number of occurrences of each phrase-term within a document collection can be computed. To compute the collection frequency, Structured Query Language (SQL) scripts are required to retrieve the data from the hybrid query index, not the hybrid token index, as only the hybrid query index contained the phrase-term. Thereafter the computations are written to the data store. The token based, the term based, and the phrase-term based collection frequencies for both methods are now presented.

Pilot 1 results – Token based collection frequency

For the inverted index method, of the 282 tokens acquired from the text, 170 were distinct, and for the hybrid index method, of the 283 tokens acquired from the text, 170 were distinct. Of these, the first top ten ranked token based collection frequencies, ranked in descending order, for both methods are now presented.



Inverted index method

Rank	Token	cf
1	the	20
2	of	15
2	to	15
3	and	12
4	that	7
5	a	5
5	-	5
6	be	4
6	sleep	4
6	we	4

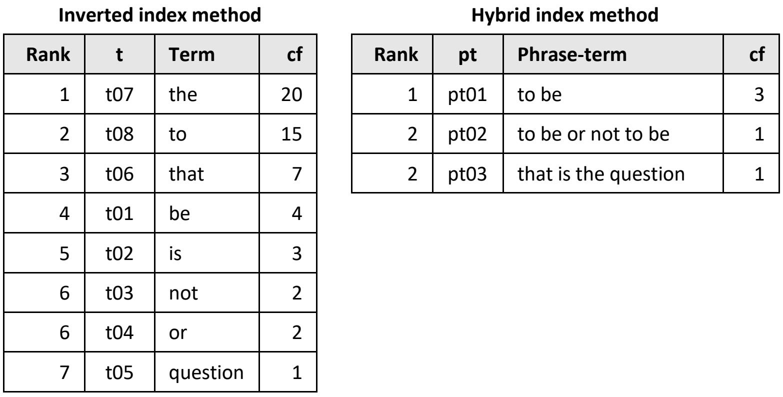
Hybrid index method

Rank	Token	cf
1	the	20
2	of	15
2	to	15
3	and	12
4	that	7
5	a	5
5	s	5
5	sleep	5
6	be	4
6	we	4

Figure A.16: Pilot 1 – First top ten ranked token collection frequencies

Pilot 1 results – Term and phrase-term based collection frequency

The term based collection frequencies, ranked in descending order, for the inverted index method and the phrase-term collection frequencies for the hybrid index method are now presented in Figure A.17.



Inverted index method

Rank	t	Term	cf
1	t07	the	20
2	t08	to	15
3	t06	that	7
4	t01	be	4
5	t02	is	3
6	t03	not	2
6	t04	or	2
7	t05	question	1

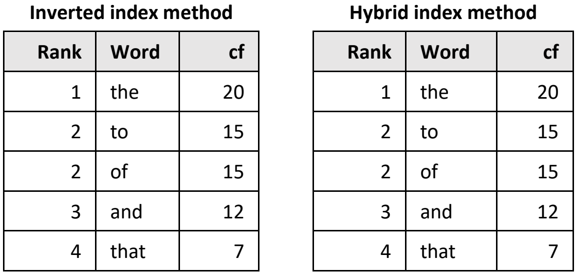
Hybrid index method

Rank	pt	Phrase-term	cf
1	pt01	to be	3
2	pt02	to be or not to be	1
2	pt03	that is the question	1

Figure A.17: Pilot 1 – Ranked term/phrase-term collection frequencies

Pilot 1 results – Stop words

The concept of stop words is a good way to describe the use of collection frequency best. Stop words, according to Ha et al. (2002), are the most frequently occurring tokens normally ignored within a collection. Using the collection frequency at the token level, the number of occurrences of each token within a document collection can be computed. A method of presenting these data is to provide a ranking table. For Pilot 1, the top five stop words ranked in descending order for both the inverted and hybrid index methods are provided in Figure A.18.



Inverted index method

Rank	Word	cf
1	the	20
2	to	15
2	of	15
3	and	12
4	that	7

Hybrid index method

Rank	Word	cf
1	the	20
2	to	15
2	of	15
3	and	12
4	that	7

Figure A.18: Pilot 1 – Top five stop words

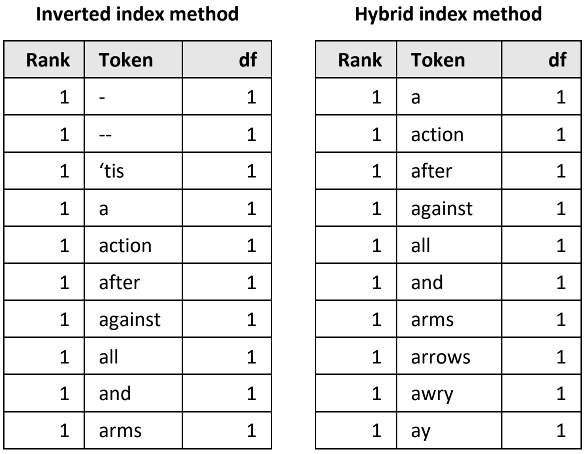
The collection frequency (cf_t) for the token *'the'* is 20 as this token occurs 20 times in the document collection. These data compare favourably with the work of Ha et al. (2002) where their top five tokens from a Wall Street document collection were: *'the', 'of', 'to', 'a',* and *'and'.*

A.4.2.5 Pilot 1 results – Document frequency

Document frequency is defined as the number of documents in which a term or phrase-term occurs. In this research for the inverted index method using single-word terms, document frequency is denoted by df_t and for the hybrid indexing method using multi-word phrase-terms by df_{pt}. Owing to space limitations, only the first ten token-based document frequencies, for both indexing methods, are presented in Figure A.19.

Pilot 1 results – Token based document frequency

The first top ten ranked token-based document frequencies, ranked in descending order, for both methods are now presented.



Inverted index method

Rank	Token	df
1	-	1
1	--	1
1	'tis	1
1	a	1
1	action	1
1	after	1
1	against	1
1	all	1
1	and	1
1	arms	1

Hybrid index method

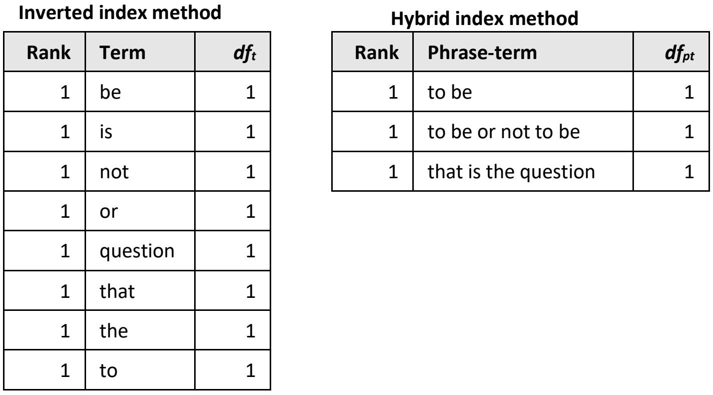
Rank	Token	df
1	a	1
1	action	1
1	after	1
1	against	1
1	all	1
1	and	1
1	arms	1
1	arrows	1
1	awry	1
1	ay	1

Figure A.19: Pilot 1 – Ranked token document frequencies

Pilot 1 results – Term and phrase-term based document frequencies

Using the document frequency at the token level, the number of occurrences of each token within a document can be computed. To compute the document frequency, SQL scripts were required to acquire the data from the hybrid token index and thereafter write them to a new table in the data store. As there is only one document in this collection, the document frequency for all tokens would be equal to one, and therefore the document frequency (*df*) for the token *'the'* is one, as this token occurs in one document, the only document in the collection. But in this study, single-word terms within a query are not equal to a token. Multi-word terms are used as phrase-terms and the source data for this does not reside in the hybrid token index but in the hybrid query index.

Therefore, because of the design of this IRS, in the data analysis the collection frequency is kept, but the design for document frequency computations are reworked. If we take the first phrase-term pt_{01} *'to be'* we then need to compute, at the phrase-term level, the number of occurrences of each phrase-term within a document, and therefore would again use SQL but acquire the data from the hybrid query index. Therefore, in this study the document frequency (df_{pt}) is defined as the number of documents in which phrase-term *pt* occurs. From this document collection, the document frequencies for each of the three-phase terms $df_{pt01,\ d01}$, $df_{pt02,\ d01}$, $df_{pt03,\ d01}$, and $df_{pt04,\ d01}$ are all equal to one, as they all occur in the single document d_{01} within the collection, at least once. The document frequencies for the three phrase-terms searched for in this document collection ranked in descending order are provided in Figure A.20. The term based collection frequencies, ranked in descending order, for the inverted index method and the phrase-term collection frequencies for the hybrid index method are now presented.



Inverted index method

Rank	Term	df_t
1	be	1
1	is	1
1	not	1
1	or	1
1	question	1
1	that	1
1	the	1
1	to	1

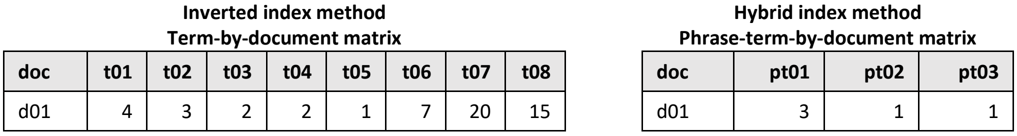
Hybrid index method

Rank	Phrase-term	df_{pt}
1	to be	1
1	to be or not to be	1
1	that is the question	1

Figure A.20: Pilot 1 – Ranked term/phrase-term document frequencies

A.4.2.6 Term frequency, phrase-term frequency and matrices

For the inverted index, the term-by-document matrix is a method of capturing the number of times each term within a query occurs in each document. The rows represent the documents, the columns represent the terms, and the values in each cell represent the term frequency. Alternatively, for the hybrid index, the phrase-term-by-document matrix is a method of capturing the number of times each phrase-term within a query occurs in each document. The rows represent the documents, the columns represent the phrase-terms, and the values in each cell represent the phrase-term frequency. Figure A.21 presents the results in Pilot 1 for both matrices for the two methods.



Inverted index method
Term-by-document matrix

doc	t01	t02	t03	t04	t05	t06	t07	t08
d01	4	3	2	2	1	7	20	15

Hybrid index method
Phrase-term-by-document matrix

doc	pt01	pt02	pt03
d01	3	1	1

Figure A.21: Pilot 1 – Matrices

Referring to the term-by-document matrix $tf_{t01,\ d01}$ = 4 indicating that term t_{01} occurs in document d_{01} four times, $tf_{t02,\ d01}$ = 3 indicating that term t_{02} occurs in document d_{01} three times, etc. Referring to the phrase-term-by-document matrix $ptf_{pt01,\ d01}$ = 3 indicating that phrase-term pt_{01} occurs in document d_{01} three times, and the remaining two phrase-terms occur once in document d_{01}.

A.4.3 Performance measurements

In this section, the three information-need-by-document matrices for the user, inverted index method, and hybrid index method are presented. Based on these matrices, the 2x2 contingency tables for both methods are created and presented. These matrices and contingency tables form the basis for the computation of the performance measurements. Thereafter the computations for Precision, Recall, and F-measure for both methods are performed and presented. Finally, the full evaluation for both methods is performed presenting summarised statistics of the computations in both tabular and graphical formats.

To develop and build the document statistics, the same development tools used to build the indices are used to expand the data store to accommodate additional tables to store these statistical data. Computational algorithms making use of SQL are used to acquire data from the hybrid token index within the data store, which compute and then store the results. Referring to Figure A.22, the traditional 2x2 contingency table (Cleverdon & Keen, 1966; Cleverdon, 1967) is expanded to accommodate additional computations. There are two tests: the first test is to determine whether the IRS retrieves a document by matching a query (related to an information need) to a document or not, and the second test is for the user who judges whether a document is relevant to a query (information need) or not. Reading the contingency table from left-to-right and top-to-bottom, the first cell relates to true positive (*tp*) and is defined as the number of user relevant documents retrieved by the IRS. Critically, these documents are the ones searched for relevant to an information need. Moving to the right, false positive (*fp*) is the value that must be kept as low as possible to limit the user's perusal of non-relevant documents and is defined as the number of user non-relevant documents retrieved by the IRS. Depending on the indexing method used, the IRS retrieves the document because the term or phrase exists in it. The false negative (*fn*) is the number of documents relevant to the user not retrieved by the IRS. This value should be as low as possible, indicating the effectiveness of the IRS at detecting the occurrence of a term or phrase-term within a document. The true negative (*tn*) measurement refers to the number of a user's documents judged non-relevant that are not retrieved by the IRS. The sum of true positive and false negative, the number of relevant documents in a collection, as judged by the user, is represented by *tp* + *fn.* The sum of false positive and true negative, the number of non-relevant documents judged by the user, is represented by *fp* + *tn*. The sum of the positives, true positive and false positive, is the number of documents retrieved by the IRS, represented by *tp* + *fp,* and the sum of the negatives, false negative and true negative, the number of documents not retrieved by the IRS, is represented by *fn* + *tn*. N represents the number of documents in the collection and is the sum of *tp* + *fp* + *fn* + *tn*.

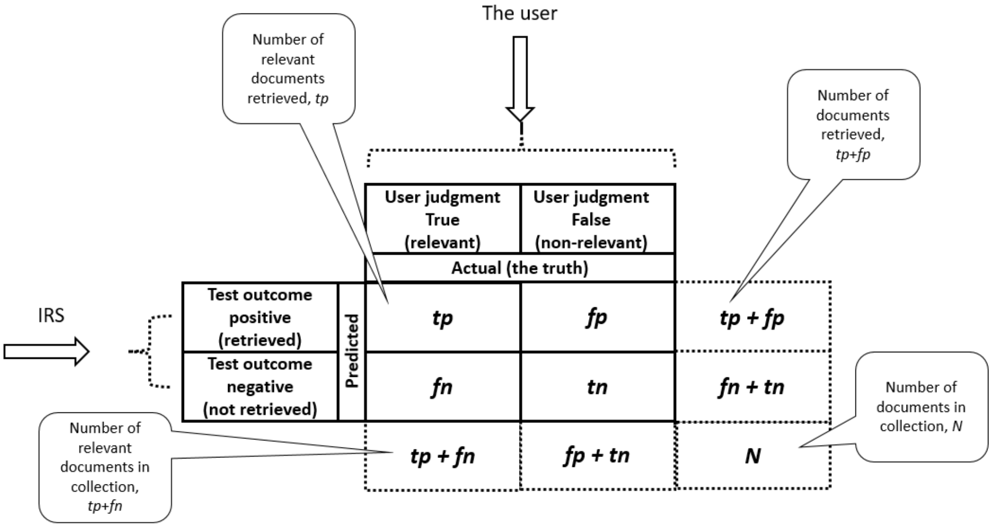

The user
Number of relevant documents retrieved, tp
Number of documents retrieved, tp+fp
User judgment True (relevant)
User judgment False (non-relevant)
Actual (the truth)
IRS
Test outcome positive (retrieved)
Test outcome negative (not retrieved)
Predicted
tp
fp
tp + fp
fn
tn
fn + tn
tp + fn
fp + tn
N
Number of documents in collection, N
Number of relevant documents in collection, tp+fn

Figure A.22: Pilot 1 – The expanded contingency table

By rearranging the values from this table, the values can be presented as a user information-need-by-document matrix. Listed below (Figure A.23) are the results from the user's information-need-by-matrix (the documents that the user has judged relevant to the information needs) together with the information-need-by-document matrix produced by IRS-I, the IRS utilising the inverted indexes (the documents that IRS-I has retrieved from the collection), and finally the information-need-by-document matrix produced by IRS-H, the IRS utilising the hybrid indexes (the documents that IRS-H has retrieved from the collection).

User

doc	in01	in02	in03	in04
d01	1	1	1	1

IRS-I

Inverted index method

doc	in01	in02	in03	in04
d01	1	1	1	1

IRS-H

Hybrid index method

doc	in01	in02	in03	in04
d01	1	1	1	1

Figure A.23: Pilot 1 – User and IRS information-need-by-document matrices

Using these data from the matrices, the performance measurements for both IRSs can now be computed by applying the following rules:

i) If the user judged the document as relevant (relevant = true) and the IRS retrieved the document (retrieved = positive) then *tp* = 1 else *tp* = 0.

ii) If the user judged the document as non-relevant (relevant = false) and the IRS retrieved the document (retrieved = positive) then *fp* = 1 else *fp* = 0.

iii) If the user judged the document as relevant (relevant = true) and the IRS did not retrieve the document (retrieved = negative) then *tn* = 1 else *tn* = 0.

iv) If the user judged the document as non-relevant (relevant = false) and the IRS did not retrieve the document (retrieved = negative) then *fn* = 1 else *fn* = 0.

Taking the values from the matrices where the value of 1 represents ‘true’ for the user’s judgement or positive for the IRS’s judgement and the value 0 represents ‘false’ or negative, the rules become:

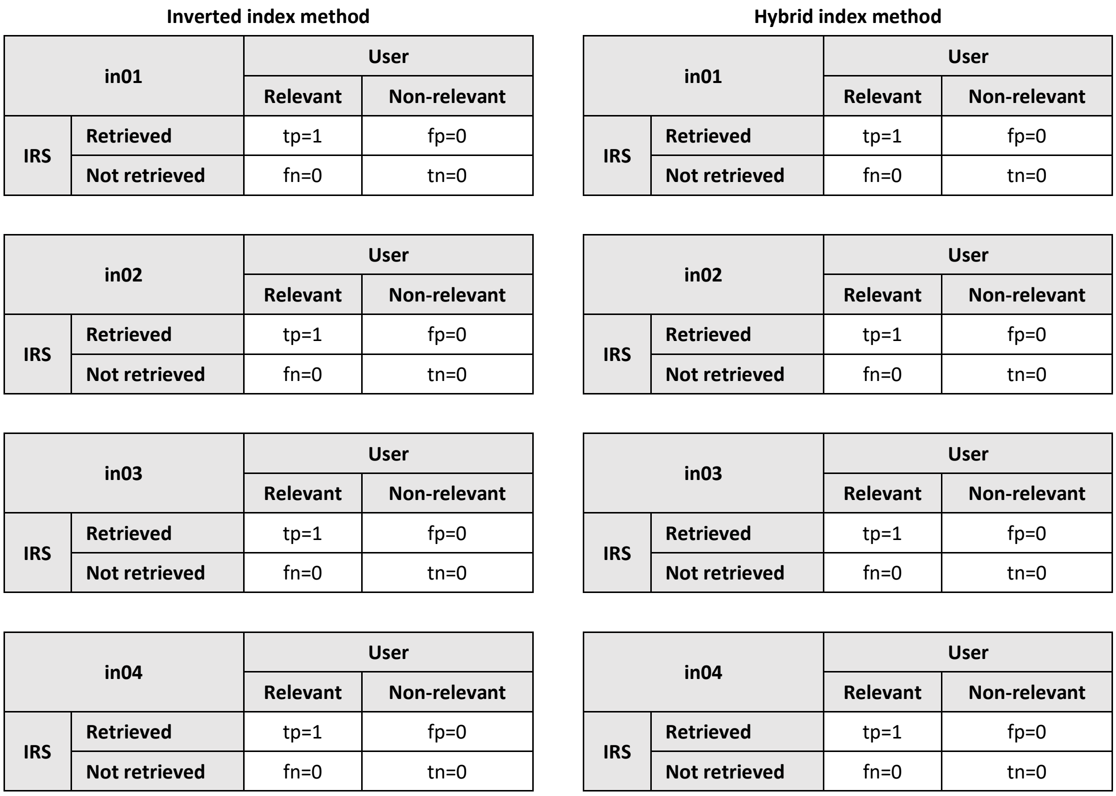
i) If user in_{01} = 1 and IRS in_{01} = 1 then *tp* = 1 else *tp* = 0.

ii) If user in_{01} = 0 and IRS in_{01} = 1 then *fp* = 1 else *fp* = 0.

iii) If user in_{01} = 1 and IRS in_{01} = 0 then *tn* = 1 else *tn* = 0.

iv) If user in_{01} = 0 and IRS in_{01} = 0 then *fn* = 1 else *fn* = 0.

The results are now presented as 2x2 contingency tables (Figure A.24).



Inverted index method

in01		User	
		Relevant	Non-relevant
IRS	Retrieved	tp=1	fp=0
	Not retrieved	fn=0	tn=0

in02		User	
		Relevant	Non-relevant
IRS	Retrieved	tp=1	fp=0
	Not retrieved	fn=0	tn=0

in03		User	
		Relevant	Non-relevant
IRS	Retrieved	tp=1	fp=0
	Not retrieved	fn=0	tn=0

in04		User	
		Relevant	Non-relevant
IRS	Retrieved	tp=1	fp=0
	Not retrieved	fn=0	tn=0

Hybrid index method

in01		User	
		Relevant	Non-relevant
IRS	Retrieved	tp=1	fp=0
	Not retrieved	fn=0	tn=0

in02		User	
		Relevant	Non-relevant
IRS	Retrieved	tp=1	fp=0
	Not retrieved	fn=0	tn=0

in03		User	
		Relevant	Non-relevant
IRS	Retrieved	tp=1	fp=0
	Not retrieved	fn=0	tn=0

in04		User	
		Relevant	Non-relevant
IRS	Retrieved	tp=1	fp=0
	Not retrieved	fn=0	tn=0

Figure A.24: Pilot 1 – Information needs 2x2 contingency tables

Calculating Precision

Precision is a measurement of how well a search query is structured using words to express an information need of the user. It uses a mathematical formula comparing user relevant documents to IRS retrieved and not retrieved documents. Precision is defined as $P = tp/(tp + fp)$ and is the ratio of the number of user relevant documents retrieved by the IRS and the number of documents retrieved by the IRS. In this research, the results for Precision are represented as percentages.

For the inverted index method, Precision for *in₀₁* is therefore

$$P_{in01} = \frac{tp}{tp+fp} = \frac{1}{1+0} = 1\ or\ 100\%$$

As the data are equal in all cases:

$$P_{in02} = 1\ or\ 100\%$$

$$P_{in03} = 1\ or\ 100\%$$

$$P_{in04} = 1\ or\ 100\%$$

For the hybrid index method, these data are again identical and therefore for each of the four information needs *P* = 1 or 100%:

$$P_{in01} = 1\ or\ 100\%$$

$$P_{in02} = 1\ or\ 100\%$$

$$P_{in03} = 1\ or\ 100\%$$

$$P_{in04} = 1\ or\ 100\%$$

Calculating Recall

Recall is a measurement of how well an IRS's indexing system handles text from documents. Recall is defined as $R = tp/(tp + fn)$ and is the ratio of the number of user relevant documents retrieved by the IRS and the number of user relevant documents in the collection. In this research, the results for Recall are represented as percentages.

For the inverted index method, Recall for *in_{01}* is therefore

$$R_{in01} = \frac{tp}{tp+fn} = \frac{1}{1+0} = 1\ or\ 100\%$$

As the data are equal in all cases:

$$R_{in02} = 1\ or\ 100\%$$

$$R_{in03} = 1\ or\ 100\%$$

$$R_{in04} = 1\ or\ 100\%$$

For the hybrid index method, the data are again identical and therefore for each of the four information needs *R* = 1 or 100%:

$$R_{in01} = 1\ or\ 100\%$$

$$R_{in02} = 1\ or\ 100\%$$

$$R_{in03} = 1\ or\ 100\%$$

$$R_{in04} = 1\ or\ 100\%$$

Calculating F-measure

F-measure is a measurement of how effective an IRS is in retrieving relevant documents and not retrieving non-relevant documents. F-measure is defined as $F = 2PR/((P + R))$ and is the ratio of twice the product of Precision and Recall and the sum of Precision and Recall. In this research, the results for F-measure are represented as percentages.

For the inverted index method, F-measure for *in₀₁* is therefore

$$F_{in01} = \frac{2PR}{(P+R)} = \frac{2*1*1}{1+0} = 1\ or\ 100\%$$

As the data are equal in all cases:

$$F_{in02} = 1\ or\ 100\%$$

$$F_{in03} = 1\ or\ 100\%$$

$$F_{in04} = 1\ or\ 100\%$$

For the hybrid index method, these data are again identical and therefore for each of the four information needs *F* = 1 or 100%:

$$F_{in01} = 1\ or\ 100\%$$

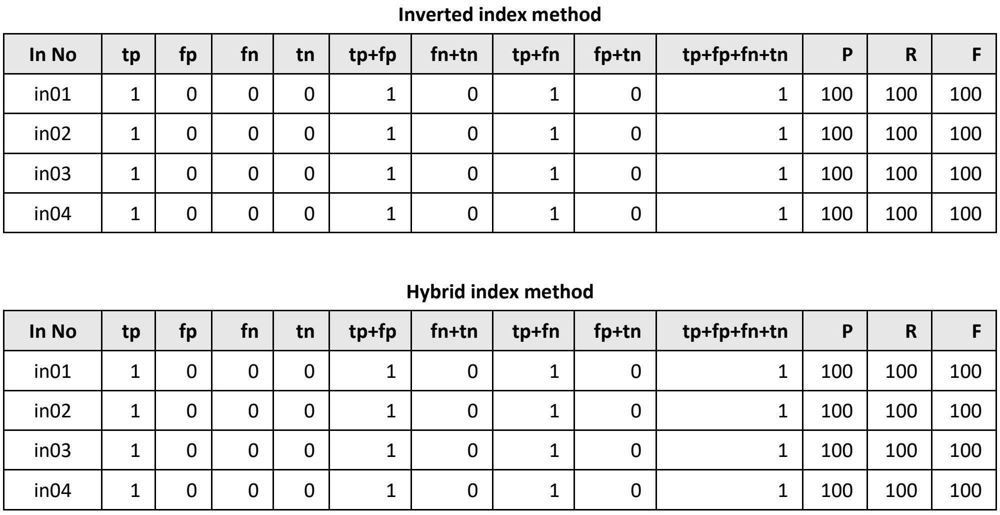
$$F_{in02} = 1\ or\ 100\%$$

$$F_{in03} = 1\ or\ 100\%$$

$$F_{in04} = 1\ or\ 100\%$$

A.5 Evaluation

These performance measurements are all listed in table form for both the inverted and hybrid indexing methods below. The first column holds the information need number, the second the number of true positives, the third the number of false positives, the fourth the number of false negatives, the fifth the number of true negatives, the sixth all the positives, the seventh all the negatives, the eighth the true positives plus false negatives, the ninth the false positives plus true negatives, the tenth the number of documents in the collection where N = *tp* + *fp* + *fn* + *tn*, the eleventh, twelfth and thirteenth columns represent the Precision, Recall, and F-measure values represented as percentages.



Inverted index method

In No	tp	fp	fn	tn	tp+fp	fn+tn	tp+fn	fp+tn	tp+fp+fn+tn	P	R	F
in01	1	0	0	0	1	0	1	0	1	100	100	100
in02	1	0	0	0	1	0	1	0	1	100	100	100
in03	1	0	0	0	1	0	1	0	1	100	100	100
in04	1	0	0	0	1	0	1	0	1	100	100	100

Hybrid index method

In No	tp	fp	fn	tn	tp+fp	fn+tn	tp+fn	fp+tn	tp+fp+fn+tn	P	R	F
in01	1	0	0	0	1	0	1	0	1	100	100	100
in02	1	0	0	0	1	0	1	0	1	100	100	100
in03	1	0	0	0	1	0	1	0	1	100	100	100
in04	1	0	0	0	1	0	1	0	1	100	100	100

Figure A.25: Pilot 1 – Performance measurements

In Figure A.26, the four information needs for Pilot 1 are presented for each of the IRSs. The diamonds represent the F-measure values for IRS-I, and the squares represent the F-measure values for IRS-H.

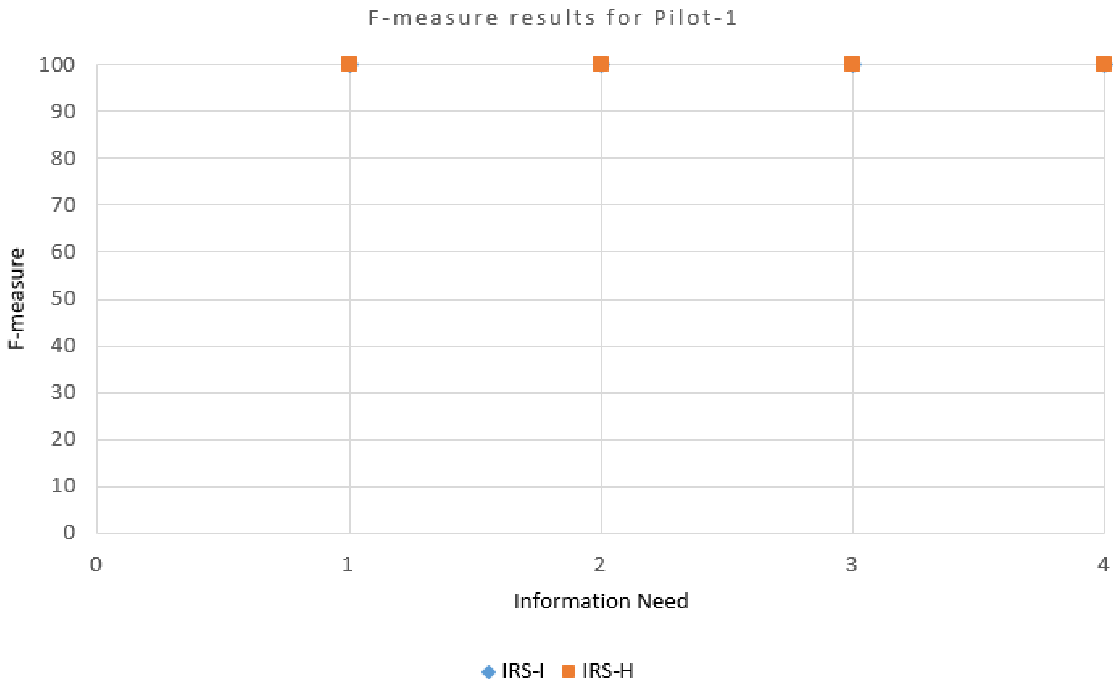


Figure A.26: Pilot 1 – IRS-I and IRS-H performance measurements

In Pilot 1, four information needs were defined by the user and after perusing the single document in the collection, the user judged the document that was relevant to all four information needs. Both IRS-I and IRS-H retrieved the document for all four queries as these systems judged the document that was relevant to all four information needs. All four queries, representing the four information needs, produced identical data for both the inverted and hybrid indexing methods with the performance measurements of Precision, Recall, and F-measure all achieving 100 percent.

A.6 Summary

At this stage there is currently no evidence to prove that the functionality of IRS-H is more effective than IRS-I or vice versa. However, the evidence does suggest that the functionality in handling vocabulary mismatch using the concepts of phrase-terms and Token IDs in the hybrid indexing method is equally as effective as when using the inverted indexing method as the results are the same. The design findings for Pilot 1 are summarised in Table A.25.

Table A.25: Pilot 1 – Summary of design findings

Pilot	Stage	Finding
1	Information gathering	Pilot 1 was based on the book Hamlet Act 3 Scene 1 written by William Shakespeare (Shakespeare, 2018).
1	Information gathering	The inverted index was replaced by the pair of hybrid indices: the hybrid token index and the hybrid query index (Figure 4.2).
1	Information gathering	Content acquisition: the content from the single two-page document from Hamlet Act 3 Scene 1 was acquired from the pdf document and converted to text successfully.
1	Information gathering	Text transformation: text was case folded to lowercase, special characters were removed, and the tokens of text identified between delimiters were tokenised successfully.
1	Information gathering	Hybrid token index: the population of the document numbers and unique token IDs were allocated successfully. Thereafter the hybrid token index was populated with the tokens, document numbers, and unique token IDs.
1	Search engine	Phrase-term: four phrase-terms provided by the user (the researcher in this pilot) were presented correctly, all in lowercase without special characters.
1	Search engine	Phrase-term query: these phrase-terms were expressed as four queries, three singular and one expanded query.
1	Search engine	Hybrid query index: the four queries were presented to the hybrid query index, which was thereafter populated with the phrase-terms, and unique begin and end token IDs.
1	Search engine	The hybrid query index interrogated the hybrid token index successfully and where a match was found (a phrase-term existed in a document) the document number was returned and the hybrid query index updated accordingly.
1	Design	Phrase-term frequency (ptf) needed to replace term frequency (tf), as by design it was the number of phrase-terms that were required to be calculated rather than single terms used in the inverted indexing method.
1	Design	Converting ptf values to binary and the population of the phrase-term-by-document matrix (rather than the term-by-document matrix used in the inverted indexing method) with these values was successful.
1	Design	Stopping, the removal of stop words, the use of stemming, classifiers and suffix stripping were needless in this design as the tokens were not to be changed in any way, thus preventing an exact match.
1	Design	The IRS was able to match phrase-terms expressed in queries, held within the hybrid query index, to phrase-terms within the text of document held within the hybrid token index, exactly.
1	Design	Performance measurements were unusable as the judgment results from the user were unavailable and therefore not tested.
1	Design	The sequentially generated number ranges made by the IRS were limiting. This applied to the document number and the indexes' unique token ID. To remedy this issue the number ranges were expanded accordingly.
1	Design	The document length was a limiting factor that disallowed any benefit IRS-H may have had over IRS-I and vice versa. To remedy this issue the document length was increased.
1	Design	At this stage at the end of Pilot 1 there was no evidence to suggest that the functionality of IRS-H was more effective than IRS-I and vice versa.

Reflecting back to Figure 3.10 of Hevner (2007:2) and Figure 3.11 that explains the three design cycles for this research in Chapter Three, the design cycle is now complete for Pilot 1. From the measurements and the IRS evaluation, any design and build issues evidenced were now, through the design cycle, driven to perform Pilot 2.

APPENDIX B: PILOT 2 ULYSSES

B.1 Design issues and build updates

From a design perspective, four issues became evident while evaluating Pilot 1: two design issues, one data issue, and one data analysis issue. The first design issue was the limitation in both IRSs' generated document number, and the second was the limitation in the IRS-H generated Token ID number. The third issue was that the evaluation results were identical and therefore not all combinations of 2x2 contingency tables were tested rigorously.

For Pilot 2, the build updates included increasing the document number by two digits to accommodate 10,000 numbers rather than the limited 100, and re-formatting the IRS-H generated Token ID from three digits to eight, thus accommodating 10,000,000 tokens rather than the limited 1,000.

The data issue related to various un-tested outcome combinations possible using the 2x2 contingency table. Only *tp* was tested as true whereas *fp*, *fn*, and *tn* were not. Therefore, the information needs in this Pilot 2 were purposively selected to test the various combinations. In particular, *fn* where the user judges an information need as non-relevant but the IRS retrieves the document as it matches the term or phrase within the query to the document. Therefore, in_{05} was purposively judged non-relevant to evaluate both IRSs for this outcome.

Further, analysing the data to present the results for collection frequency (excluding stop words), and document frequency, provided no value to this research in the computations performed. Only the term frequency and phrase-term frequency added value, as these were required for the matrices. Therefore, collection frequency and document frequency computations were omitted from this research for Pilot 2 onwards.

B.2 Comparative evaluation and results

Moving now into Pilot 2, in this empirically comparative evaluation and results section, the preparation of the Pilot 2 test collection is presented, followed by the data analysis and finally the performance measurements. The final evaluation compared the results of IRS-I using the inverted index method to IRS-H.

B.2.1 Test collection preparation

After the design and build of the IRSs, the text collection was prepared to evaluate Pilot 2 rigorously. The document collection was collated by selecting a single 666-page document, the book *Ulysses* written by James Joyce. Therefore, document collection N remained at 1. For Pilot 2, 26 information needs were compiled by the user (the researcher in this pilot), 20 referring to people's names, and six referring to lengthy or unusual words. These were all purposively selected, firstly to test word ordinality and proximity in people's names, and secondly to test whether the IRSs could accommodate rather lengthy and unusual words used by the book’s author. In addition, the user established these information needs and thereafter performed his judgments whether these information needs containing these people’s names and unusual words were relevant to the document. Table B.1 presents the information needs that were provided by the user, including the relevancy judgment. A tick represents ‘relevant’ while a cross represents non-relevant.

Table B.1: Pilot 2 – User judged information needs

Document number - d0001 Please indicate whether this document is relevant to any of the following information needs (please tick)		
In No	**Information Need**	**Relevant**
in01	I want to find all documents relevant to the person's name alderman cowley	√
in02	I want to find all documents relevant to the person's name ben dollard	√
in03	I want to find all documents relevant to the person's name buck mulligan	√
in04	I want to find all documents relevant to the person's name councillor abraham lyon	√
in05	I want to find all documents relevant to the person's name father cowley	X
in06	I want to find all documents relevant to the person's name jimmy henry	√
in07	I want to find all documents relevant to the person's name john fanning	√
in08	I want to find all documents relevant to the person's name john wyse nolan	√
in09	I want to find all documents relevant to the person's name lord edward street	√
in10	I want to find all documents relevant to the person's name martin cunningham	√
in11	I want to find all documents relevant to the person's name miss douce	√
in12	I want to find all documents relevant to the person's name miss kennedy	√
in13	I want to find all documents relevant to the person's name mr boylan	√
in14	I want to find all documents relevant to the person's name mr dedalus	√
in15	I want to find all documents relevant to the person's name mr m e solomons	√
in16	I want to find all documents relevant to the person's name mr owen	X
in17	I want to find all documents relevant to the person's name mr power	√
in18	I want to find all documents relevant to the person's name mr thomas kernan	√
in19	I want to find all documents relevant to the person's name reverend hugh c love	√
in20	I want to find all documents relevant to the person's name william humble	√
in21	I want to find all documents relevant to the term wavyavyeavyheavyeavyevyevyhair	√
in22	I want to find all documents relevant to the term frseeeeeeeeeeeeeeeeeeeefrong	√
in23	I want to find all documents relevant to the term honorificabilitudinitatibus	√
in24	I want to find all documents relevant to the term whorusalaminyourhighhohhhh	√
in25	I want to find all documents relevant to the term theolologicophilolological	√
in26	I want to find all documents relevant to the term handsomemarriedwomanrubbedagainstwidebehindinclonskeatram	√

According to the user there were two information needs that were judged non-relevant, *$in_{05,}$* representing the information need *“I want to find all documents relevant to the person's name father cowley”* and *$in_{16,}$* representing the information need *“I want to find all documents relevant to the person's name mr owen*”.

The next activity in preparing the test collection was selecting the phrase-terms to be used in the hybrid indexing method queries and selecting the terms to be used in the inverted indexing method queries. To evaluate the document collection for Pilot 2 using the hybrid indexing method, 20 multi-word phrase-terms and six single-word phrase-terms were used in the queries, to express the 26 information needs. Each information need had one phrase-term allocated to it. Each phrase-term and its associated information need are listed in Table B.2.

Table B.2: Pilot 2 – IRS-H: Phrase-terms per information need

In No	pt	Phrase-term
in01	pt01	alderman cowley
in02	pt02	ben dollard
in03	pt03	buck mulligan
in04	pt04	councillor abraham lyon
in05	pt05	father cowley
in06	pt06	jimmy henry
in07	pt07	john fanning
in08	pt08	john wyse nolan
in09	pt09	lord edward street
in10	pt10	martin cunningham
in11	pt11	miss douce
in12	pt12	miss kennedy
in13	pt13	mr boylan
in14	pt14	mr dedalus
in15	pt15	mr m e solomons
in16	pt16	mr owen
in17	pt17	mr power
in18	pt18	mr thomas kernan
in19	pt19	reverend hugh c love
in20	pt20	william humble
in21	pt21	wavyavyeavyheavyeavyevyevyhair
in22	pt22	frseeeeeeeeeeeeeeeeeeeefrong
in23	pt23	honorificabilitudinitatibus
in24	pt24	whorusalaminyourhighhohhhh
in25	pt25	theolologicophilolological
in26	pt26	handsomemarriedwomanrubbedagainstwidebehindinclonskeatram

To evaluate the document collection for Pilot 2 using the inverted indexing method, 46 single-word terms were used in the queries to express the 26 information needs. Each information need had one or more terms allocated to it. Each term is listed in Table B.3 and each term and its associated information need are listed in Table B.4.

Table B.3: Pilot 2 – IRS-I: Terms

t	Term
t01	abraham
t02	alderman
t03	ben
t04	boylan
t05	buck
t06	c
t07	councillor

t	Term
t08	cowley
t09	cunningham
t10	dedalus
t11	dollard
t12	douce
t13	e
t14	edward
t15	fanning
t16	father
t17	frseeeeeeeeeeeeeeeeeeeefrong
t18	handsomemarriedwomanrubbedagainstwidebehindinclonskeatram
t19	henry
t20	honorificabilitudinitatibus
t21	hugh
t22	humble
t23	jimmy
t24	john
t25	kennedy
t26	kernan
t27	lord
t28	love
t29	lyon
t30	m
t31	martin
t32	miss
t33	mr
t34	mulligan
t35	nolan
t36	owen
t37	power
t38	reverend
t39	solomons
t40	street
t41	theolologicophilolological
t42	thomas
t43	wavyavyeavyheavyeavyevyevyhair
t44	whorusalaminyourhighhohhhh
t45	william
t46	wyse

Table B.4: Pilot 2 – IRS-I: Terms per information need

In No	t	Term
in01	t02	alderman
in01	t08	cowley
in02	t03	ben
in02	t11	dollard
in03	t05	buck
in03	t34	mulligan

In No	t	Term
in04	t01	abraham
in04	t07	councillor
in04	t29	lyon
in05	t08	cowley
in05	t16	father
in06	t19	henry
in06	t23	jimmy
in07	t15	fanning
in07	t24	john
in08	t24	john
in08	t35	nolan
in08	t46	wyse
in09	t14	edward
in09	t27	lord
in09	t40	street
in10	t09	cunningham
in10	t31	martin
in11	t12	douce
in11	t32	miss
in12	t25	kennedy
in12	t32	miss
in13	t04	boylan
in13	t33	mr
in14	t10	dedalus
in14	t33	mr
in15	t13	e
in15	t30	m
in15	t33	mr
in15	t39	solomons
in16	t33	mr
in16	t36	owen
in17	t33	mr
in17	t37	power
in18	t26	kernan
in18	t33	mr
in18	t42	thomas
in19	t06	c
in19	t21	hugh
in19	t28	love
in19	t38	reverend
in20	t22	humble
in20	t45	william
in21	t43	wavyavyeavyheavyeavyevyevyhair
in22	t17	frseeeeeeeeeeeeeeeeeeeefrong
in23	t20	honorificabilitudinitatibus
in24	t44	whorusalaminyourhighhohhhh
in25	t41	theolologicophilolological
in26	t18	handsomemarriedwomanrubbedagainstwidebehindinclonskeatram

The next activity was to present the queries that express each information need to the search engine, using both indexing methods. For the hybrid index method, the queries were structured as phrase-terms. As this pilot uses single phrase-terms, it is unnecessary to use the logical OR operator in the queries. The 26 information needs and their related 26 queries, all having a one-to-one relationship, are presented in Table B.5.

Table B.5: Pilot 2 – IRS-H: Query per information need

In No	q	Query
in01	q01	"alderman cowley"
in02	q02	"ben dollard"
in03	q03	"buck mulligan"
in04	q04	"councillor abraham lyon"
in05	q05	"father cowley"
in06	q06	"jimmy henry"
in07	q07	"john fanning"
in08	q08	"john wyse nolan"
in09	q09	"lord edward street"
in10	q10	"martin cunningham"
in11	q11	"miss douce"
in12	q12	"miss kennedy"
in13	q13	"mr boylan"
in14	q14	"mr dedalus"
in15	q15	"mr m e solomons"
in16	q16	"mr owen"
in17	q17	"mr power"
in18	q18	"mr thomas kernan"
in19	q19	"reverend hugh c love"
in20	q20	"william humble"
in21	q21	"wavyavyeavyheavyeavyevyevyhair"
in22	q22	"frseeeeeeeeeeeeeeeeeeeefrong"
in23	q23	"honorificabilitudinitatibus"
in24	q24	"whorusalaminyourhighhohhhh"
in25	q25	"theolologicophilolological"
in26	q26	"handsomemarriedwomanrubbedagainstwidebehindinclonskeatram"

For the inverted index method the queries are structured using the terms from the bag of words concept and these queries are structured where each distinct term is separated by the logical OR operator. The 26 information needs and their related 26 queries, all having a one-to-one relationship, are presented in Table B.6.

Table B.6: Pilot 2 – IRS-I: Query per information need

In No	q	Query
in01	q01	alderman OR cowley
in02	q02	ben OR dollard
in03	q03	buck OR mulligan
in04	q04	abraham OR councillor OR lyon
in05	q05	cowley OR father

In No	q	Query
in06	q06	henry OR jimmy
in07	q07	fanning OR john
in08	q08	john OR nolan OR wyse
in09	q09	edward OR lord OR street
in10	q10	cunningham OR martin
in11	q11	douce OR miss
in12	q12	kennedy OR miss
in13	q13	boylan OR mr
in14	q14	dedalus OR mr
in15	q15	e OR m OR mr OR solomons
in16	q16	mr OR owen
in17	q17	mr OR power
in18	q18	kernan OR mr OR thomas
in19	q19	c OR hugh OR love OR reverend
in20	q20	humble OR william
in21	q21	wavyavyeavyheavyeavyevyevyhair
in22	q22	frseeeeeeeeeeeeeeeeeeeeefrong
in23	q23	honorificabilitudinitatibus
in24	q24	whorusalaminyourhighhohhhh
in25	q25	theolologicophilolological
in26	q26	handsomemarriedwomanrubbedagainstwidebehindinclonskeatram

In summary, for Pilot 2 the test collection comprised a single 666-page document Ulysses with 26 information needs, 26 queries, 26 phrase-terms using the hybrid index method, and 46 single-word terms using the inverted index method.

B.2.2 Data analysis

In this data analysis section, the list of file names within the document collection are presented followed by the token and query indices for both the inverted and hybrid indexing methods. Thereafter the top five stop words are presented. Finally, the term-by-document matrix with the computed values of term frequency for the inverted index method, and the phrase-term-by-document matrix with the computed values of phrase-term frequency for the hybrid index method are presented.

B.2.2.1 Pilot 2 results – File names

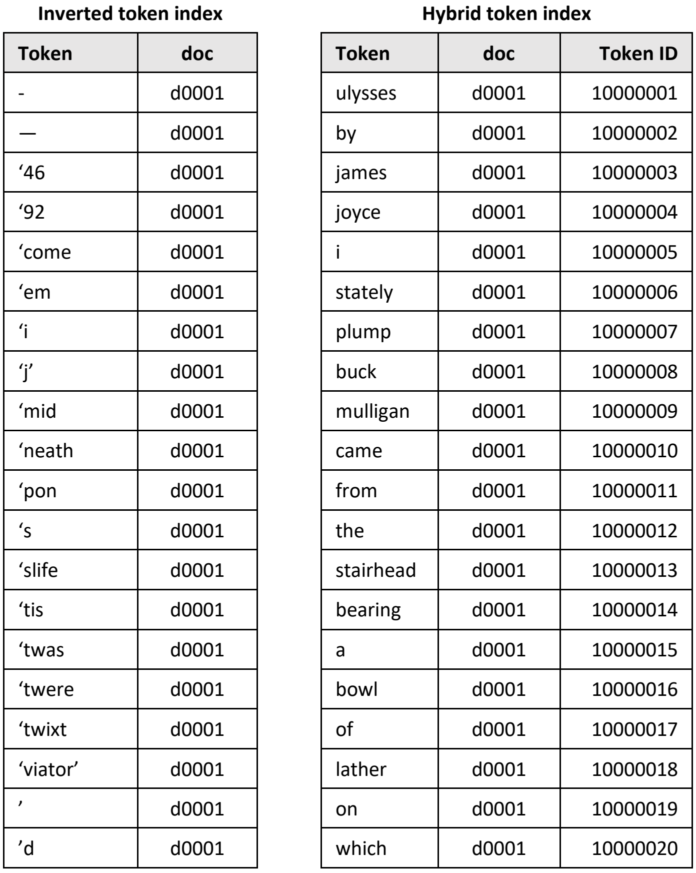
For Pilot 2 the single file name for the document collection is presented in Table B.7.

Table B.7: Pilot 2 – File names

doc	File Name	Path
d0001	Ulysses.txt	C:\Thesis\Pilot 2\

B.2.2.2 Pilot 2 results – The token indices

The inverted token index and the hybrid token index are now presented. The inverted token index contained 30,889 distinct tokens, all having a relationship with a single document d_{0001}. The hybrid token index contained 270,598 non-distinct tokens, all having a relationship with a single document *d₀₀₀₁*. In this pilot, and similar to Pilot 1, the advantage the inverted token index had over the hybrid token index was fewer records, and the advantage the hybrid token index had over the inverted token index was the addition of the unique Token ID that preserved word ordinality and proximity. Figure B.1 presents the results of the first 20 tokens in sequential order for both token indexing methods.



Inverted token index

Token	doc
-	d0001
—	d0001
‘46	d0001
‘92	d0001
‘come	d0001
‘em	d0001
‘i	d0001
‘j’	d0001
‘mid	d0001
‘neath	d0001
‘pon	d0001
‘s	d0001
‘slife	d0001
‘tis	d0001
‘twas	d0001
‘twere	d0001
‘twixt	d0001
‘viator’	d0001
’	d0001
’d	d0001

Hybrid token index

Token	doc	Token ID
ulysses	d0001	10000001
by	d0001	10000002
james	d0001	10000003
joyce	d0001	10000004
i	d0001	10000005
stately	d0001	10000006
plump	d0001	10000007
buck	d0001	10000008
mulligan	d0001	10000009
came	d0001	10000010
from	d0001	10000011
the	d0001	10000012
stairhead	d0001	10000013
bearing	d0001	10000014
a	d0001	10000015
bowl	d0001	10000016
of	d0001	10000017
lather	d0001	10000018
on	d0001	10000019
which	d0001	10000020

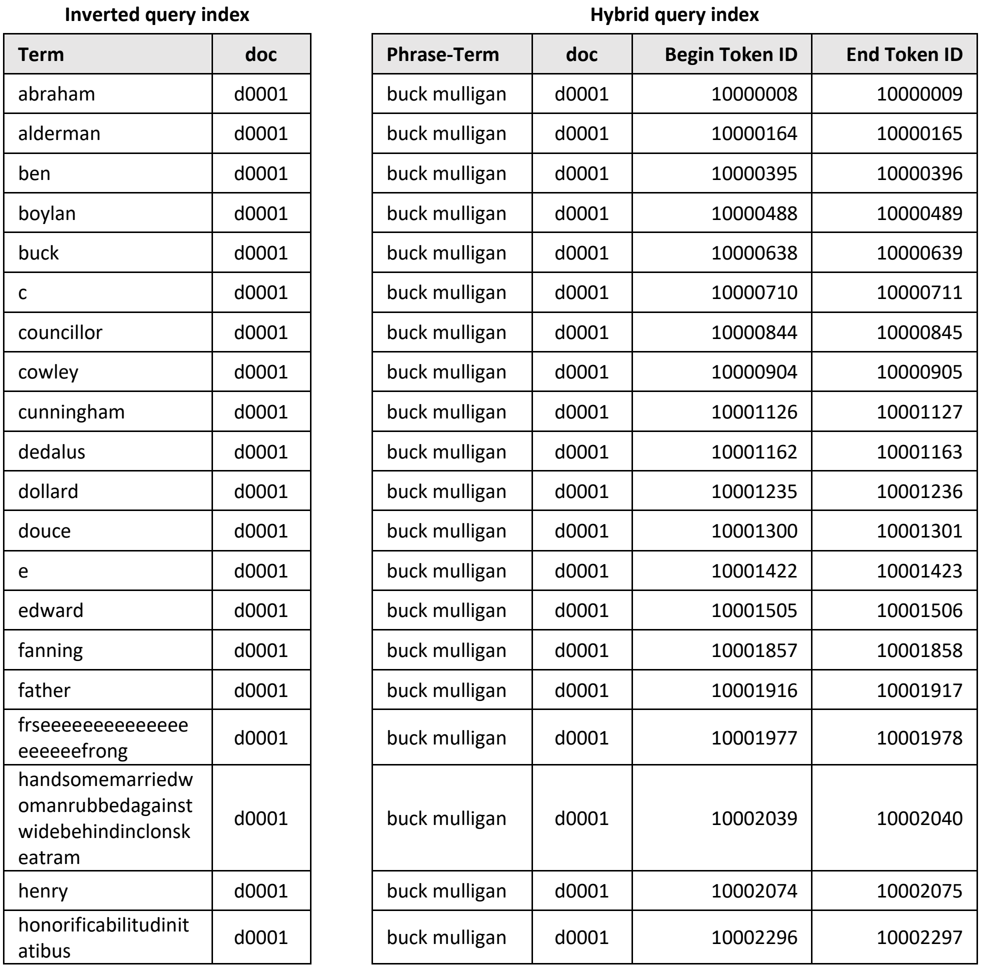
Figure B.1: Pilot 2 – The token indexes

Note that the tokens in the inverted token index are in alphabetical order while the tokens in the hybrid token index are in the same order as they appear in the text.

B.2.2.3 Pilot 2 results – The query indices

The inverted query index and the hybrid query index are now presented. From the many words used within the queries, the inverted query index contained 46 distinct terms all having a relationship with a single document *d₀₀₀₁*. The hybrid query index contained 516 non-distinct phrase-terms, all having a relationship with a single document *d₀₀₀₁*. In this pilot, the advantage the hybrid query index had over the inverted query index is the addition of the begin Token ID and end Token ID.

Figure B.2 presents the results of the first 20 records in sequential order for both query index methods.



Inverted query index

Term	doc
abraham	d0001
alderman	d0001
ben	d0001
boylan	d0001
buck	d0001
c	d0001
councillor	d0001
cowley	d0001
cunningham	d0001
dedalus	d0001
dollard	d0001
douce	d0001
e	d0001
edward	d0001
fanning	d0001
father	d0001
frseeeeeeeeeeeeeeeeeeeefrong	d0001
handsomemarriedwomanrubbedagainstwidebehindinclonskeatram	d0001
henry	d0001
honorificabilitudinitatibus	d0001

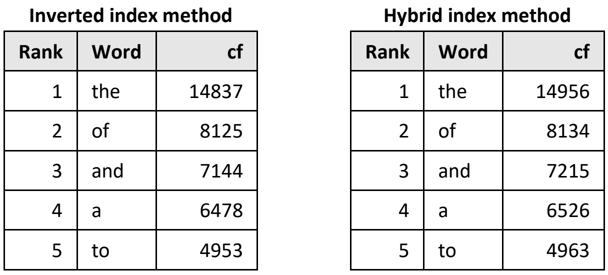
Hybrid query index

Phrase-Term	doc	Begin Token ID	End Token ID
buck mulligan	d0001	10000008	10000009
buck mulligan	d0001	10000164	10000165
buck mulligan	d0001	10000395	10000396
buck mulligan	d0001	10000488	10000489
buck mulligan	d0001	10000638	10000639
buck mulligan	d0001	10000710	10000711
buck mulligan	d0001	10000844	10000845
buck mulligan	d0001	10000904	10000905
buck mulligan	d0001	10001126	10001127
buck mulligan	d0001	10001162	10001163
buck mulligan	d0001	10001235	10001236
buck mulligan	d0001	10001300	10001301
buck mulligan	d0001	10001422	10001423
buck mulligan	d0001	10001505	10001506
buck mulligan	d0001	10001857	10001858
buck mulligan	d0001	10001916	10001917
buck mulligan	d0001	10001977	10001978
buck mulligan	d0001	10002039	10002040
buck mulligan	d0001	10002074	10002075
buck mulligan	d0001	10002296	10002297

Figure B.2: Pilot 2 – The query indices

B.2.2.4 Pilot 2 results – Stop words

For the inverted index method, of the 266,102 tokens acquired from the text, 30,889 were distinct, and for the hybrid index method, of the 270,598 tokens acquired from the text, 29,375 were distinct. Of these, the top five stop words ranked in descending order for both the inverted and hybrid index methods are provided in Figure B.3.



Inverted index method

Rank	Word	cf
1	the	14837
2	of	8125
3	and	7144
4	a	6478
5	to	4953

Hybrid index method

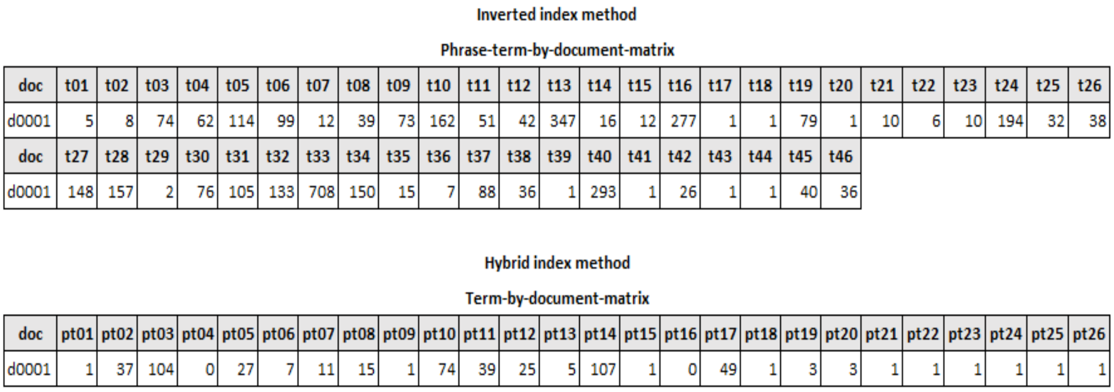
Rank	Word	cf
1	the	14956
2	of	8134
3	and	7215
4	a	6526
5	to	4963

Figure B.3: Pilot 2 – Top five stop words

The collection frequencies for the token *'the'* are therefore 14,837 and 14,956 for the inverted and hybrid indexing methods respectively. The data suggests that the token collection frequency using the hybrid indexing method is always higher than the inverted indexing method.

B.2.2.5 Term frequency, phrase-term frequency and matrices

Figure B.4 presents the term-by-document matrix and the phrase-term-by-document matrix for the inverted and hybrid indexing methods respectively.



Inverted index method

Phrase-term-by-document-matrix

doc	t01	t02	t03	t04	t05	t06	t07	t08	t09	t10	t11	t12	t13	t14	t15	t16	t17	t18	t19	t20	t21	t22	t23	t24	t25	t26
d0001	5	8	74	62	114	99	12	39	73	162	51	42	347	16	12	277	1	1	79	1	10	6	10	194	32	38

doc	t27	t28	t29	t30	t31	t32	t33	t34	t35	t36	t37	t38	t39	t40	t41	t42	t43	t44	t45	t46
d0001	148	157	2	76	105	133	708	150	15	7	88	36	1	293	1	26	1	1	40	36

Hybrid index method

Term-by-document-matrix

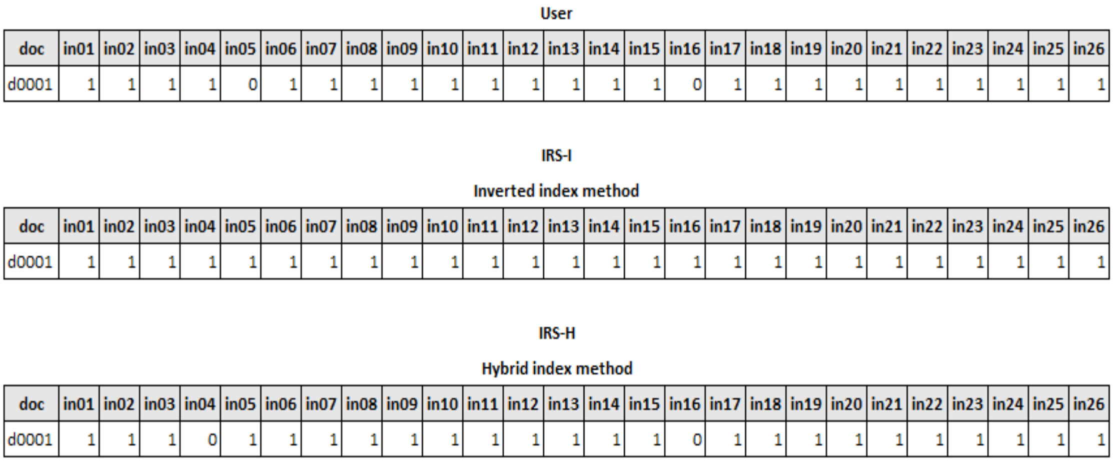
doc	pt01	pt02	pt03	pt04	pt05	pt06	pt07	pt08	pt09	pt10	pt11	pt12	pt13	pt14	pt15	pt16	pt17	pt18	pt19	pt20	pt21	pt22	pt23	pt24	pt25	pt26
d0001	1	37	104	0	27	7	11	15	1	74	39	25	5	107	1	0	49	1	3	3	1	1	1	1	1	1

Figure B.4: Pilot 2 – Matrices

Referring to the term-by-document matrix $tf_{t01,\ d0001}$ = 5 indicating that term t_{01} occurs in document d_{0001} five times, $tf_{t02,\ d0001}$ = 8 indicating that term t_{02} occurs in document d_{0001} eight times, etc. Referring to the phrase-term-by-document matrix $ptf_{pt01,\ d0001}$ = 1 indicating that phrase-term pt_{01} occurs in document d_{0001} once, $ptf_{pt02,\ d0001}$ = 37 indicating that phrase-term pt_{02} occurs in document d_{0001} 37 times, etc.

B.2.3 Performance measurements

Listed below (Figure B.5) are the results from the user's judged information-need-by-document matrix together with the information-need-by-document matrix produced by IRS-I, and finally the information-need-by-document matrix produced by IRS-H.



User

doc	in01	in02	in03	in04	in05	in06	in07	in08	in09	in10	in11	in12	in13	in14	in15	in16	in17	in18	in19	in20	in21	in22	in23	in24	in25	in26
d0001	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1

IRS-I

Inverted index method

doc	in01	in02	in03	in04	in05	in06	in07	in08	in09	in10	in11	in12	in13	in14	in15	in16	in17	in18	in19	in20	in21	in22	in23	in24	in25	in26
d0001	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

IRS-H

Hybrid index method

doc	in01	in02	in03	in04	in05	in06	in07	in08	in09	in10	in11	in12	in13	in14	in15	in16	in17	in18	in19	in20	in21	in22	in23	in24	in25	in26
d0001	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1

Figure B.5: Pilot 2 – User and IRS information-need-by-document matrices

B.3 Evaluation

These performance measurements for Pilot 2 are all listed in table form for IRS-I as well as IRS-H in Table B.8 and Table B.9 respectively.

Table B.8: Pilot 2 – IRS-I: performance measurements

IRS-I

In No	tp	fp	fn	tn	tpfp	fntn	tpfn	fptn	tpfpfntn	P	R	F
in01	1	0	0	0	1	0	1	0	1	100	100	100
in02	1	0	**0**	0	1	0	1	0	1	100	100	100
in03	1	0	0	0	1	0	1	0	1	100	100	100
in04	1	0	0	0	1	0	1	0	1	100	100	100
in05	0	1	0	0	1	0	0	1	1	0	0	0
in06	1	0	0	0	1	0	1	0	1	100	100	100
in07	1	0	0	0	1	0	1	0	1	100	100	100
in08	1	0	0	0	1	0	1	0	1	100	100	100
in09	1	0	0	0	1	0	1	0	1	100	100	100
in10	1	0	0	0	1	0	1	0	1	100	100	100
in11	1	0	0	0	1	0	1	0	1	100	100	100
in12	1	0	0	0	1	0	1	0	1	100	100	100
in13	1	0	0	0	1	0	1	0	1	100	100	100
in14	1	0	0	0	1	0	1	0	1	100	100	100
in15	1	0	0	0	1	0	1	0	1	100	100	100
in16	0	1	0	0	1	0	0	1	1	0	0	0
in17	1	0	0	0	1	0	1	0	1	100	100	100
in18	1	0	0	0	1	0	1	0	1	100	100	100
in19	1	0	0	0	1	0	1	0	1	100	100	100
in20	1	0	0	0	1	0	1	0	1	100	100	100
in21	1	0	0	0	1	0	1	0	1	100	100	100
in22	1	0	0	0	1	0	1	0	1	100	100	100
in23	1	0	0	0	1	0	1	0	1	100	100	100
in24	1	0	0	0	1	0	1	0	1	100	100	100
in25	1	0	0	0	1	0	1	0	1	100	100	100
in26	1	0	0	0	1	0	1	0	1	100	100	100

Table B.9: Pilot 2 – IRS-H: performance measurements

IRS-H

In No	tp	fp	fn	tn	tpfp	fntn	tpfn	fptn	tpfpfntn	P	R	F
in01	1	0	0	0	1	0	1	0	1	100	100	100
in02	1	0	0	0	1	0	1	0	1	100	100	100
in03	1	0	0	0	1	0	1	0	1	100	100	100
in04	0	0	1	0	0	1	1	0	1	0	0	0
in05	0	1	0	0	1	0	0	1	1	0	0	0
in06	1	0	0	0	1	0	1	0	1	100	100	100
in07	1	0	0	0	1	0	1	0	1	100	100	100
in08	1	0	0	0	1	0	1	0	1	100	100	100
in09	1	0	0	0	1	0	1	0	1	100	100	100
in10	1	0	0	0	1	0	1	0	1	100	100	100
in11	1	0	0	0	1	0	1	0	1	100	100	100
in12	1	0	0	0	1	0	1	0	1	100	100	100
in13	1	0	0	0	1	0	1	0	1	100	100	100
in14	1	0	0	0	1	0	1	0	1	100	100	100
in15	1	0	0	0	1	0	1	0	1	100	100	100
in16	0	0	0	1	0	1	0	1	1	0	0	0

IRS-H

In No	tp	fp	fn	tn	tpfp	fntn	tpfn	fptn	tpfpfntn	P	R	F
in17	1	0	0	0	1	0	1	0	1	100	100	100
in18	1	0	0	0	1	0	1	0	1	100	100	100
in19	1	0	0	0	1	0	1	0	1	100	100	100
in20	1	0	0	0	1	0	1	0	1	100	100	100
in21	1	0	0	0	1	0	1	0	1	100	100	100
in22	1	0	0	0	1	0	1	0	1	100	100	100
in23	1	0	0	0	1	0	1	0	1	100	100	100
in24	1	0	0	0	1	0	1	0	1	100	100	100
in25	1	0	0	0	1	0	1	0	1	100	100	100
in26	1	0	0	0	1	0	1	0	1	100	100	100

In Figure B.6, the 26 information needs for Pilot 2 are presented for each of the IRSs. The diamonds represent the F-measure values for IRS-I, and the squares represent the F-measure values for IRS-H.

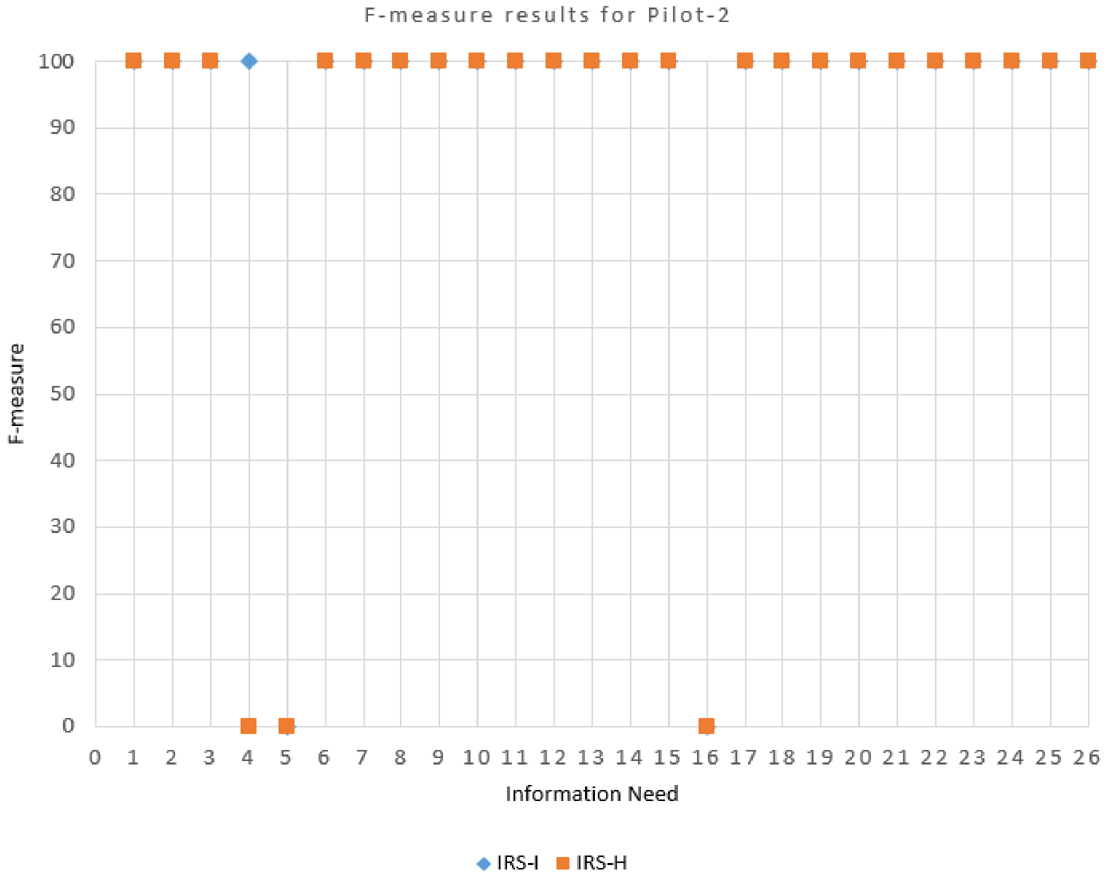


Figure B.6: Pilot 2 – IRS-I and IRS-H performance measurements

Referring to the performance measurement graph for Pilot 2 in Figure B.6, the 26 information needs were defined by the user, and after perusing the single document in the collection, the user judged the document was relevant to 24 of the 26 information needs, *in*$_{05}$ and *in*$_{16}$ were judged non-relevant. Thus, there are two exceptions related to relevancy and the third exception is *in*$_{04}$ where the two IRSs disagreed with each other.

i) in_{04} – the phrase-term ‘councillor abraham lyon’ existed in the text but the OCR erred in converting the token ‘councillor’ as ‘councillo r’. As there was only one occurrence of ‘councillor’, as part of a phrase together with ‘abraham lyon’ in the text, IRS-H did not retrieve the document, but as it was judged relevant by the user, *fn* was set to 1. However as the phrase-term ‘abraham leon’ existed in the text and as ‘councillor’ appeared elsewhere in the text, IRS-I did retrieve the document using the Boolean OR operator, thus setting tp to 1.

ii) in_{05} – the phrase-term ‘father cowley’ existed in the text but the user judged this information need as non-relevant, setting *fp* to 1 for both IRSs.

iii) in_{16} – the phrase-term ‘mr owen’ did not exist in the text and as the user judged this information need as non-relevant for IRS-H, tn was set to 1. However, as the terms ‘mr’ and ‘owen’ existed individually in the document, IRS-I set *fp* to 1. In this case, IRS-I retrieves the document as it believes it is relevant when it is not, thus creating unnecessary additional reading for the user to determine its relevancy for the information need. In this research, it is these *fps* that need to be reduced, thus saving the researcher’s time in perusing the documents within the collection.

All other information needs had a tp set to 1 i.e. in_{05} and in_{16} using IRS-I and in_{04}, in_{05} and in_{16} using IRS-H.

B.4 Summary

At this stage at the end of Pilot 2, there was evidence to suggest that the functionality of IRS-H was more effective than IRS-I but this needed further investigation, testing, and evaluation with input from participating users. Both IRSs have their merits and after evaluation, any differentiations between the two are supported by theory and the data. The evidence in Pilot 2 does not suggest any differentiation in handling vocabulary mismatch as this pilot made use of non-expanded queries, hence the reasoning for the final Pilot 3 that specifically addresses the topic of vocabulary mismatch.

The design findings for Pilot 2 are summarised in Table B.10.

Table B.10: Pilot 2 – Summary of design findings

Pilot	Stage	Finding
2	Information gathering	Pilot 2 was based on the book Ulysses written by James Joyce (Joyce, 1932).
2	Information gathering	Content acquisition: document length was increased by altering the content to a single 666-page document, the book Ulysses. The content was acquired from the pdf document and converted to text successfully. However, on a few occasions the text was converted incorrectly by the OCR software.
2	Information gathering	Hybrid token index: the number ranges for the document number and the token ID were expanded as these were limiting factors in Pilot 1. In addition, the token field in the index was expanded to accommodate larger sized tokens. For example, the token ‘handsomemarriedwomanrubbedagainstwidebehindinclonskeatram’
2	Search engine	Phrase-term: 26 phrase-terms provided by the user (the researcher in this pilot) were presented correctly, all in lowercase without special characters.
2	Search engine	Phrase-term query: these phrase-terms were expressed as 26 queries, six of which used single word phrase-terms.

Pilot	Stage	Finding
2	Search engine	Hybrid query index: the 26 queries were presented to the hybrid query index, which was thereafter populated with the phrase-terms, and unique begin and end token IDs.
2	Search engine	The hybrid query index interrogated the hybrid token index successfully and where a match was found (a phrase-term existed in a document) the document number was returned and the hybrid query index updated with the document number accordingly.
2	Design	Phrase-term frequency (ptf) was maintained.
2	Design	Converting ptf values to binary and the population of the phrase-term-by-document matrix with these values remained successful.
2	Design	The IRS was able to match phrase-terms expressed in queries to those in documents exactly.
2	Design	Performance measurements remained unusable, as the judgment results from the user were unavailable and therefore not tested.
2	Design	At this stage at the end of Pilot 2, there was evidence to suggest that the functionality of IRS-H was more effective than IRS-I but this needed further investigation, testing, and evaluation with input from participating users.

Reflecting back to Figure 3.10 of Hevner (2007:2) and Figure 3.11 that explains the three design cycles for this research in Chapter Three, the design cycle is now complete for Pilot 2. From the measurements and the IRS evaluation, any design and build issues evidenced were now, through the design cycle, driven to perform Pilot 3.

APPENDIX C: PILOT 3 VOCABULARY MISMATCH

C.1 Design and test objectives

From a design perspective and using the results from Pilot 2, it became apparent that four additional objectives were needed in order to test the two IRSs rigorously – two design objectives and two test objectives.

The first objective for Pilot 3 was to re-design the length of the token field in the token index and the query index for both IRSs. The length was originally set at 40 characters, not expecting tokens acquired from the text, or words in the English language, to be longer than that. The field had to be increased to 60 characters to accommodate the 57 character token '*handsomemarriedwomanrubbedagainstwidebehindinclonskeatram'* from the text of Ulysses. Because these token indices increase in size depending on the number of tokens that exist within the document collection, then for design purposes, it is best to keep the token field as small as possible. However, in this research, 60 characters became an acceptable size from a design perspective.

The second objective was to re-design the method of handling special characters existing within the text, during the text acquisition stage that occurs during the information gathering process. According to the original design in Pilot 1 and Pilot 2, unwanted special characters were rejected using special routines, but to specify, all unwanted characters became a challenge, as many were easily missed and/or not catered for, or the Windows operating system could not identify them. Therefore, the converse was applied where only letters and numbers were accepted and the routine was updated accordingly for the hybrid indexing method. Similarly, for the inverted indexing method, the same conditions were applied but hyphens and apostrophes were preserved.

The third objective was to use a mix of single phrase-term queries and expanded queries and to test this mix. The expanded queries were required in order to try to retrieve documents that contained a number of differing phrases that described the same, or similar, concepts.

The fourth objective was to test for plurals, the use of synonyms and antonyms, and to differentiate between different versions of English language spelling (British English versus US English).

C.2 Comparative evaluation and results

In this empirically comparative evaluation and results section, the preparation of the test collection is presented, followed by the data analysis and finally the performance measurements for Pilot 3. Similar to Pilot 2, the final evaluation compares the results of IRS-I to those of IRS-H.

C.2.1 Test collection preparation

After the design and build of the IRSs, the test collection was prepared to evaluate Pilot 3 rigorously. The document collection was collated by selecting 20 documents; therefore, document collection N equalled 20. For Pilot 3, 14 information needs were compiled by the user, all referring to the problem of vocabulary mismatch.

Here the tests included differentiating in English language spelling, plurals, expanded queries, and the use of synonyms and antonyms. Table C.1 presents the information needs that were provided by the user (this researcher) including the relevancy judgment (document d_{0001} is used as an example). A tick represents 'relevant' while a cross represents non-relevant.

Table C.1: Pilot 3 – User judged information needs

Document number - d0001		
Please indicate whether this document is relevant to any of the following information needs (please tick)		
In No	**Information Need**	**Relevant**
in01	I want to find all documents relevant to term mismatch	√
in02	I want to find all documents relevant to vocabulary agreement	√
in03	I want to find all documents relevant to vocabulary gap	√
in04	I want to find all documents relevant to vocabulary limitation	x
in05	I want to find all documents relevant to vocabulary mismatch	√
in06	I want to find all documents relevant to vocabulary normalisation	√
in07	I want to find all documents relevant to vocabulary problem	√
in08	I want to find all documents relevant to vocabulary mismatch problem	√
in09	I want to find all documents relevant to vocabulary mismatch phrases	√
in10	I want to find all documents relevant to vocabulary mismatch remains a problem	√
in11	I want to find all documents relevant to vocabulary mismatch is still a problem	√
in12	I want to find all documents relevant to vocabulary mismatch remains unresolved	√
in13	I want to find all documents relevant to vocabulary mismatch is still unresolved	√
in14	I want to find all documents relevant to vocabulary mismatch where it remains a problem	√

According to the user, only one information need was judged non-relevant, $in_{04,}$ representing the information need *"I want to find all documents relevant to vocabulary limitation".*

The next activity in preparing the test collection was selecting the phrase-terms to be used in the hybrid indexing method queries and selecting the terms to be used in the inverted indexing method queries. To evaluate the document collection for Pilot 3 using the hybrid indexing method, 14 multi-word phrase-terms were used in the queries to express the 14 information needs. Each information need had one or more phrase-terms allocated to it.

Each phrase-term is presented in Table C.2 and its associated information need is presented in Table C.3.

Table C.2: Pilot 3 – IRS-H: Phrase-terms

pt	Phrase
pt01	term mismatch
pt02	vocabulary agreement
pt03	vocabulary gap
pt04	vocabulary limitation
pt05	vocabulary limitations
pt06	vocabulary mismatch
pt07	vocabulary mismatch is still a problem
pt08	vocabulary mismatch is still unresolved
pt09	vocabulary mismatch problem
pt10	vocabulary mismatch remains a problem
pt11	vocabulary mismatch remains unresolved
pt12	vocabulary normalisation
pt13	vocabulary normalization
pt14	vocabulary problem

Table C.3: Pilot 3 – IRS-H: Phrase-terms per information need

In No	pt	Phrase-term
in01	pt01	term mismatch
in02	pt02	vocabulary agreement
in03	pt03	vocabulary gap
in04	pt04	vocabulary limitation
in04	pt05	vocabulary limitations
in05	pt06	vocabulary mismatch
in06	pt13	vocabulary normalization
in06	pt12	vocabulary normalisation
in07	pt14	vocabulary problem
in08	pt09	vocabulary mismatch problem
in09	pt02	vocabulary agreement
in09	pt03	vocabulary gap
in09	pt04	vocabulary limitation
in09	pt05	vocabulary limitations
in09	pt06	vocabulary mismatch
in09	pt13	vocabulary normalization
in09	pt12	vocabulary normalisation
in09	pt14	vocabulary problem
in09	pt09	vocabulary mismatch problem
in10	pt10	vocabulary mismatch remains a problem
in11	pt07	vocabulary mismatch is still a problem
in12	pt11	vocabulary mismatch remains unresolved
in13	pt08	vocabulary mismatch is still unresolved
in14	pt10	vocabulary mismatch remains a problem
in14	pt07	vocabulary mismatch is still a problem
in14	pt11	vocabulary mismatch remains unresolved
in14	pt08	vocabulary mismatch is still unresolved

To evaluate the document collection for Pilot 3 using the inverted indexing method, 15 single-word terms were used in the queries to express the 14 information needs. Each information need had one or more terms allocated to it. Each term is presented in Table C.4 and each term and its associated information need is presented in Table C.5.

Table C.4: Pilot 3 – IRS-I: Terms

t	Term
t01	a
t02	agreement
t03	gap
t04	is
t05	limitation
t06	limitations
t07	mismatch
t08	normalisation
t09	normalization
t10	problem
t11	remains
t12	still
t13	term
t14	unresolved
t15	vocabulary

Table C.5: Pilot 3 – IRS-I: Terms per information need

In No	t	Term
in01	t07	mismatch
in01	t13	term
in02	t02	agreement
in02	t15	vocabulary
in03	t03	gap
in03	t15	vocabulary
in04	t05	limitation
in04	t06	limitations
in04	t15	vocabulary
in05	t07	mismatch
in05	t15	vocabulary
in06	t08	normalisation
in06	t09	normalization
in06	t15	vocabulary
in07	t10	problem
in07	t15	vocabulary
in08	t07	mismatch
in08	t10	problem
in08	t15	vocabulary
in09	t02	agreement
in09	t03	gap
in09	t05	limitation
in09	t06	limitations
in09	t07	mismatch
in09	t08	normalisation
in09	t09	normalization
in09	t10	problem
in09	t15	vocabulary
in10	t01	a
in10	t07	mismatch
in10	t10	problem
in10	t11	remains
in10	t15	vocabulary
in11	t01	a
in11	t04	is
in11	t07	mismatch
in11	t10	problem
in11	t12	still
in11	t15	vocabulary
in12	t07	mismatch
in12	t11	remains
in12	t14	unresolved
in12	t15	vocabulary
in13	t04	is
in13	t07	mismatch
in13	t12	still
in13	t14	unresolved
in13	t15	vocabulary
in14	t01	a
in14	t04	is
in14	t07	mismatch
in14	t10	problem
in14	t11	remains
in14	t12	still
in14	t14	unresolved
in14	t15	vocabulary

The next activity was to present the queries that express each information need to the search engine using both indexing methods. For the hybrid index method, the queries were structured as phrase-terms. As this pilot uses multiple phrase-terms, to expand the queries, it was necessary to use the logical OR operator in the queries. The 14 information needs and their related 14 queries all having a one-to-one relationships are presented in Table C.6.

Table C.6: Pilot 3 – IRS-H: Query per information need

In No	q	Query
in01	q01	"term mismatch"
in02	q02	"vocabulary agreement"
in03	q03	"vocabulary gap"
in04	q04	"vocabulary limitation" OR "vocabulary limitations"
in05	q05	"vocabulary mismatch"
in06	q06	"vocabulary normalization" OR "vocabulary normalisation"
in07	q07	"vocabulary problem"
in08	q08	"vocabulary mismatch problem"
in09	q09	"vocabulary agreement" OR "vocabulary gap" OR "vocabulary limitation" OR "vocabulary limitations" OR "vocabulary mismatch" OR "vocabulary normalization" OR "vocabulary normalisation" OR "vocabulary problem" OR "vocabulary mismatch problem"
in10	q10	"vocabulary mismatch remains a problem"
in11	q11	"vocabulary mismatch is still a problem"
in12	q12	"vocabulary mismatch remains unresolved"
in13	q13	"vocabulary mismatch is still unresolved"
in14	q14	"vocabulary mismatch remains a problem" OR "vocabulary mismatch is still a problem" OR "vocabulary mismatch remains unresolved" OR "vocabulary mismatch is still unresolved"

For the inverted index method, the queries were structured using the terms from the bag of words concept and these queries were structured where each distinct term was separated by the logical OR operator. The 14 information needs and their related 14 queries all having a one-to-one relationships are presented in Table C.7.

Table C.7: Pilot 3 – IRS-I: Query per information need

In No	q	Query
in01	q01	mismatch OR term
in02	q02	agreement OR vocabulary
in03	q03	gap OR vocabulary
in04	q04	limitation OR limitations OR vocabulary
in05	q05	mismatch OR vocabulary
in06	q06	normalisation OR normalization OR vocabulary
in07	q07	problem OR vocabulary
in08	q08	mismatch OR problem OR vocabulary
in09	q09	agreement OR gap OR limitation OR limitations OR mismatch OR normalisation OR normalization OR problem OR vocabulary
in10	q10	a OR mismatch OR problem OR remains OR vocabulary
in11	q11	a OR is OR mismatch OR problem OR still OR vocabulary
in12	q12	mismatch OR remains OR unresolved OR vocabulary
in13	q13	is OR mismatch OR still OR unresolved OR vocabulary
in14	q14	a OR is OR mismatch OR problem OR remains OR still OR unresolved OR vocabulary

In summary, for Pilot 3 the test collection comprised a collection of 20 documents with 14 information needs, 14 queries, 14 phrase-terms using the hybrid index method, and 15 single-word terms using the inverted index method.

C.2.2 Data analysis

In this data analysis section, the list of file names within the document collection is presented, followed by the token and query indices for both the inverted and hybrid indexing methods. Thereafter the top five stop words are presented. Finally, the term-by-document matrix with the computed values of term frequency for the inverted index method, and the phrase-term-by-document matrix with the computed values of phrase-term frequency for the hybrid index method are presented.

C.2.2.1 Pilot 3 results – File names

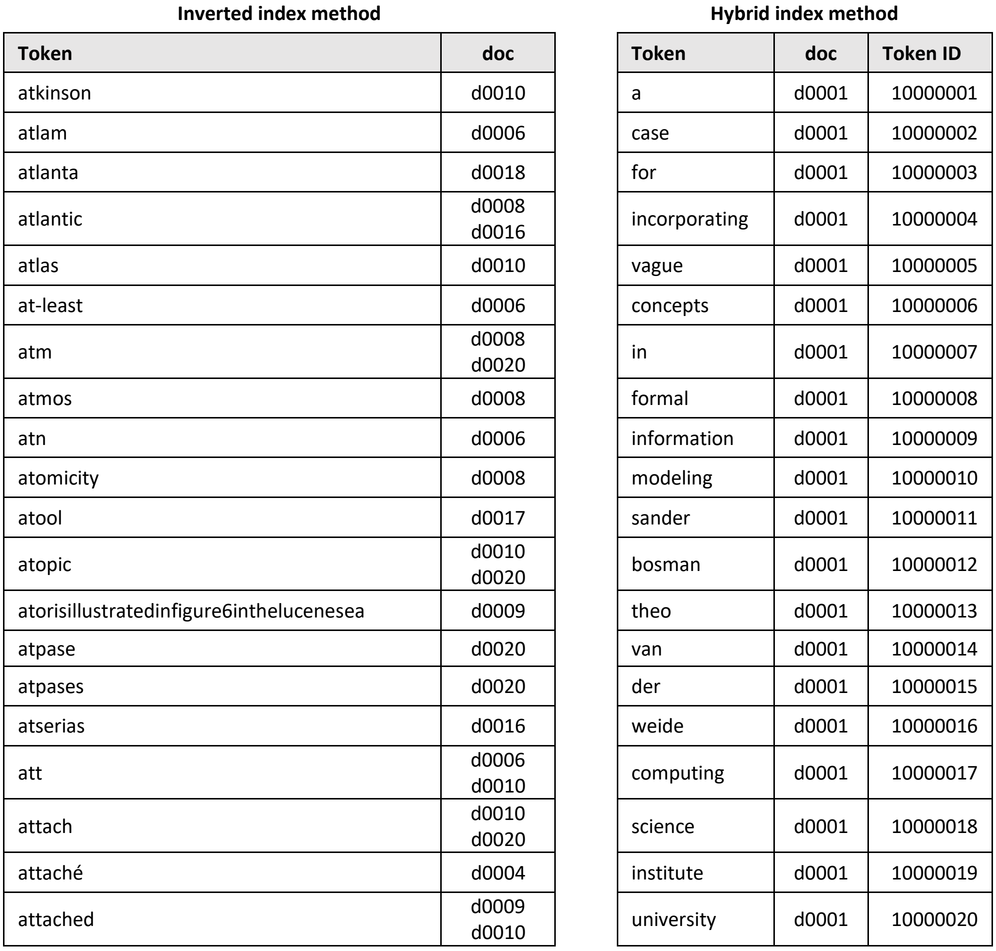
For Pilot 3, the 20 documents in the collection are presented in Table C.8.

Table C.8: Pilot 3 – File names

doc	File Name	Path
d0001	A case for incorporating vague concepts in formal information modeling.txt	C:\Thesis\Pilot 3\
d0002	A coefficient of agreement for nominal scales (Cohen 1960 Kappa).txt	C:\Thesis\Pilot 3\
d0003	A communication perspective on the international information and knowledge system.txt	C:\Thesis\Pilot 3\
d0004	A comparative analysis of critical issues facing Canadian information systems personnel - a national and global perspective.txt	C:\Thesis\Pilot 3\
d0005	A national survey of physician industry relationships.txt	C:\Thesis\Pilot 3\
d0006	A novel neighborhood based document smoothing model for information retrieval_art_10.1007_s10791-012-9202-3.txt	C:\Thesis\Pilot 3\
d0007	A Porters Five Forces Approach to the Australian Private Hospital Industry.txt	C:\Thesis\Pilot 3\
d0008	A Practical Guide to Big Data.txt	C:\Thesis\Pilot 3\
d0009	Augmenting and Structuring User Queries to Support Efficient Free-Form Code Search (2015).txt	C:\Thesis\Pilot 3\
d0010	Autoantibodies related to type 1 diabetes in children(T130).txt	C:\Thesis\Pilot 3\
d0011	Automated mapping of clinical terms into SNOMED-CT. An application to codify procedures in pathology.txt	C:\Thesis\Pilot 3\
d0012	Automatic term mismatch diagnosis for selective query expansion (Zhao 2012).txt	C:\Thesis\Pilot 3\
d0013	Combining evidence for Web retrieval using the inference network model an experimental study (2004).txt	C:\Thesis\Pilot 3\
d0014	Combining Grounded Theory and Case Study Methods in IT Outsourcing Study.txt	C:\Thesis\Pilot 3\
d0015	Discovering Latent Topical Structure by Second-Order Similarity Analysis.txt	C:\Thesis\Pilot 3\
d0016	Expansion for information retrieval contribution of word sense disambiguation and semantic relatedness (PhD 2011).txt	C:\Thesis\Pilot 3\
d0017	Exploring criteria for successful query expansion in the genomic domain.txt	C:\Thesis\Pilot 3\
d0018	Mining document, concept, and term associations for effective biomedical retrieval introducing MeSH-enhanced retrieval models_art_10.1007_s10791-015-9264-0.txt	C:\Thesis\Pilot 3\
d0019	On the Vocabulary Agreement in Software Issue Descriptions (2016).txt	C:\Thesis\Pilot 3\
d0020	Proof of concept - Concept-based biomedical information retrieval.txt	C:\Thesis\Pilot 3\

C.2.2.2 Pilot 3 results – The token indices

The inverted token index and the hybrid token index are now presented. The inverted token index contained 22,152 distinct tokens. The hybrid token index contained 336,514 non-distinct tokens. In this pilot, and similar to Pilot 2, the advantage that the inverted token index had over the hybrid token index was fewer records, and the advantage that the hybrid token index had over the inverted token index was the addition of the unique Token ID that preserved word ordinality and proximity. Figure C.1 presents the results of randomly selected 20 tokens in sequential order for the inverted indexing method and the first 20 tokens for the hybrid indexing method.



Inverted index method

Token	doc
atkinson	d0010
atlam	d0006
atlanta	d0018
atlantic	d0008 d0016
atlas	d0010
at-least	d0006
atm	d0008 d0020
atmos	d0008
atn	d0006
atomicity	d0008
atool	d0017
atopic	d0010 d0020
atorisillustratedinfigure6inthelucenesea	d0009
atpase	d0020
atpases	d0020
atserias	d0016
att	d0006 d0010
attach	d0010 d0020
attaché	d0004
attached	d0009 d0010

Hybrid index method

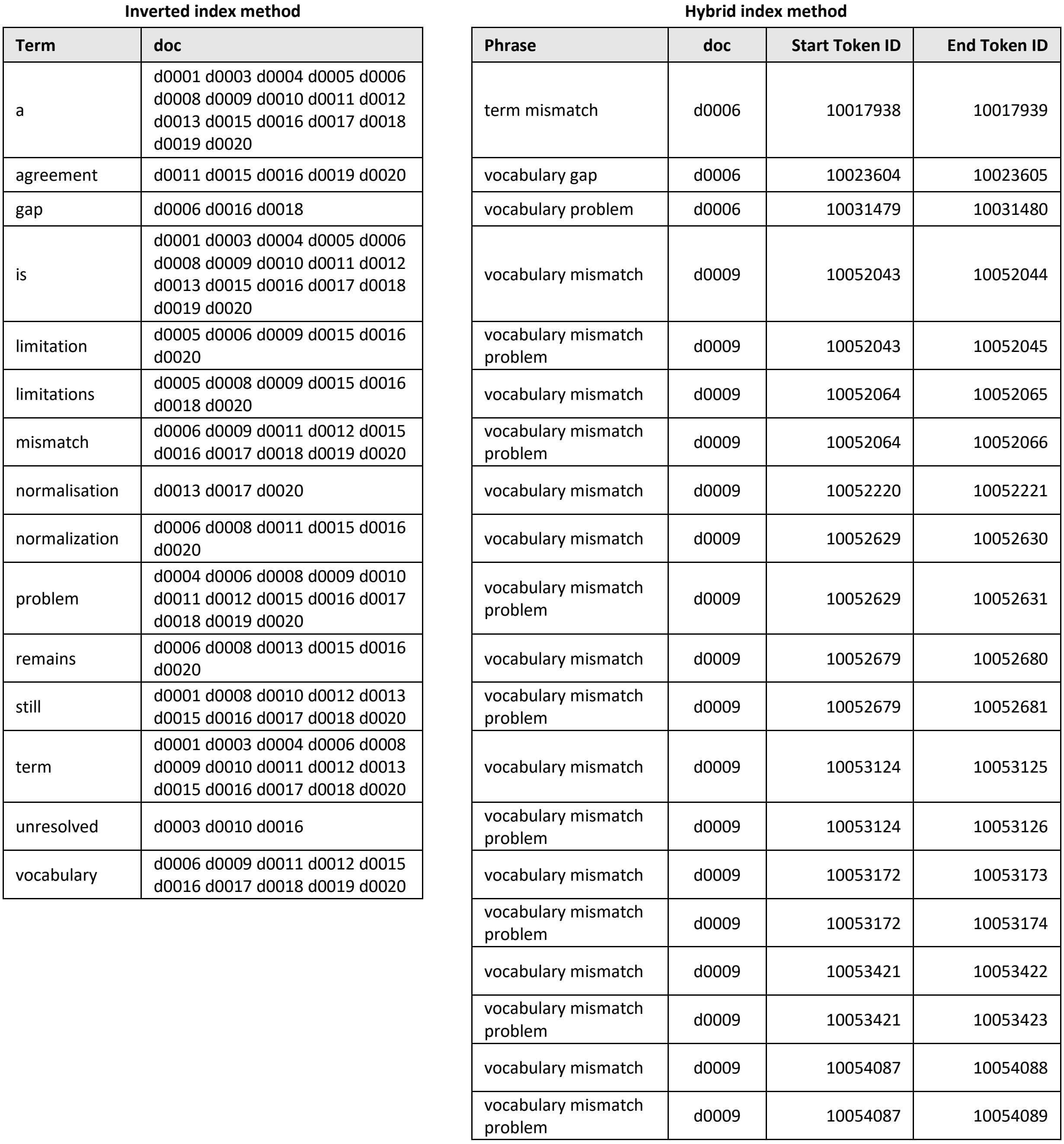
Token	doc	Token ID
a	d0001	10000001
case	d0001	10000002
for	d0001	10000003
incorporating	d0001	10000004
vague	d0001	10000005
concepts	d0001	10000006
in	d0001	10000007
formal	d0001	10000008
information	d0001	10000009
modeling	d0001	10000010
sander	d0001	10000011
bosman	d0001	10000012
theo	d0001	10000013
van	d0001	10000014
der	d0001	10000015
weide	d0001	10000016
computing	d0001	10000017
science	d0001	10000018
institute	d0001	10000019
university	d0001	10000020

Figure C.1: Pilot 3 – The token indices

Note that the tokens in the inverted token index are in alphabetical order while the tokens in the hybrid token index are in the same order as they appeared in the text. Referring to the inverted token index, a few tokens were acquired from the text in unexpected formats, for example, the token *'atorisillustratedinfigure6inthelucenesea'* from document d_{0009} where the space between words was omitted by the OCR conversion software. Token *'atool'* from document d_{0017} was acquired correctly as it formed part of an URL in the text and *'atpase'* document d_{0020} was also acquired correctly as it was originally *'ATPase'*. Although words are expected to be acquired from text, all tokens are extracted and these are best described as chunks of data. Referring to the hybrid token index, the token *'modeling'* from document d_{0001} uses US English rather than British English spelling and the token *'bosman'* from document d_{0001} is a person's surname in the Afrikaans language.

C.2.2.3 Pilot 3 results – The query indices

The inverted query index and the hybrid query index are now presented. From the many words used within the queries, the inverted query index contained 15 distinct terms. The hybrid query index contained 137 non-distinct phrase-terms. In this pilot, the advantages the hybrid query index had over the inverted query index were again fewer records and the addition of the begin Token ID and end Token ID. Figure C.2 presents the results of the 15 records using the inverted index method and the first 20 records in sequential order for the hybrid indexing methods.



Inverted index method

Term	doc
a	d0001 d0003 d0004 d0005 d0006 d0008 d0009 d0010 d0011 d0012 d0013 d0015 d0016 d0017 d0018 d0019 d0020
agreement	d0011 d0015 d0016 d0019 d0020
gap	d0006 d0016 d0018
is	d0001 d0003 d0004 d0005 d0006 d0008 d0009 d0010 d0011 d0012 d0013 d0015 d0016 d0017 d0018 d0019 d0020
limitation	d0005 d0006 d0009 d0015 d0016 d0020
limitations	d0005 d0008 d0009 d0015 d0016 d0018 d0020
mismatch	d0006 d0009 d0011 d0012 d0015 d0016 d0017 d0018 d0019 d0020
normalisation	d0013 d0017 d0020
normalization	d0006 d0008 d0011 d0015 d0016 d0020
problem	d0004 d0006 d0008 d0009 d0010 d0011 d0012 d0015 d0016 d0017 d0018 d0019 d0020
remains	d0006 d0008 d0013 d0015 d0016 d0020
still	d0001 d0008 d0010 d0012 d0013 d0015 d0016 d0017 d0018 d0020
term	d0001 d0003 d0004 d0006 d0008 d0009 d0010 d0011 d0012 d0013 d0015 d0016 d0017 d0018 d0020
unresolved	d0003 d0010 d0016
vocabulary	d0006 d0009 d0011 d0012 d0015 d0016 d0017 d0018 d0019 d0020

Hybrid index method

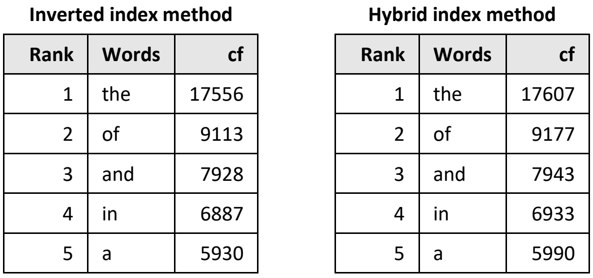
Phrase	doc	Start Token ID	End Token ID
term mismatch	d0006	10017938	10017939
vocabulary gap	d0006	10023604	10023605
vocabulary problem	d0006	10031479	10031480
vocabulary mismatch	d0009	10052043	10052044
vocabulary mismatch problem	d0009	10052043	10052045
vocabulary mismatch	d0009	10052064	10052065
vocabulary mismatch problem	d0009	10052064	10052066
vocabulary mismatch	d0009	10052220	10052221
vocabulary mismatch	d0009	10052629	10052630
vocabulary mismatch problem	d0009	10052629	10052631
vocabulary mismatch	d0009	10052679	10052680
vocabulary mismatch problem	d0009	10052679	10052681
vocabulary mismatch	d0009	10053124	10053125
vocabulary mismatch problem	d0009	10053124	10053126
vocabulary mismatch	d0009	10053172	10053173
vocabulary mismatch problem	d0009	10053172	10053174
vocabulary mismatch	d0009	10053421	10053422
vocabulary mismatch problem	d0009	10053421	10053423
vocabulary mismatch	d0009	10054087	10054088
vocabulary mismatch problem	d0009	10054087	10054089

Figure C.2: Pilot 3 – The query indices

Note that the inverted query index only refers to 17 documents for the term *'a'*. One would have expected this stop word *'a'* appearing in all 20 documents within the collection. After further investigation it was discovered that two documents, d_{0002} and d_{0007}, did not OCR convert from pdf to text correctly, the files contained no text, and a third document, d_{0014}, had only a small fraction of the text converted. As a result of identifying these unusable files it was deemed beneficial during the full evaluation to install a verification process to identify zero length text files to effectively pre-validate the document collection before evaluation commenced.

C.2.2.4 Pilot 3 results – Stop words

For the inverted index method, of the 329,719 tokens acquired from the text, 22,152 were distinct, and for the hybrid index method, of the 336,514 tokens acquired from the text, 20,005 were distinct. Of these, the top five stop words ranked in descending order for both the inverted and hybrid index methods are presented in Figure C.3.



Inverted index method

Rank	Words	cf
1	the	17556
2	of	9113
3	and	7928
4	in	6887
5	a	5930

Hybrid index method

Rank	Words	cf
1	the	17607
2	of	9177
3	and	7943
4	in	6933
5	a	5990

Figure C.3: Pilot 3 – Top five stop words

The collection frequencies for the token *'the'* are therefore 17,556 and 17,607 for the inverted and hybrid indexing methods respectively. The top five ranked stop words are identical with differing collection frequencies owing to the differing data transformation processes.

C.2.2.5 Term frequency, phrase-term frequency and matrices

Table C.9 presents the term-by-document matrix and Table C.10 the phrase-term-by-document matrix for the inverted and hybrid indexing methods respectively.

Table C.9: Pilot 3 – IRS-I: Term-by-document matrix

doc	t01	t02	t03	t04	t05	t06	t07	t08	t09	t10	t11	t12	t13	t14	t15
d0001	77	0	0	44	0	0	0	0	0	0	0	1	2	0	0
d0002	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0003	24	0	0	10	0	0	0	0	0	0	0	0	1	1	0
d0004	90	0	0	4	0	0	0	0	0	1	0	0	2	0	0
d0005	59	0	0	4	1	2	0	0	0	0	0	0	0	0	0
d0006	327	0	1	196	1	0	1	0	4	8	4	0	134	0	2
d0007	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0008	400	0	0	167	0	1	0	0	3	2	1	3	4	0	0
d0009	419	0	0	141	2	2	20	0	0	22	0	0	3	0	21
d0010	696	0	0	258	0	0	0	0	0	3	0	9	2	1	0
d0011	142	1	0	82	0	0	4	0	1	2	0	0	51	0	2
d0012	135	0	0	134	0	0	40	0	0	17	0	14	126	0	6
d0013	124	0	0	155	0	0	0	2	0	0	1	6	8	0	0
d0014	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

doc	t01	t02	t03	t04	t05	t06	t07	t08	t09	t10	t11	t12	t13	t14	t15
d0015	267	1	0	238	2	1	6	0	1	10	7	13	15	0	11
d0016	688	1	11	463	2	3	11	0	2	23	2	7	48	1	15
d0017	282	0	0	201	0	0	2	15	0	13	0	2	92	0	9
d0018	168	0	5	204	0	2	3	0	0	6	0	4	42	0	14
d0019	55	66	0	66	0	0	1	0	0	9	0	0	0	0	90
d0020	1977	4	0	930	2	18	8	47	3	27	4	26	195	0	129

Table C.10: Pilot 3 – IRS-H: Phrase-term-by-document matrix

doc	pt01	pt02	pt03	pt04	pt05	pt06	pt07	pt08	pt09	pt10	pt11	pt12	pt13	pt14
d0001	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0002	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0003	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0004	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0005	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0006	1	0	1	0	0	0	0	0	0	0	0	0	0	1
d0007	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0008	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0009	1	0	0	0	0	17	0	0	16	0	0	0	0	1
d0010	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0011	0	0	0	0	0	2	0	0	0	0	0	0	0	0
d0012	16	0	0	0	0	4	0	0	0	0	0	0	0	1
d0013	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0014	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0015	0	0	0	0	0	6	0	0	2	0	0	0	0	1
d0016	0	0	0	0	0	10	0	0	6	0	0	0	0	1
d0017	0	0	0	0	0	2	0	0	1	0	0	0	0	1
d0018	0	0	1	0	0	2	0	0	0	0	0	0	0	3
d0019	0	27	0	0	0	1	0	0	1	0	0	0	0	1
d0020	0	0	0	0	0	6	0	0	3	0	0	0	0	1

Referring to the term-by-document matrix $tf_{t01,\ d0001}$ = 77 indicating that term t_{01} occurs in document d_{0001} 77 times, $tf_{t04,\ d0003}$ = 10 indicating that term t_{04} occurs in document d_{0003} ten times, etc. Referring to the phrase-term-by-document matrix $ptf_{pt06,\ d0009}$ = 17 indicating that phrase-term pt_{06} occurs in document d_{0009} 17 times, $ptf_{pt06,\ d0016}$ = 10 indicating that phrase-term pt_{06} occurs in document d_{0016} ten times, etc.

C.2.3 Performance measurements

Listed below in the following three tables C.11, C.12 and C.13 are the results, in binary, from the user's judged information-need-by-document matrix together with the information-need-by-document matrix produced by IRS-I, and finally, the information-need-by-document matrix produced by IRS-H.

Table C.11: Pilot 3 – User information-need-by-document matrix

User

doc	in01	in02	in03	in04	in05	in06	in07	in08	in09	in10	in11	in12	in13	in14
d0001	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0002	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0003	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0004	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0005	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0006	1	0	1	0	0	0	1	0	1	0	0	0	0	0
d0007	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0008	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0009	1	0	0	0	1	0	1	1	1	0	0	0	0	0
d0010	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0011	0	0	0	0	1	0	0	0	1	0	0	0	0	0
d0012	1	0	0	0	1	0	1	0	1	0	0	0	0	0
d0013	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0014	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0015	0	0	0	0	1	0	1	1	1	0	0	0	0	0
d0016	0	0	0	0	1	0	1	1	1	0	0	0	0	0
d0017	0	0	0	0	1	0	1	1	1	0	0	0	0	0
d0018	0	0	1	0	1	0	1	0	1	0	0	0	0	0
d0019	0	1	0	0	1	0	1	1	1	0	0	0	0	0
d0020	0	0	0	0	1	0	1	1	1	0	0	0	0	0

Table C.12: Pilot 3 – IRS-I: information-need-by-document matrix

IRS-I

Inverted index method

doc	in01	in02	in03	in04	in05	in06	in07	in08	in09	in10	in11	in12	in13	in14
d0001	1	0	0	0	0	0	0	0	0	1	1	0	1	1
d0002	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0003	1	0	0	0	0	0	0	0	0	1	1	1	1	1
d0004	1	0	0	0	0	0	1	1	1	1	1	0	1	1
d0005	0	0	0	1	0	0	0	0	1	1	1	0	1	1
d0006	1	1	1	1	1	1	1	1	1	1	1	1	1	1
d0007	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0008	1	0	0	1	0	1	1	1	1	1	1	1	1	1
d0009	1	1	1	1	1	1	1	1	1	1	1	1	1	1
d0010	1	0	0	0	0	0	1	1	1	1	1	1	1	1
d0011	1	1	1	1	1	1	1	1	1	1	1	1	1	1
d0012	1	1	1	1	1	1	1	1	1	1	1	1	1	1
d0013	1	0	0	0	0	1	0	0	1	1	1	1	1	1
d0014	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0015	1	1	1	1	1	1	1	1	1	1	1	1	1	1
d0016	1	1	1	1	1	1	1	1	1	1	1	1	1	1
d0017	1	1	1	1	1	1	1	1	1	1	1	1	1	1
d0018	1	1	1	1	1	1	1	1	1	1	1	1	1	1
d0019	1	1	1	1	1	1	1	1	1	1	1	1	1	1
d0020	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Table C.13: Pilot 3 – IRS-H: information-need-by-document matrix

IRS-H

Hybrid index method

doc	in01	in02	in03	in04	in05	in06	in07	in08	in09	in10	in11	in12	in13	in14
d0001	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0002	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0003	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0004	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0005	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0006	1	0	1	0	0	0	1	0	1	0	0	0	0	0
d0007	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0008	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0009	1	0	0	0	1	0	1	1	1	0	0	0	0	0
d0010	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0011	0	0	0	0	1	0	0	0	1	0	0	0	0	0
d0012	1	0	0	0	1	0	1	0	1	0	0	0	0	0
d0013	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0014	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0015	0	0	0	0	1	0	1	1	1	0	0	0	0	0
d0016	0	0	0	0	1	0	1	1	1	0	0	0	0	0
d0017	0	0	0	0	1	0	1	1	1	0	0	0	0	0
d0018	0	0	1	0	1	0	1	0	1	0	0	0	0	0
d0019	0	1	0	0	1	0	1	1	1	0	0	0	0	0
d0020	0	0	0	0	1	0	1	1	1	0	0	0	0	0

C.3 Evaluation

These performance measurements for Pilot 3 are all listed in table form for the IRS-I as well as IRS-H in Table C.14 and Table C.15 respectively.

Table C.14: Pilot 3 – IRS-I: performance measurements

In No	tp	fp	fn	tn	tpfp	fntn	tpfn	fptn	tpfpfntn	P	R	F
in01	3	13	0	4	16	4	3	17	20	19	100	32
in02	1	9	0	10	10	10	1	19	20	10	100	18
in03	2	8	0	10	10	10	2	18	20	20	100	33
in04	0	12	0	8	12	8	0	20	20	0	0	0
in05	9	1	0	10	10	10	9	11	20	90	100	95
in06	0	12	0	8	12	8	0	20	20	0	0	0
in07	9	4	0	7	13	7	9	11	20	69	100	82
in08	6	7	0	7	13	7	6	14	20	46	100	63
in09	10	5	0	5	15	5	10	10	20	67	100	80
in10	0	17	0	3	17	3	0	20	20	0	0	0
in11	0	17	0	3	17	3	0	20	20	0	0	0
in12	0	14	0	6	14	6	0	20	20	0	0	0
in13	0	17	0	3	17	3	0	20	20	0	0	0
in14	0	17	0	3	17	3	0	20	20	0	0	0

Table C.15: Pilot 3 – IRS-H: performance measurements

In No	tp	fp	fn	tn	tpfp	fntn	tpfn	fptn	tpfpfntn	P	R	F
in01	3	0	0	17	3	17	3	17	20	100	100	100
in02	1	0	0	19	1	19	1	19	20	100	100	100
in03	2	0	0	18	2	18	2	18	20	100	100	100
in04	0	0	0	20	0	20	0	20	20	0	0	0
in05	9	0	0	11	9	11	9	11	20	100	100	100
in06	0	0	0	20	0	20	0	20	20	0	0	0
in07	9	0	0	11	9	11	9	11	20	100	100	100
in08	6	0	0	14	6	14	6	14	20	100	100	100
in09	10	0	0	10	10	10	10	10	20	100	100	100
in10	0	0	0	20	0	20	0	20	20	0	0	0
in11	0	0	0	20	0	20	0	20	20	0	0	0
in12	0	0	0	20	0	20	0	20	20	0	0	0
in13	0	0	0	20	0	20	0	20	20	0	0	0
in14	0	0	0	20	0	20	0	20	20	0	0	0

Note that for IRS-H, *fp* and *fn* values were all zero while in IRS-I the *fn* values were all zero. Because IRS-H *fp* values were zero, all documents judged relevant by the user were retrieved by IRS-H exactly with no differentiation. However for IRS-I *fp* values ranged between 1 and 17, suggesting discrepancies between user and IRS-I judgements.

In Figure C.4, the 14 information needs for Pilot 3 are presented for each of the IRSs. The diamonds represent the F-measure values for IRS-I, and the squares represent the F-measure values for IRS-H.

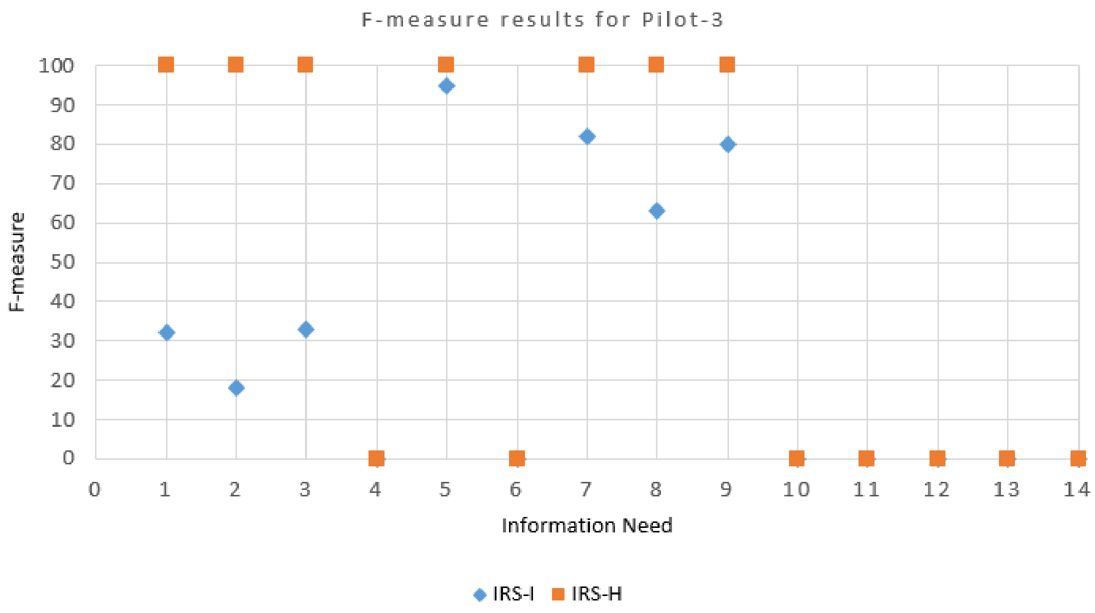


Figure C.4: Pilot 3 – IRS-I and IRS-H performance measurements

Referring to the performance measurement graph for Pilot 3 in Figure C.4, the computed F-measure values for the 14 information needs defined by the user are presented for both IRS-I and IRS-H. Between the two methods, there was agreement with 7 of the 14 or 50% of the information needs albeit the values are all zero. Of the remaining seven information needs, IRS-H achieved an F-measure of 100% while the values for IRS-I ranged between 18% and 95%. For IRS-I the Recall values were all 100%, suggesting the terms within the queries were identified and acquired from the documents exactly.

However, what is significant was the ability of IRS-H to match the phrase-terms in the queries exactly to those in the documents. These were verified by using Adobe's PDF advanced find search to physically check whether these phrase-terms existed in the documents or not. IRS-H and the user agreed on the judgments made, hence the 100% values for Precision, Recall, and F-measure.

C.4 Summary

At this stage at the end of Pilot 3, there was evidence to suggest that the functionality of IRS-H was more effective than IRS-I but this needed further investigation, testing, and evaluation with input from participating users. The design findings for Pilot 3 are summarised in Table C.16.

Table C.16: Pilot 3 – Summary of design findings

Pilot	Stage	Finding
3	Information gathering	Pilot 3 was based on a sample 20 journal articles, conference papers, and theses.
3	Information gathering	Content acquisition: the document collection was increased from a single document to 20 documents. The contents acquired from the pdf documents were converted to text successfully.
3	Search engine	Phrase-term: 14 phrase-terms provided by the user (the researcher in this pilot) were presented correctly, all in lowercase without special characters.
3	Search engine	Phrase-term query: these phrase-terms were expressed as 14 queries, four of which were expanded queries.
3	Search engine	Hybrid query index: the 14 queries were presented to the hybrid query index, which was thereafter populated with the phrase-terms, and unique begin and end token IDs.
3	Search engine	The hybrid query index interrogated the hybrid token index successfully and where a match was found (a phrase-term existed in a document) the document number was returned and the hybrid query index updated with the document number accordingly.
3	Design	Phrase-term frequency (ptf) was maintained.
3	Design	Converting ptf values to binary and the population of the phrase-term-by-document matrix with these values remained successful.
3	Design	IRS-H was able to match phrase-terms expressed in queries to those in documents exactly.
3	Design	IRS-H was able to maintain word ordinality and word proximity.
3	Design	At this stage at the end of Pilot 3, there was evidence to suggest that the functionality of IRS-H was more effective than IRS-I but this needed further investigation, testing, and evaluation with input from participating users.

These design findings from Appendices A, B and C for Pilot tests 1, 2, and 3 are utilised in Volume I, Chapter Four, sections 4.41, 4.42 and 4.43 respectively.

APPENDIX D: DOCUMENTS, QUERIES AND PHRASE-TERMS RESULTS

Appendix D follows on from Chapter Four and contains the full data tables for the results of this thesis.

Table D.1: Document table results

doc	File Name	Directory
d0001	a data quality measurement information model based on iso--iec 15939.txt	C:\Thesis\PhD-2018\Data\Txt\
d0002	A Design Science Research Methodology for Information Systems Research - Peffers.txt	C:\Thesis\PhD-2018\Data\Txt\
d0003	A framework for outsourcing ISiT security services.txt	C:\Thesis\PhD-2018\Data\Txt\
d0004	A framework for rigorously identifying research gaps in qualitative lit review.txt	C:\Thesis\PhD-2018\Data\Txt\
d0005	A Framework for Techniques for Information Technology.txt	C:\Thesis\PhD-2018\Data\Txt\
d0006	AD Agile QL-QR6XNEAs.txt	C:\Thesis\PhD-2018\Data\Txt\
d0007	Adebasin (suspect).txt	C:\Thesis\PhD-2018\Data\Txt\
d0008	AFramework4CorporateHouseholding.txt	C:\Thesis\PhD-2018\Data\Txt\
d0009	Agency double dance Rose 1-1-rose.txt	C:\Thesis\PhD-2018\Data\Txt\
d0010	Agency Rose double-dance 10.1.1.201.9129.txt	C:\Thesis\PhD-2018\Data\Txt\
d0011	ANT Hansen-etal-2004_Actor_Network_Theory_and_Information_Systems_ITP[1].txt	C:\Thesis\PhD-2018\Data\Txt\
d0012	ANT IT Elderly Care The_values[1].txt	C:\Thesis\PhD-2018\Data\Txt\
d0013	ANT semiotics Law 2009[1].txt	C:\Thesis\PhD-2018\Data\Txt\
d0014	ANT space 1-s2.0-S0143622805000275-main[1].txt	C:\Thesis\PhD-2018\Data\Txt\
d0015	ANT4DPaper1Heeks.txt	C:\Thesis\PhD-2018\Data\Txt\
d0016	ANT4DWorkingPaper2FaikEtAl.txt	C:\Thesis\PhD-2018\Data\Txt\
d0017	Belangrike articles.txt	C:\Thesis\PhD-2018\Data\Txt\
d0018	best_practices_for_data_stew_153470[1].txt	C:\Thesis\PhD-2018\Data\Txt\
d0019	BI&DQ iciq08.txt	C:\Thesis\PhD-2018\Data\Txt\
d0020	Big Data Publication 39879.txt	C:\Thesis\PhD-2018\Data\Txt\
d0021	Burell Morgan Design JohanssonWoodilla_DMIProceedings[1].txt	C:\Thesis\PhD-2018\Data\Txt\
d0022	burrel-morgan-explained.txt	C:\Thesis\PhD-2018\Data\Txt\
d0023	Burrell_and_Morgan_4_Paradigms_v2lsu.txt	C:\Thesis\PhD-2018\Data\Txt\
d0024	Burton-Jones Using IS effectively.txt	C:\Thesis\PhD-2018\Data\Txt\
d0025	Business Process Management 14637151311294831.txt	C:\Thesis\PhD-2018\Data\Txt\
d0026	Capabilities of Sen Evans 2002.txt	C:\Thesis\PhD-2018\Data\Txt\
d0027	Capability and theory RobeynsJHDoncapabilities.txt	C:\Thesis\PhD-2018\Data\Txt\
d0028	Capability care giving barjis_2013_DSS.txt	C:\Thesis\PhD-2018\Data\Txt\
d0029	CHEC tender for research into graduate destinations.txt	C:\Thesis\PhD-2018\Data\Txt\
d0030	choudrie.txt	C:\Thesis\PhD-2018\Data\Txt\
d0031	CIS Adoption.txt	C:\Thesis\PhD-2018\Data\Txt\
d0032	Co-development copda2014_submission_01.txt	C:\Thesis\PhD-2018\Data\Txt\
d0033	CoDesign and software development A Framework For Behavioral Studies of Technology Framing In Information Systems Design 2.txt	C:\Thesis\PhD-2018\Data\Txt\
d0034	CoDesign and software development A Framework For Behavioral Studies of Technology Framing In Information Systems Design.txt	C:\Thesis\PhD-2018\Data\Txt\
d0035	CoDesignInterface1500030a.txt	C:\Thesis\PhD-2018\Data\Txt\
d0036	Community engagement wellbeing A_guide_to_community-centred_approaches_for_health_and_wellbeing.txt	C:\Thesis\PhD-2018\Data\Txt\
d0037	Community From Digital Divide to Digital Inclusion and Beyond _ Nemer _ The Journal of Community Informatics.txt	C:\Thesis\PhD-2018\Data\Txt\
d0038	Community Guide's Social Environment and Health Model (1).txt	C:\Thesis\PhD-2018\Data\Txt\
d0039	CommunityBasedParticipatoryResearchSA.txt	C:\Thesis\PhD-2018\Data\Txt\
d0040	CommunityEngagementHC.txt	C:\Thesis\PhD-2018\Data\Txt\
d0041	CommunityOfPracticeNursing.txt	C:\Thesis\PhD-2018\Data\Txt\

doc	File Name	Directory
d0042	Context Davison and Marthinsons art_10.1057_jit.2015.19.txt	C:\Thesis\PhD-2018\Data\Txt\
d0043	Context dillon2[1].txt	C:\Thesis\PhD-2018\Data\Txt\
d0044	Context HC.txt	C:\Thesis\PhD-2018\Data\Txt\
d0045	CSCW Fitzpatrick-Ellingsen-CSCW.txt	C:\Thesis\PhD-2018\Data\Txt\
d0046	CSCW The_concept_of_practice_Whats_the_point.txt	C:\Thesis\PhD-2018\Data\Txt\
d0047	Curriculum Informatics Jobs J Am Med Inform Assoc-2012-Ohno-Machado-919[1].txt	C:\Thesis\PhD-2018\Data\Txt\
d0048	Curriculum ISEDJv10n2p15[1].txt	C:\Thesis\PhD-2018\Data\Txt\
d0049	Curriculum ISEDJv10n3p35[1].txt	C:\Thesis\PhD-2018\Data\Txt\
d0050	DataIntegration.txt	C:\Thesis\PhD-2018\Data\Txt\
d0051	dataquality-vocab-lwdm2011.txt	C:\Thesis\PhD-2018\Data\Txt\
d0052	data_quality_part2.txt	C:\Thesis\PhD-2018\Data\Txt\
d0053	Delphi OkoliPawlowski2004DelphiPostprint.txt	C:\Thesis\PhD-2018\Data\Txt\
d0054	Design and delivery of social networked learning.txt	C:\Thesis\PhD-2018\Data\Txt\
d0055	Design ethnography.txt	C:\Thesis\PhD-2018\Data\Txt\
d0056	design ethnography[1].txt	C:\Thesis\PhD-2018\Data\Txt\
d0057	Design research and meaning making eScholarship UC item 0mr972w6.txt	C:\Thesis\PhD-2018\Data\Txt\
d0058	Design research philosophy worldviews1-s2.0-S0142694X08000203-main.txt	C:\Thesis\PhD-2018\Data\Txt\
d0059	Design research Pragmatism GG-EDSS2011.txt	C:\Thesis\PhD-2018\Data\Txt\
d0060	Design research science Reich art_10.1007_s00163-013-0163-3.txt	C:\Thesis\PhD-2018\Data\Txt\
d0061	Design research service design customer experience JOSM2012.txt	C:\Thesis\PhD-2018\Data\Txt\
d0062	Design Theories in Information Systems - A Need for Multi-Groundi.txt	C:\Thesis\PhD-2018\Data\Txt\
d0063	Design theory Gregor.txt	C:\Thesis\PhD-2018\Data\Txt\
d0064	Design theory papers.txt	C:\Thesis\PhD-2018\Data\Txt\
d0065	Design theory process pragmatism GG-EDSS2012.txt	C:\Thesis\PhD-2018\Data\Txt\
d0066	Development Service delivery indicators 4284-service-delivery-indicators.txt	C:\Thesis\PhD-2018\Data\Txt\
d0067	Development Service Delivery Indicators What_is_SDI.txt	C:\Thesis\PhD-2018\Data\Txt\
d0068	Digital services eras Technology driven evolution of design practices envisioning the role of design in the digital era.txt	C:\Thesis\PhD-2018\Data\Txt\
d0069	Digital services HOFE14ExperiencesInApplyingServiceDesign.txt	C:\Thesis\PhD-2018\Data\Txt\
d0070	DQ Context 0912f513914cb6e4f5000000.txt	C:\Thesis\PhD-2018\Data\Txt\
d0071	DQ Healthcare.txt	C:\Thesis\PhD-2018\Data\Txt\
d0072	DQ Methodologies for Data Quality Assessment and Improvement[1].txt	C:\Thesis\PhD-2018\Data\Txt\
d0073	DQConceptualModel 2.txt	C:\Thesis\PhD-2018\Data\Txt\
d0074	DSR and Service Design Hofemann_Raatikainen_Myllärniemi_Norja_Experiences_in_Applying_Service_Design_to_Digital_Services.txt	C:\Thesis\PhD-2018\Data\Txt\
d0075	DSR design of IT artefact ejbrm-volume10-issue2-article281 (6).txt	C:\Thesis\PhD-2018\Data\Txt\
d0076	DSR Evaluation 20080023.txt	C:\Thesis\PhD-2018\Data\Txt\
d0077	DSR Evaluation Framework Comprehensive_Framework_for_Evaluation_in_DSR_offprint (1).txt	C:\Thesis\PhD-2018\Data\Txt\
d0078	DSR IS Design Research Framework - A Critical Realist Perspective.txt	C:\Thesis\PhD-2018\Data\Txt\
d0079	DSR Peffers-etal-2008_DesignScienceResearchMethodology[1].txt	C:\Thesis\PhD-2018\Data\Txt\
d0080	eHealth 130522 HIMJ Ruxwana online_2014.txt	C:\Thesis\PhD-2018\Data\Txt\
d0081	eHealth challenges paper-based records.txt	C:\Thesis\PhD-2018\Data\Txt\
d0082	EHR Data_Model_Paper_Final Version as Uploaded.txt	C:\Thesis\PhD-2018\Data\Txt\
d0083	EHR DE.txt	C:\Thesis\PhD-2018\Data\Txt\
d0084	EHR Socio-technical art05[1].txt	C:\Thesis\PhD-2018\Data\Txt\
d0085	EHR structure.txt	C:\Thesis\PhD-2018\Data\Txt\
d0086	Emotion annotation sentimentMKZ.txt	C:\Thesis\PhD-2018\Data\Txt\
d0087	Emotion annotation short text NRC-Sentiment-JAIR-2014.txt	C:\Thesis\PhD-2018\Data\Txt\
d0088	Ethics 4D 1049-2946-1-PB (2).txt	C:\Thesis\PhD-2018\Data\Txt\

doc	File Name	Directory
d0089	Ethics design wellbeing Manzini 060828-design-ethics-sustainability.txt	C:\Thesis\PhD-2018\Data\Txt\
d0090	Evaluation Design_principles.txt	C:\Thesis\PhD-2018\Data\Txt\
d0091	Evaluation ICT4D Mobile Phones di_wp39.txt	C:\Thesis\PhD-2018\Data\Txt\
d0092	Evaluation IS ejsr_37_2_05[1].txt	C:\Thesis\PhD-2018\Data\Txt\
d0093	evaluation4D Heeks .txt	C:\Thesis\PhD-2018\Data\Txt\
d0094	ISEDJv9n6p11.txt	C:\Thesis\PhD-2018\Data\Txt\
d0095	rCollaboration choreographies soca.txt	C:\Thesis\PhD-2018\Data\Txt\
d0096	structuration theory.txt	C:\Thesis\PhD-2018\Data\Txt\
d0097	Studying the Impact of Personality Constructs on Employees'.txt	C:\Thesis\PhD-2018\Data\Txt\
d0098	Thesis_Mongezi_Mati_18Nov 15.txt	C:\Thesis\PhD-2018\Data\Txt\
d0099	Thesis_Regis_Muchemwa_Wismar_Final_06Oct2015 changes.txt	C:\Thesis\PhD-2018\Data\Txt\
d0100	Visual Methods and the World Technique - The Importance of the Elicitation Interview in Understanding Non-traditional Students- Journeys through University.txt	C:\Thesis\PhD-2018\Data\Txt\

Table D.2: Information need table results

In No	Information Need
in01	I want to find all documents relevant to design science research
in02	I want to find all documents relevant to qualitative research
in03	I want to find all documents relevant to quantitative research
in04	I want to find all documents relevant to clinical guidelines
in05	I want to find all documents relevant to cloud computing
in06	I want to find all documents relevant to conceptual frameworks
in07	I want to find all documents relevant to research ethics
in08	I want to find all documents relevant to design research methods
in09	I want to find all documents relevant to data quality
in10	I want to find all documents relevant to electronic health records
in11	I want to find all documents relevant to design science
in12	I want to find all documents relevant to design sciences
in13	I want to find all documents relevant to design science research
in14	I want to find all documents relevant to design science methodology
in15	I want to find all documents relevant to the design method
in16	I want to find all documents relevant to design research
in17	I want to find all documents relevant to design science research paradigm
in18	I want to find all documents relevant to design science research paradigms
in19	I want to find all documents relevant to qualitative method
in20	I want to find all documents relevant to qualitative analysis
in21	I want to find all documents relevant to qualitative research
in22	I want to find all documents relevant to qualitative research design
in23	I want to find all documents relevant to qualitative research method
in24	I want to find all documents relevant to qualitative research methods
in25	I want to find all documents relevant to qualitative research methodology
in26	I want to find all documents relevant to quantitative method
in27	I want to find all documents relevant to quantitative analysis
in28	I want to find all documents relevant to quantitative research
in29	I want to find all documents relevant to quantitative research design
in30	I want to find all documents relevant to quantitative research method
in31	I want to find all documents relevant to quantitative research methods
in32	I want to find all documents relevant to quantitative research methodology
in33	I want to find all documents relevant to clinical guideline

In No	Information Need
in34	I want to find all documents relevant to clinical guidelines
in35	I want to find all documents relevant to clinical guidelines in primary care
in36	I want to find all documents relevant to clinical guidelines in family practice
in37	I want to find all documents relevant to clinical guidelines for operations
in38	I want to find all documents relevant to clinical guidelines for stroke management
in39	I want to find all documents relevant to cloud computing
in40	I want to find all documents relevant to cloud computing types
in41	I want to find all documents relevant to cloud computing models
in42	I want to find all documents relevant to cloud computing service models
in43	I want to find all documents relevant to conceptual framework
in44	I want to find all documents relevant to conceptual frameworks
in45	I want to find all documents relevant to conceptual framework in research
in46	I want to find all documents relevant to conceptual frameworks in research
in47	I want to find all documents relevant to conceptual model
in48	I want to find all documents relevant to conceptual models
in49	I want to find all documents relevant to research ethics
in50	I want to find all documents relevant to ethics in research
in51	I want to find all documents relevant to research ethics principles
in52	I want to find all documents relevant to design method
in53	I want to find all documents relevant to design methods
in54	I want to find all documents relevant to design practice
in55	I want to find all documents relevant to design research methods
in56	I want to find all documents relevant to design research method
in57	I want to find all documents relevant to design research philosophy
in58	I want to find all documents relevant to design research pragmatism
in59	I want to find all documents relevant to design theory
in60	I want to find all documents relevant to data quality
in61	I want to find all documents relevant to data qualities
in62	I want to find all documents relevant to data quality methodology
in63	I want to find all documents relevant to data quality methodologies
in64	I want to find all documents relevant to data quality model
in65	I want to find all documents relevant to data quality models
in66	I want to find all documents relevant to data quality conceptual models
in67	I want to find all documents relevant to data quality conceptual model
in68	I want to find all documents relevant to data quality framework
in69	I want to find all documents relevant to data quality frameworks
in70	I want to find all documents relevant to electronic health record
in71	I want to find all documents relevant to electronic health records
in72	I want to find all documents relevant to e health record
in73	I want to find all documents relevant to e health records
in74	I want to find all documents relevant to electronic patient record
in75	I want to find all documents relevant to electronic patient records

Table D.3: Query table results

q	Query
q01	"design science"
q02	"design sciences"
q03	"design science research"
q04	"design science methodology"

q	Query
q05	"the design method"
q06	"design research"
q07	"design science research paradigm"
q08	"design science research paradigms"
q09	"qualitative method"
q10	"qualitative analysis"
q11	"qualitative research"
q12	"qualitative research design"
q13	"qualitative research method"
q14	"qualitative research methods"
q15	"qualitative research methodology"
q16	"quantitative method"
q17	"quantitative analysis"
q18	"quantitative research"
q19	"quantitative research design"
q20	"quantitative research method"
q21	"quantitative research methods"
q22	"quantitative research methodology"
q23	"clinical guideline"
q24	"clinical guidelines"
q25	"clinical guidelines in primary care"
q26	"clinical guidelines in family practice"
q27	"clinical guidelines for operations"
q28	"clinical guidelines for stroke management"
q29	"cloud computing"
q30	"cloud computing types"
q31	"cloud computing models"
q32	"cloud computing service models"
q33	"conceptual framework"
q34	"conceptual frameworks"
q35	"conceptual framework in research"
q36	"conceptual frameworks in research"
q37	"conceptual model"
q38	"conceptual models"
q39	"research ethics"
q40	"ethics in research"
q41	"research ethics principles"
q42	"design method"
q43	"design methods"
q44	"design practice"
q45	"design research methods"
q46	"design research method"
q47	"design research philosophy"
q48	"design research pragmatism"
q49	"design theory"
q50	"data quality"
q51	"data qualities"
q52	"data quality methodology"

q	Query
q53	"data quality methodologies"
q54	"data quality model"
q55	"data quality models"
q56	"data quality conceptual models"
q57	"data quality conceptual model"
q58	"data quality framework"
q59	"data quality frameworks"
q60	"electronic health record"
q61	"electronic health records"
q62	"e health record"
q63	"e health records"
q64	"electronic patient record"
q65	"electronic patient records"
q66	"design science" OR "design sciences" OR "design science research" OR "design science methodology" OR "the design method" OR "design research" OR "design science research paradigm" OR "design science research paradigms"
q67	"qualitative method" OR "qualitative analysis" OR "qualitative research" OR "qualitative research design" OR "qualitative research method" OR "qualitative research methods" OR "qualitative research methodology"
q68	"quantitative method" OR "quantitative analysis" OR "quantitative research" OR "quantitative research design" OR "quantitative research method" OR "quantitative research methods" OR "quantitative research methodology"
q69	"clinical guideline" OR "clinical guidelines" OR "clinical guidelines in primary care" OR "clinical guidelines in family practice" OR "clinical guidelines for operations" OR "clinical guidelines for stroke management"
q70	"cloud computing" OR "cloud computing types" OR "cloud computing models" OR "cloud computing service models"
q71	"conceptual framework" OR "conceptual frameworks" OR "conceptual framework in research" OR "conceptual frameworks in research" OR "conceptual model" OR "conceptual models"
q72	"research ethics" OR "ethics in research" OR "research ethics principles"
q73	"design method" OR "design methods" OR "design practice" OR "design research methods" OR "design research method" OR "design research philosophy" OR "design research pragmatism" OR "design theory"
q74	"data quality" OR "data qualities" OR "data quality methodology" OR "data quality methodologies" OR "data quality model" OR "data quality models" OR "data quality conceptual models" OR "data quality conceptual model" OR "data quality framework" OR "data quality frameworks"
q75	"electronic health record" OR "electronic health records" OR "e health record" OR "e health records" OR "electronic patient record" OR "electronic patient records"

Table D.4: Phrase-term table results

pt	Phrase-term
pt01	design science
pt02	design sciences
pt03	design science research
pt04	design science methodology
pt05	the design method
pt06	design research
pt07	design science research paradigm
pt08	design science research paradigms
pt09	qualitative method
pt10	qualitative analysis
pt11	qualitative research
pt12	qualitative research design
pt13	qualitative research method
pt14	qualitative research methods
pt15	qualitative research methodology
pt16	quantitative method
pt17	quantitative analysis

pt	Phrase-term
pt18	quantitative research
pt19	quantitative research design
pt20	quantitative research method
pt21	quantitative research methods
pt22	quantitative research methodology
pt23	clinical guideline
pt24	clinical guidelines
pt25	clinical guidelines in primary care
pt26	clinical guidelines in family practice
pt27	clinical guidelines for operations
pt28	clinical guidelines for stroke management
pt29	cloud computing
pt30	cloud computing types
pt31	cloud computing models
pt32	cloud computing service models
pt33	conceptual framework
pt34	conceptual frameworks
pt35	conceptual framework in research
pt36	conceptual frameworks in research
pt37	conceptual model
pt38	conceptual models
pt39	research ethics
pt40	ethics in research
pt41	research ethics principles
pt42	design method
pt43	design methods
pt44	design practice
pt45	design research methods
pt46	design research method
pt47	design research philosophy
pt48	design research pragmatism
pt49	design theory
pt50	data quality
pt51	data qualities
pt52	data quality methodology
pt53	data quality methodologies
pt54	data quality model
pt55	data quality models
pt56	data quality conceptual models
pt57	data quality conceptual model
pt58	data quality framework
pt59	data quality frameworks
pt60	electronic health record
pt61	electronic health records
pt62	e health record
pt63	e health records
pt64	electronic patient record
pt65	electronic patient records

Table D.5: Information need query link table results

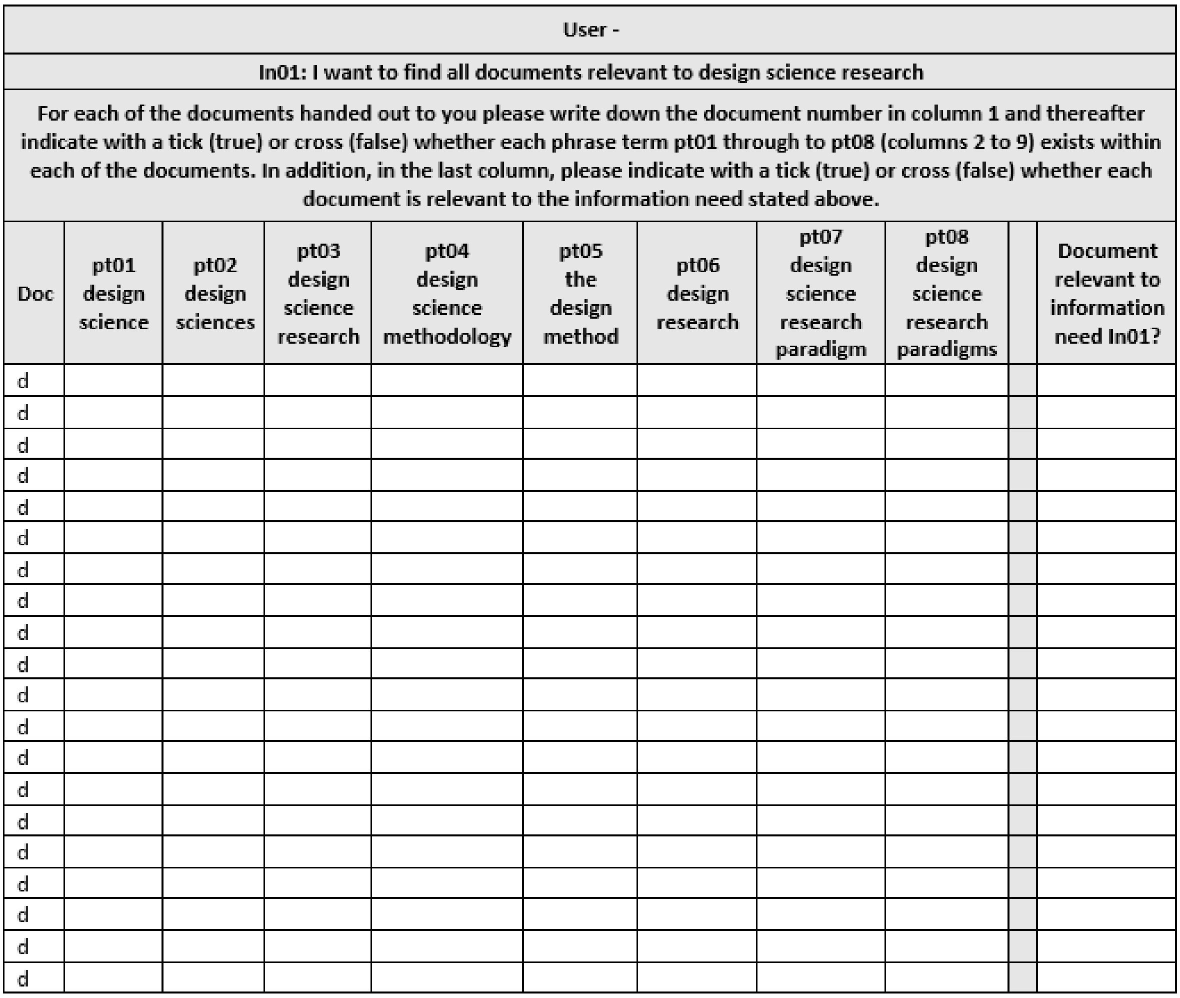
In No	q
in01	q01
in02	q02
in03	q03
in04	q04
in05	q05
in06	q06
in07	q07
in08	q08
in09	q09
in10	q10
in11	q11
in12	q12
in13	q13
in14	q14
in15	q15
in16	q16
in17	q17
in18	q18
in19	q19
in20	q20
in21	q21
in22	q22
in23	q23
in24	q24
in25	q25
in26	q26
in27	q27
in28	q28
in29	q29
in30	q30
in31	q31
in32	q32
in33	q33
in34	q34
in35	q35
in36	q36
in37	q37
in38	q38
in39	q39
in40	q40
in41	q41
in42	q42
in43	q43
in44	q44
in45	q45
in46	q46
in47	q47
in48	q48
in49	q49
in50	q50
in51	q51
in52	q52
in53	q53
in54	q54
in55	q55
in56	q56
in57	q57
in58	q58
in59	q59
in60	q60
in61	q61
in62	q62
in63	q63
in64	q64
in65	q65
in66	q66
in67	q67
in68	q68
in69	q69
in70	q70
in71	q71
in72	q72
in73	q73
in74	q74
in75	q75

Table D.6: Query phrase-term table results

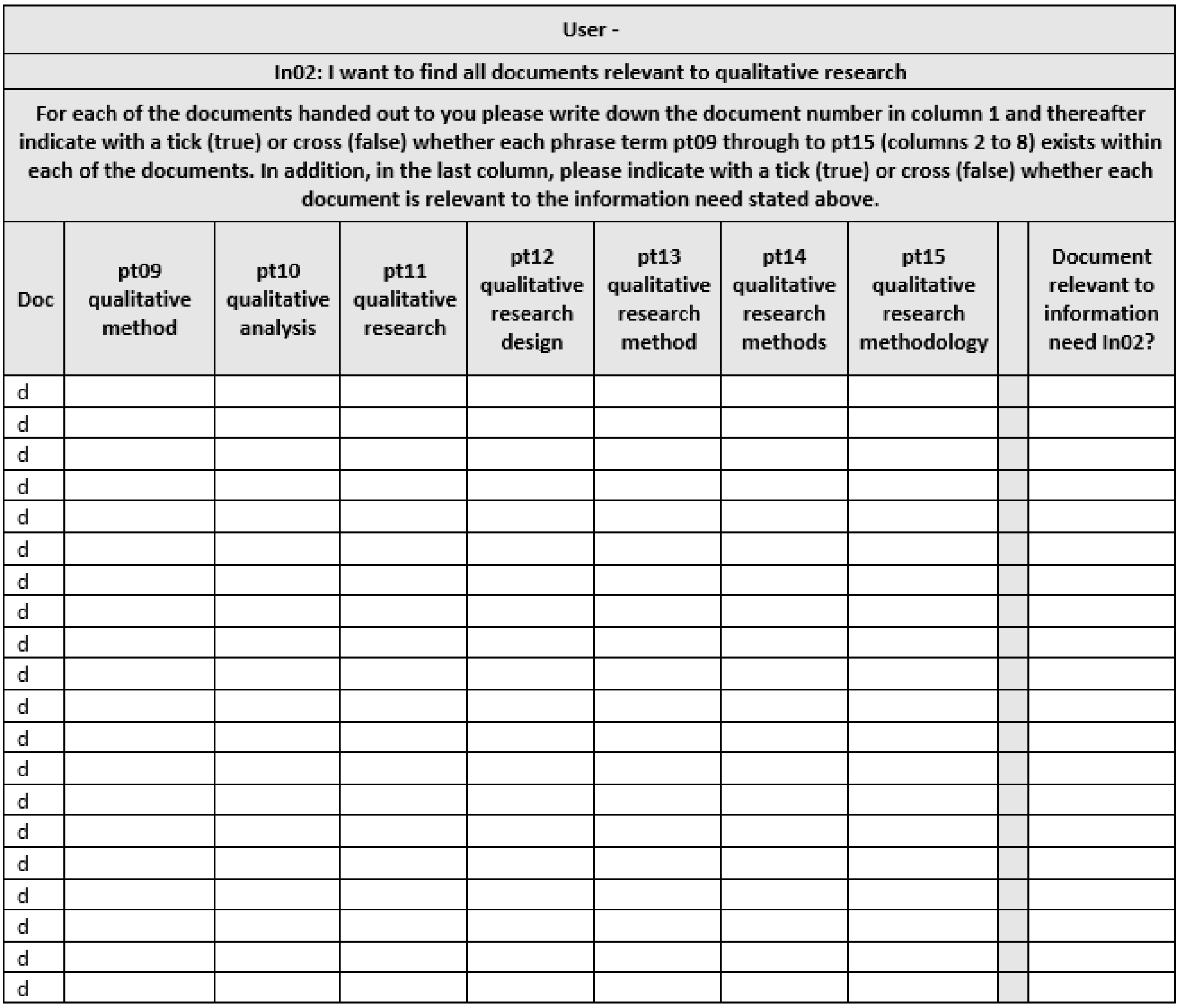
q	pt
q01	pt01
q02	pt02
q03	pt03
q04	pt04
q05	pt05
q06	pt06
q07	pt07
q08	pt08
q09	pt09
q10	pt10
q11	pt11
q12	pt12
q13	pt13
q14	pt14
q15	pt15
q16	pt16
q17	pt17
q18	pt18
q19	pt19
q20	pt20
q21	pt21
q22	pt22
q23	pt23
q24	pt24
q25	pt25
q26	pt26
q27	pt27
q28	pt28
q29	pt29
q30	pt30
q31	pt31
q32	pt32
q33	pt33
q34	pt34
q35	pt35
q36	pt36
q37	pt37
q38	pt38
q39	pt39
q40	pt40
q41	pt41
q42	pt42
q43	pt43
q44	pt44
q45	pt45
q46	pt46
q47	pt47
q48	pt48
q49	pt49
q50	pt50
q51	pt51
q52	pt52
q53	pt53
q54	pt54
q55	pt55
q56	pt56
q57	pt57
q58	pt58
q59	pt59
q60	pt60
q61	pt61
q62	pt62
q63	pt63
q64	pt64
q65	pt65
q66	pt01
q66	pt02
q66	pt03
q66	pt04
q66	pt05
q66	pt06
q66	pt07
q66	pt08
q67	pt09
q67	pt10
q67	pt11
q67	pt12
q67	pt13
q67	pt14
q67	pt15
q68	pt16
q68	pt17
q68	pt18
q68	pt19
q68	pt20
q68	pt21
q68	pt22
q69	pt23
q69	pt24
q69	pt25
q69	pt26
q69	pt27
q69	pt28
q70	pt29
q70	pt30
q70	pt31
q70	pt32
q71	pt33
q71	pt34
q71	pt35
q71	pt36
q71	pt37
q71	pt38
q72	pt39
q72	pt40
q72	pt41
q73	pt42
q73	pt43
q73	pt44
q73	pt45
q73	pt46
q73	pt47
q73	pt48
q73	pt49
q74	pt50
q74	pt51
q74	pt52
q74	pt53
q74	pt54
q74	pt55
q74	pt56
q74	pt57
q74	pt58
q74	pt59
q75	pt60
q75	pt61
q75	pt62
q75	pt63
q75	pt64
q75	pt65

APPENDIX E: USER QUESTIONNAIRE RESULTS

One page for each of the ten information needs:



User -										
In01: I want to find all documents relevant to design science research										
For each of the documents handed out to you please write down the document number in column 1 and thereafter indicate with a tick (true) or cross (false) whether each phrase term pt01 through to pt08 (columns 2 to 9) exists within each of the documents. In addition, in the last column, please indicate with a tick (true) or cross (false) whether each document is relevant to the information need stated above.										
Doc	pt01 design science	pt02 design sciences	pt03 design science research	pt04 design science methodology	pt05 the design method	pt06 design research	pt07 design science research paradigm	pt08 design science research paradigms		Document relevant to information need In01?
d										
d										
d										
d										
d										
d										
d										
d										
d										
d										
d										
d										
d										
d										
d										
d										
d										
d										
d										
d										

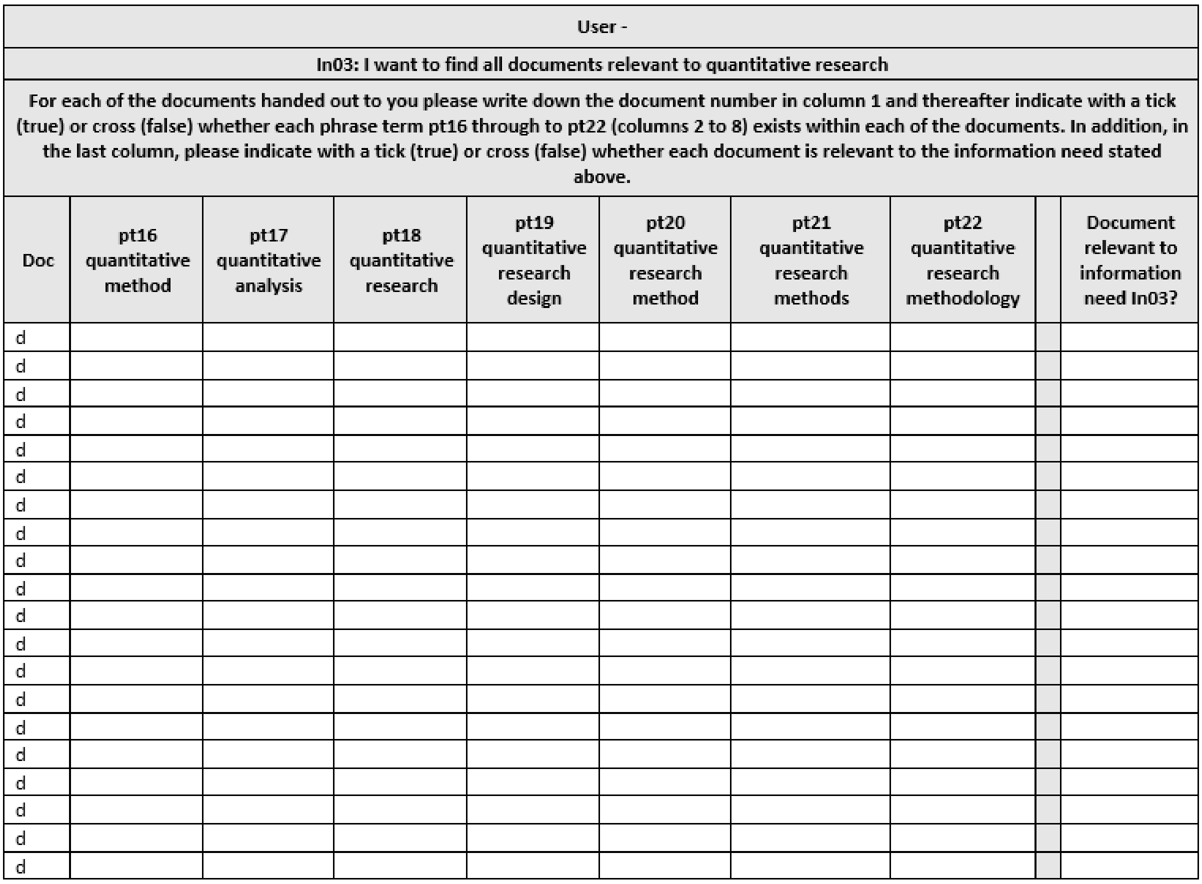


User -

In02: I want to find all documents relevant to qualitative research

For each of the documents handed out to you please write down the document number in column 1 and thereafter indicate with a tick (true) or cross (false) whether each phrase term pt09 through to pt15 (columns 2 to 8) exists within each of the documents. In addition, in the last column, please indicate with a tick (true) or cross (false) whether each document is relevant to the information need stated above.

Doc	pt09 qualitative method	pt10 qualitative analysis	pt11 qualitative research	pt12 qualitative research design	pt13 qualitative research method	pt14 qualitative research methods	pt15 qualitative research methodology		Document relevant to information need In02?
d									
d									
d									
d									
d									
d									
d									
d									
d									
d									
d									
d									
d									
d									
d									
d									
d									
d									
d									
d									

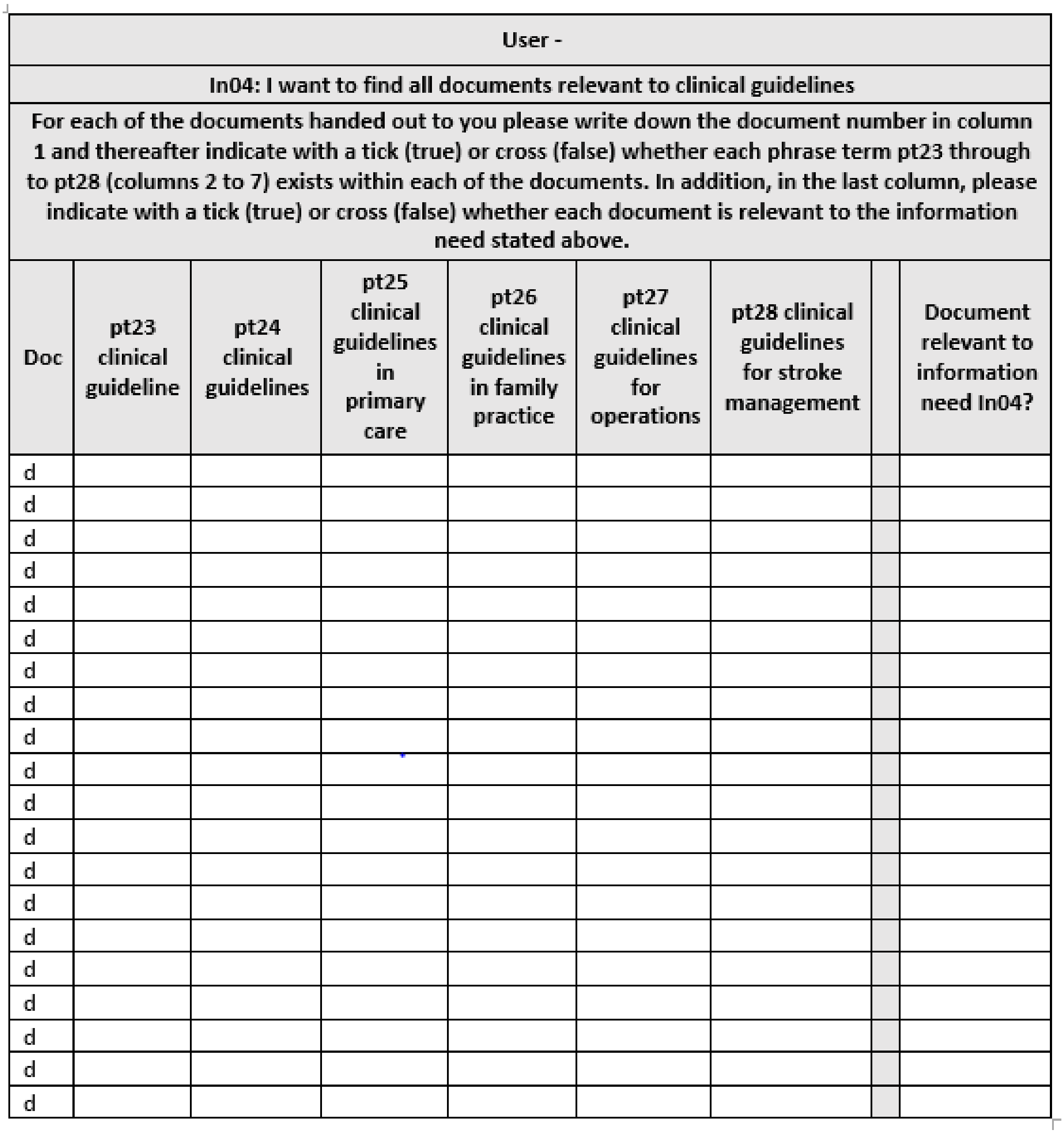


User -

In03: I want to find all documents relevant to quantitative research

For each of the documents handed out to you please write down the document number in column 1 and thereafter indicate with a tick (true) or cross (false) whether each phrase term pt16 through to pt22 (columns 2 to 8) exists within each of the documents. In addition, in the last column, please indicate with a tick (true) or cross (false) whether each document is relevant to the information need stated above.

Doc	pt16 quantitative method	pt17 quantitative analysis	pt18 quantitative research	pt19 quantitative research design	pt20 quantitative research method	pt21 quantitative research methods	pt22 quantitative research methodology		Document relevant to information need In03?
d									
d									
d									
d									
d									
d									
d									
d									
d									
d									
d									
d									
d									
d									
d									
d									
d									
d									
d									
d									

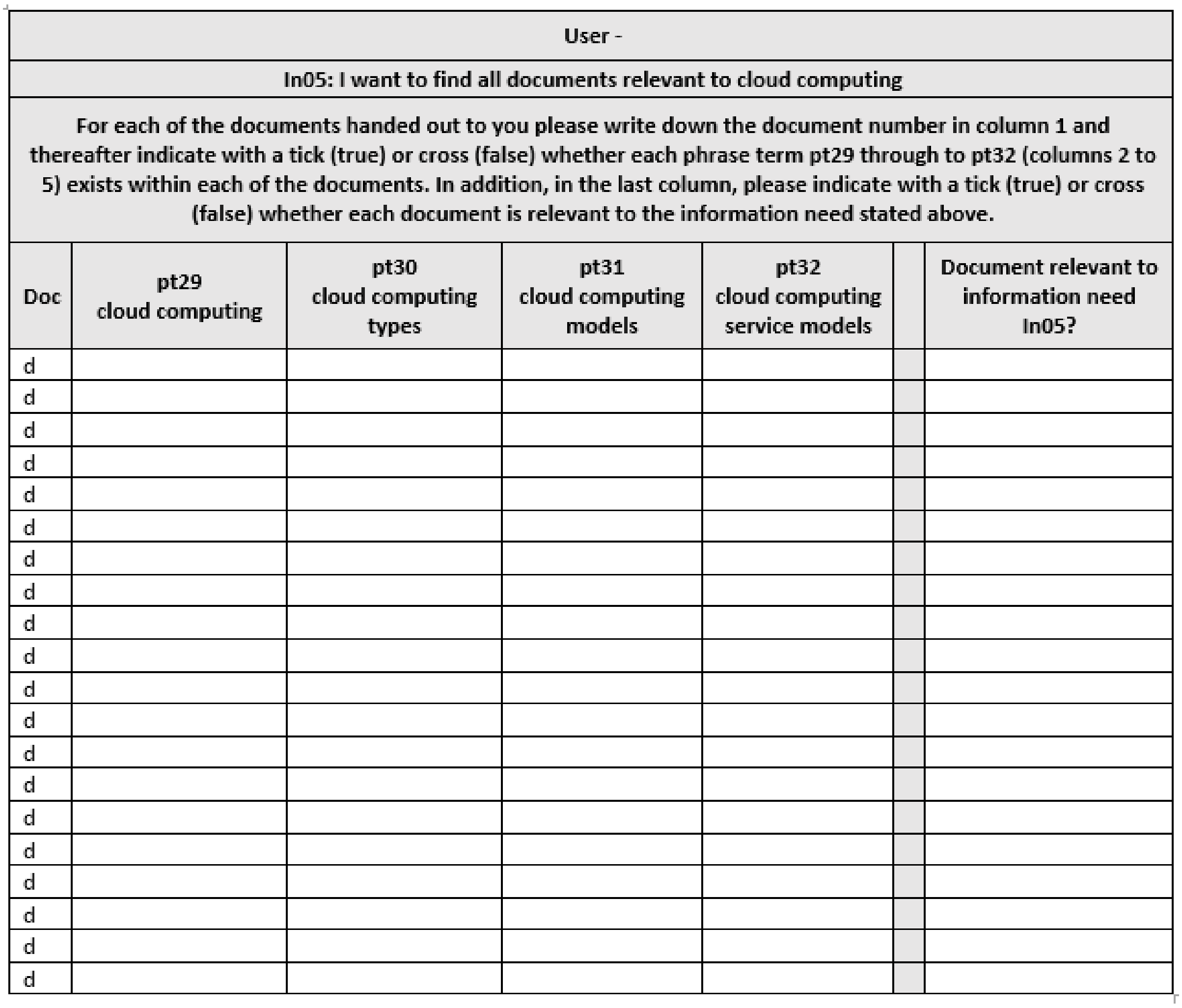


User -

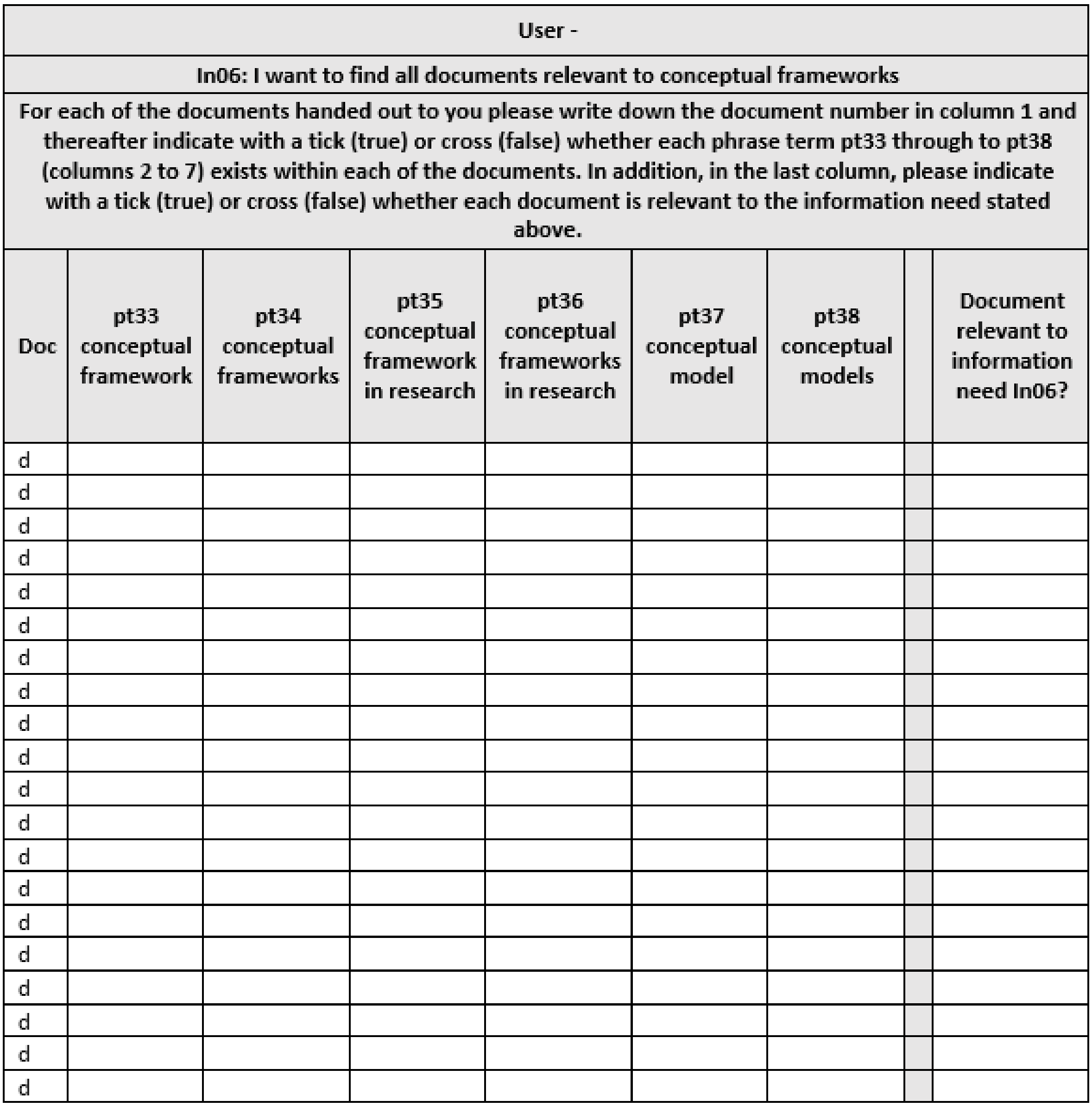
In04: I want to find all documents relevant to clinical guidelines

For each of the documents handed out to you please write down the document number in column 1 and thereafter indicate with a tick (true) or cross (false) whether each phrase term pt23 through to pt28 (columns 2 to 7) exists within each of the documents. In addition, in the last column, please indicate with a tick (true) or cross (false) whether each document is relevant to the information need stated above.

Doc	pt23 clinical guideline	pt24 clinical guidelines	pt25 clinical guidelines in primary care	pt26 clinical guidelines in family practice	pt27 clinical guidelines for operations	pt28 clinical guidelines for stroke management		Document relevant to information need In04?
d								
d								
d								
d								
d								
d								
d								
d								
d								
d								
d								
d								
d								
d								
d								
d								
d								
d								
d								
d								



<table>
<tr><th colspan="7">User -</th></tr>
<tr><th colspan="7">In05: I want to find all documents relevant to cloud computing</th></tr>
<tr><th colspan="7">For each of the documents handed out to you please write down the document number in column 1 and thereafter indicate with a tick (true) or cross (false) whether each phrase term pt29 through to pt32 (columns 2 to 5) exists within each of the documents. In addition, in the last column, please indicate with a tick (true) or cross (false) whether each document is relevant to the information need stated above.</th></tr>
<tr><th>Doc</th><th>pt29
cloud computing</th><th>pt30
cloud computing types</th><th>pt31
cloud computing models</th><th>pt32
cloud computing service models</th><th></th><th>Document relevant to information need In05?</th></tr>
<tr><td>d</td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>d</td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>d</td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>d</td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>d</td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>d</td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>d</td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>d</td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>d</td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>d</td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>d</td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>d</td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>d</td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>d</td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>d</td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>d</td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>d</td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>d</td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>d</td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>d</td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
</table>

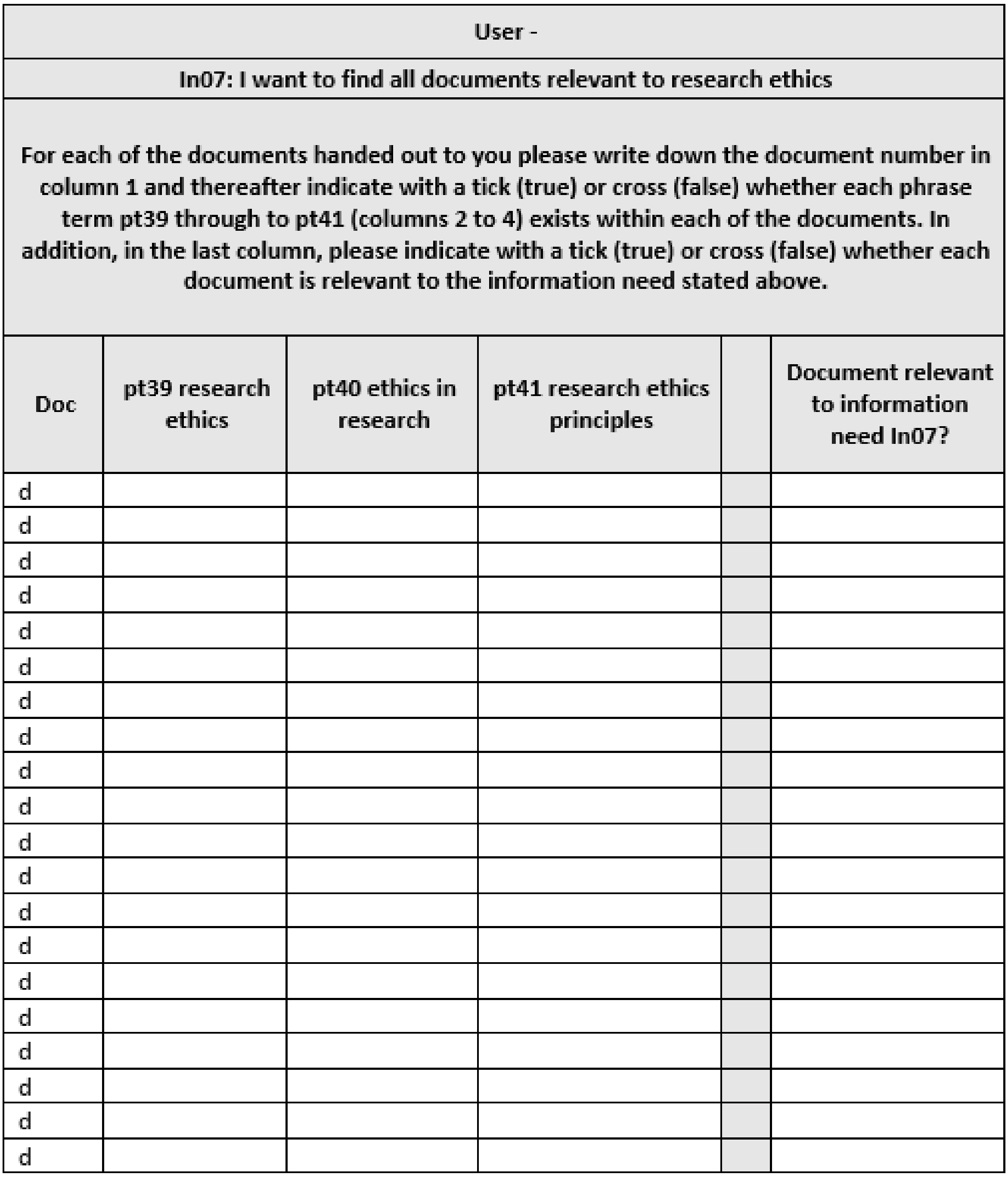


User -

In06: I want to find all documents relevant to conceptual frameworks

For each of the documents handed out to you please write down the document number in column 1 and thereafter indicate with a tick (true) or cross (false) whether each phrase term pt33 through to pt38 (columns 2 to 7) exists within each of the documents. In addition, in the last column, please indicate with a tick (true) or cross (false) whether each document is relevant to the information need stated above.

Doc	pt33 conceptual framework	pt34 conceptual frameworks	pt35 conceptual framework in research	pt36 conceptual frameworks in research	pt37 conceptual model	pt38 conceptual models		Document relevant to information need In06?
d								
d								
d								
d								
d								
d								
d								
d								
d								
d								
d								
d								
d								
d								
d								
d								
d								
d								
d								
d								

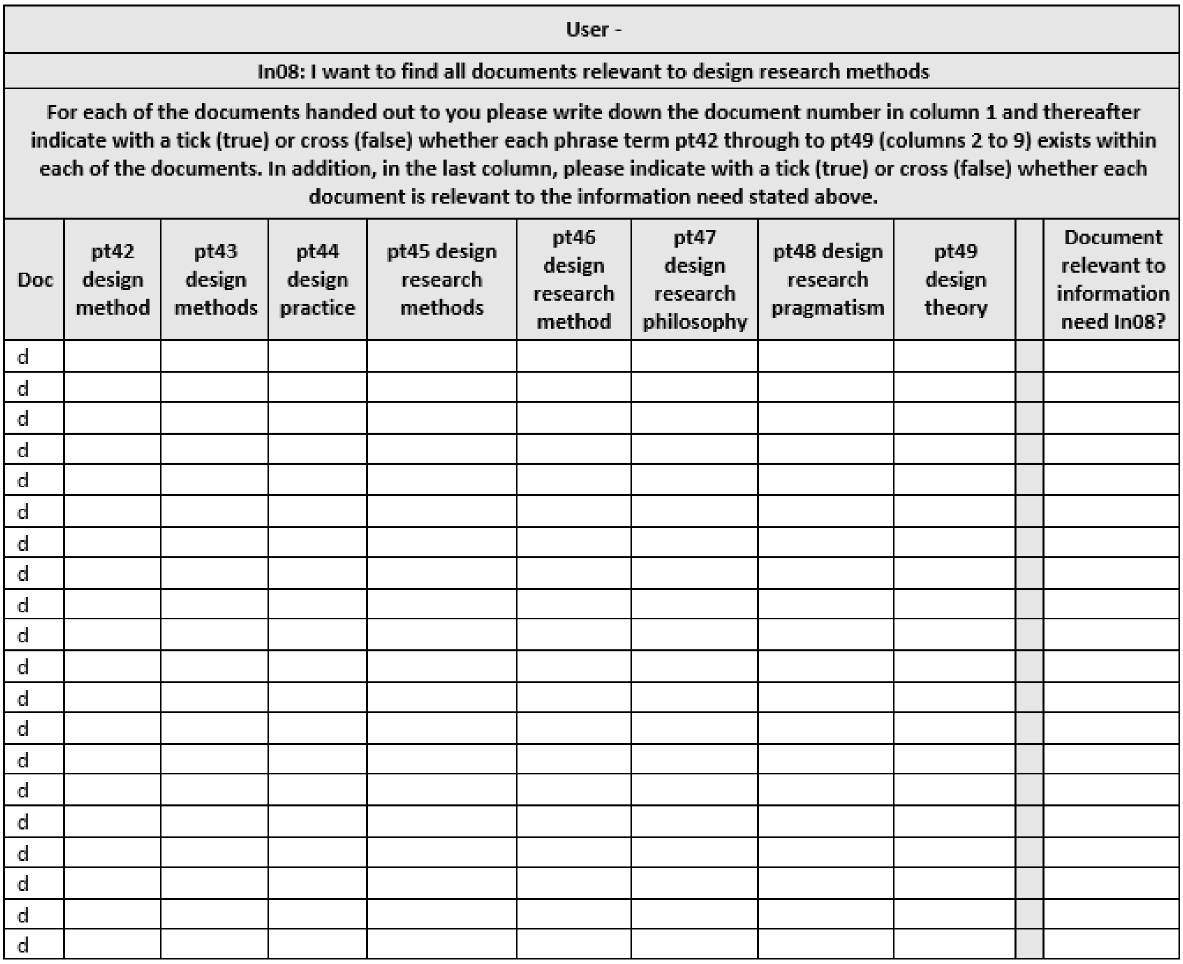


User -

In07: I want to find all documents relevant to research ethics

For each of the documents handed out to you please write down the document number in column 1 and thereafter indicate with a tick (true) or cross (false) whether each phrase term pt39 through to pt41 (columns 2 to 4) exists within each of the documents. In addition, in the last column, please indicate with a tick (true) or cross (false) whether each document is relevant to the information need stated above.

Doc	pt39 research ethics	pt40 ethics in research	pt41 research ethics principles		Document relevant to information need In07?
d					
d					
d					
d					
d					
d					
d					
d					
d					
d					
d					
d					
d					
d					
d					
d					
d					
d					
d					
d					

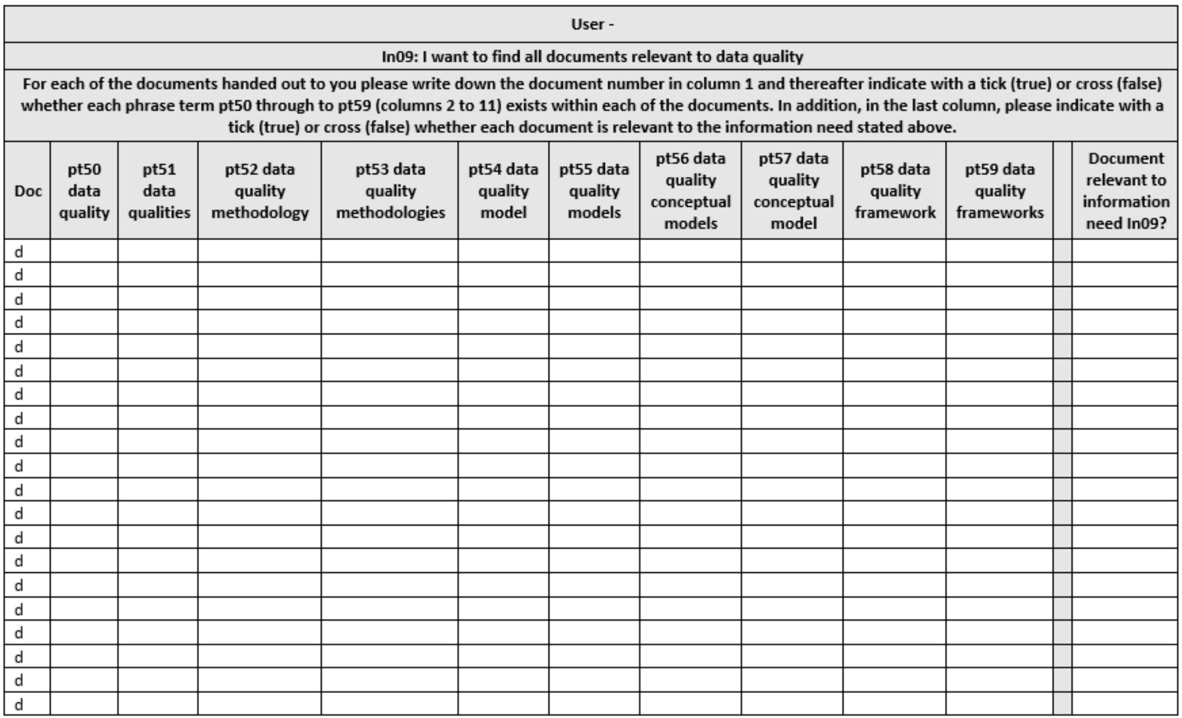


User -

In08: I want to find all documents relevant to design research methods

For each of the documents handed out to you please write down the document number in column 1 and thereafter indicate with a tick (true) or cross (false) whether each phrase term pt42 through to pt49 (columns 2 to 9) exists within each of the documents. In addition, in the last column, please indicate with a tick (true) or cross (false) whether each document is relevant to the information need stated above.

Doc	pt42 design method	pt43 design methods	pt44 design practice	pt45 design research methods	pt46 design research method	pt47 design research philosophy	pt48 design research pragmatism	pt49 design theory		Document relevant to information need In08?
d										
d										
d										
d										
d										
d										
d										
d										
d										
d										
d										
d										
d										
d										
d										
d										
d										
d										
d										
d										

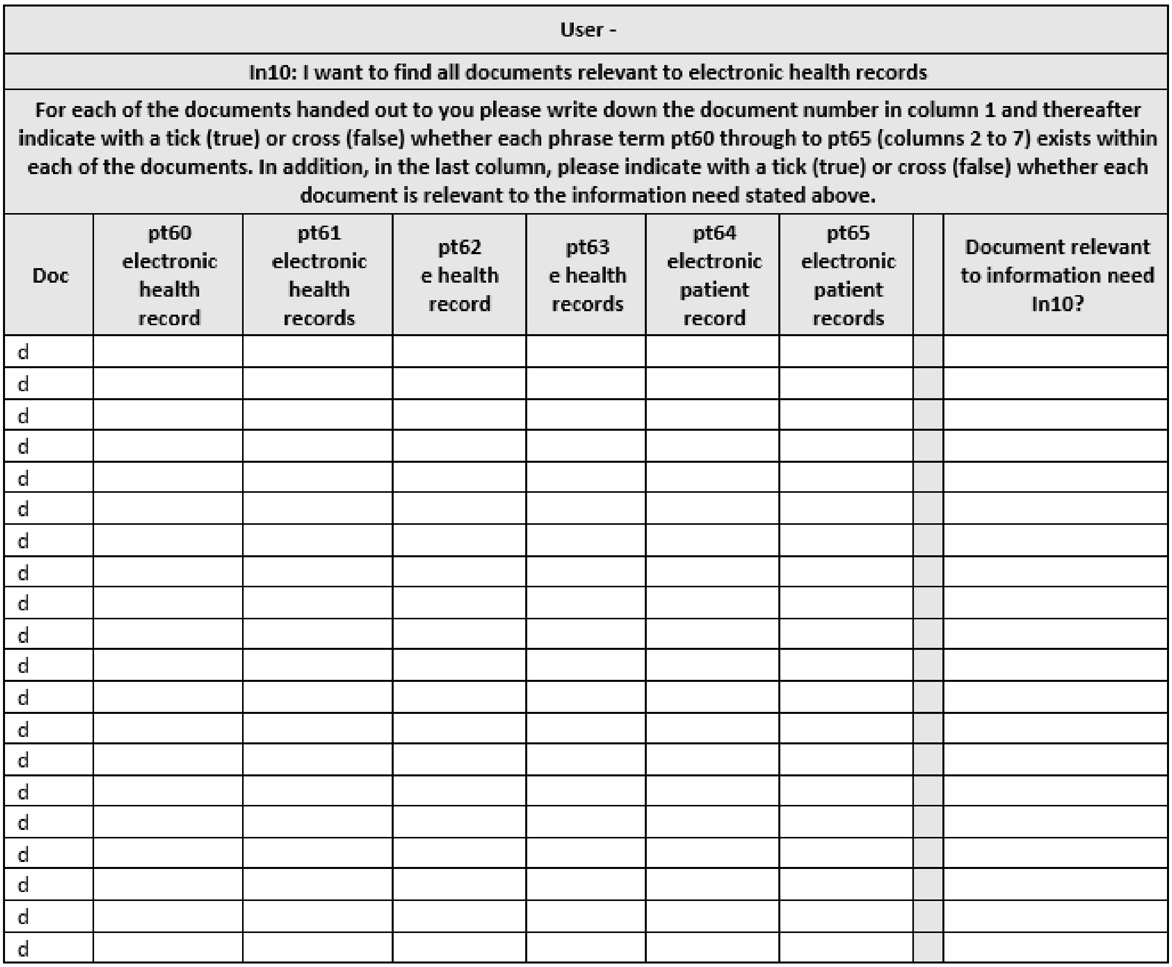


User -

In09: I want to find all documents relevant to data quality

For each of the documents handed out to you please write down the document number in column 1 and thereafter indicate with a tick (true) or cross (false) whether each phrase term pt50 through to pt59 (columns 2 to 11) exists within each of the documents. In addition, in the last column, please indicate with a tick (true) or cross (false) whether each document is relevant to the information need stated above.

Doc	pt50 data quality	pt51 data qualities	pt52 data quality methodology	pt53 data quality methodologies	pt54 data quality model	pt55 data quality models	pt56 data quality conceptual models	pt57 data quality conceptual model	pt58 data quality framework	pt59 data quality frameworks		Document relevant to information need In09?
d												
d												
d												
d												
d												
d												
d												
d												
d												
d												
d												
d												
d												
d												
d												
d												
d												
d												
d												
d												



User -

In10: I want to find all documents relevant to electronic health records

For each of the documents handed out to you please write down the document number in column 1 and thereafter indicate with a tick (true) or cross (false) whether each phrase term pt60 through to pt65 (columns 2 to 7) exists within each of the documents. In addition, in the last column, please indicate with a tick (true) or cross (false) whether each document is relevant to the information need stated above.

Doc	pt60 electronic health record	pt61 electronic health records	pt62 e health record	pt63 e health records	pt64 electronic patient record	pt65 electronic patient records		Document relevant to information need In10?
d								
d								
d								
d								
d								
d								
d								
d								
d								
d								
d								
d								
d								
d								
d								
d								
d								
d								
d								
d								

APPENDIX F: USER JUDGEMENT RESULTS

Table F.1: User information-need-by-document matrix results

doc	User	in01	in02	in03	in04	in05	in06	in07	in08	in09	in10
d0052	**A**	0	1	0	1	1	1	0	1	1	1
d0030	**A**	0	1	1	0	1	1	0	1	0	0
d0001	**A**	0	0	1	0	1	1	0	0	1	0
d0046	**A**	1	0	0	0	1	0	1	1	0	0
d0087	**A**	0	0	1	0	1	1	0	1	1	1
d0057	**A**	1	0	1	0	1	1	0	1	1	0
d0007	**A**	0	0	1	1	1	1	1	1	1	1
d0063	**A**	1	1	0	0	1	1	1	1	1	0
d0009	**A**	1	1	0	0	1	1	1	1	1	0
d0036	**A**	1	0	0	1	0	1	1	1	1	1
d0033	**A**	1	1	0	0	1	1	1	1	1	0
d0031	**A**	1	1	0	1	1	1	1	1	1	1
d0029	**A**	1	1	0	0	1	1	1	1	1	0
d0040	**A**	1	1	1	1	0	1	1	1	1	1
d0048	**A**	1	1	1	1	1	1	0	0	1	1
d0066	**A**	1	0	1	1	1	1	1	1	1	1
d0053	**A**	1	1	1	0	0	0	1	1	0	0
d0092	**A**	0	0	1	0	1	0	0	0	1	0
d0100	**A**	1	1	0	0	1	0	0	1	1	0
d0064	**A**	1	0	1	0	1	1	1	1	1	0
d0088	**B**	0	0	0	0	0	0	1	0	0	0
d0024	**B**	0	0	0	0	0	0	0	0	0	0
d0034	**B**	0	0	0	0	0	0	0	0	0	0
d0019	**B**	0	0	0	0	0	0	0	0	1	1
d0077	**B**	1	0	0	0	0	1	0	1	0	0
d0011	**B**	0	0	0	0	0	0	0	0	0	0
d0012	**B**	0	0	0	0	0	0	0	0	0	0
d0044	**B**	0	0	0	0	0	0	0	0	0	0
d0016	**B**	0	0	0	0	0	0	0	0	0	0
d0025	**B**	0	0	0	0	0	0	0	0	0	0
d0014	**B**	0	0	0	0	0	0	0	0	0	0
d0069	**B**	1	0	0	0	0	0	0	1	0	0
d0070	**B**	0	0	0	0	0	0	0	0	1	0
d0079	**B**	1	0	0	0	0	0	0	1	0	0
d0015	**B**	0	0	0	0	0	0	0	0	0	0
d0072	**B**	0	0	0	0	0	0	0	0	1	0
d0013	**B**	0	0	0	0	0	0	0	0	0	0
d0008	**B**	0	0	0	0	0	0	0	0	0	0
d0074	**B**	1	0	0	0	0	0	0	0	0	0
d0075	**B**	1	0	0	0	0	0	0	1	0	0
d0094	**C**	0	0	0	0	0	0	0	0	0	0
d0047	**C**	0	0	0	0	0	0	0	0	0	0
d0039	**C**	0	0	0	0	0	0	0	0	1	0
d0067	**C**	0	0	0	0	0	0	0	0	0	0
d0041	**C**	0	0	0	0	0	0	0	0	0	0
d0051	**C**	0	0	0	0	0	0	1	0	0	0
d0037	**C**	0	0	0	0	0	0	0	0	0	0
d0035	**C**	1	0	0	0	0	0	0	1	0	0
d0032	**C**	1	0	0	0	0	0	0	1	0	0
d0065	**C**	1	0	0	0	0	0	0	1	0	0
d0062	**C**	1	0	0	0	0	0	0	1	0	0
d0061	**C**	0	0	0	0	0	0	0	0	0	0
d0003	**C**	0	0	0	0	0	0	0	0	0	0
d0050	**C**	0	0	0	0	0	0	0	0	0	0
d0056	**C**	0	0	0	0	0	0	0	0	0	0
d0049	**C**	0	0	0	0	0	0	0	0	0	0
d0043	**C**	0	0	0	0	0	0	0	0	0	0
d0054	**C**	0	0	0	0	0	0	0	0	0	0
d0055	**C**	0	0	0	0	0	0	0	0	0	0
d0045	**C**	0	0	0	1	0	0	0	0	0	1
d0038	**D**	0	0	0	0	0	0	0	0	0	0

doc	User	in01	in02	in03	in04	in05	in06	in07	in08	in09	in10
d0020	**D**	0	0	0	0	0	0	0	0	0	0
d0018	**D**	0	0	0	0	0	0	0	0	1	0
d0090	**D**	0	0	0	0	0	0	0	0	0	0
d0089	**D**	1	0	0	0	0	0	1	0	0	0
d0086	**D**	0	0	0	0	0	0	0	0	0	0
d0085	**D**	0	0	0	1	0	0	0	0	0	1
d0084	**D**	0	0	0	0	0	0	0	0	0	0
d0083	**D**	0	0	0	0	0	0	0	0	0	0
d0023	**D**	0	0	0	0	0	0	0	0	0	0
d0082	**D**	0	0	0	0	0	1	0	0	0	1
d0080	**D**	0	0	0	0	0	0	0	0	0	0
d0022	**D**	0	0	0	0	0	0	0	0	0	0
d0096	**D**	0	0	0	0	0	0	0	0	0	0
d0004	**D**	0	0	0	0	0	0	0	0	0	0
d0095	**D**	0	0	0	0	0	0	0	0	0	0
d0091	**D**	0	0	0	0	0	0	0	0	0	0
d0099	**D**	0	0	0	0	0	0	0	0	0	0
d0098	**D**	0	0	0	0	0	0	0	0	0	0
d0093	**D**	0	0	0	0	0	0	0	0	0	0
d0017	**E**	0	0	0	0	0	0	0	0	0	0
d0026	**E**	0	0	0	0	0	0	0	0	0	0
d0081	**E**	0	1	0	0	0	0	0	0	0	0
d0021	**E**	1	0	0	0	0	0	0	0	0	0
d0078	**E**	1	0	0	0	0	0	0	0	0	0
d0073	**E**	0	0	0	0	0	1	0	0	0	0
d0010	**E**	0	0	0	0	0	0	0	0	0	0
d0059	**E**	0	0	0	0	0	0	0	0	0	0
d0060	**E**	0	0	0	0	0	0	0	0	0	0
d0006	**E**	0	0	0	0	0	0	0	0	0	0
d0097	**E**	0	0	0	0	0	0	0	0	0	0
d0005	**E**	0	0	0	0	0	1	0	0	0	0
d0028	**E**	0	0	0	0	0	0	0	0	0	0
d0042	**E**	0	0	0	0	0	0	0	0	0	0
d0076	**E**	1	0	0	0	0	0	0	1	0	0
d0068	**E**	0	0	0	0	0	0	0	1	0	0
d0071	**E**	0	0	0	0	0	1	0	0	1	0
d0027	**E**	0	0	0	0	0	0	0	0	0	0
d0058	**E**	0	0	0	0	0	0	0	1	0	0
d0002	**E**	1	0	0	0	0	0	0	1	0	0

Table F.2: User phrase-term-by-document matrix results – phrase-term pt01 to pt33

doc	User	pt01	pt02	pt03	pt04	pt05	pt06	pt07	pt08	pt09	pt10	pt11	pt12	pt13	pt14	pt15	pt16	pt17	pt18	pt19	pt20	pt21	pt22	pt23	pt24	pt25	pt26	pt27	pt28	pt29	pt30	pt31	pt32	pt33
d0052	**A**	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1
d0030	**A**	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	1	1	1	1	1
d0001	**A**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	0	0	0	0	0	0	1	1	1	1	1
d0046	**A**	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0
d0087	**A**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	0	0	0	0	0	0	1	1	1	1	1
d0057	**A**	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	1	1	1	1	1	1	1	0	0	0	0	0	0	1	1	1	1	1
d0007	**A**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
d0063	**A**	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1
d0009	**A**	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1
d0036	**A**	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	0	0	0	0	1
d0033	**A**	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1
d0031	**A**	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1
d0029	**A**	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1
d0040	**A**	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	1
d0048	**A**	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
d0066	**A**	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
d0053	**A**	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0
d0092	**A**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	0	0	0	0	0	0	1	1	1	1	0
d0100	**A**	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0
d0064	**A**	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	1
d0088	**B**	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0024	**B**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0034	**B**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0019	**B**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
d0077	**B**	1	0	1	1	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0011	**B**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0012	**B**	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0044	**B**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0016	**B**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0025	**B**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0014	**B**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0069	**B**	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0070	**B**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0079	**B**	1	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0015	**B**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0072	**B**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0013	**B**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0008	**B**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

doc	User	pt01	pt02	pt03	pt04	pt05	pt06	pt07	pt08	pt09	pt10	pt11	pt12	pt13	pt14	pt15	pt16	pt17	pt18	pt19	pt20	pt21	pt22	pt23	pt24	pt25	pt26	pt27	pt28	pt29	pt30	pt31	pt32	pt33
d0074	**B**	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0075	**B**	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0094	**C**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0047	**C**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0039	**C**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0067	**C**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0041	**C**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0051	**C**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0037	**C**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0035	**C**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0032	**C**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0065	**C**	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0062	**C**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0061	**C**	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0003	**C**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0050	**C**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0056	**C**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0049	**C**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0043	**C**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0054	**C**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0055	**C**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0045	**C**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0038	**D**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0020	**D**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0018	**D**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0090	**D**	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0089	**D**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0086	**D**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0085	**D**	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	1	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
d0084	**D**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0083	**D**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0023	**D**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0082	**D**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0080	**D**	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0022	**D**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0096	**D**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0004	**D**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0095	**D**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0091	**D**	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

doc	User	pt01	pt02	pt03	pt04	pt05	pt06	pt07	pt08	pt09	pt10	pt11	pt12	pt13	pt14	pt15	pt16	pt17	pt18	pt19	pt20	pt21	pt22	pt23	pt24	pt25	pt26	pt27	pt28	pt29	pt30	pt31	pt32	pt33
d0099	**D**	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0098	**D**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0093	**D**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0017	**E**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0026	**E**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0081	**E**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0021	**E**	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0078	**E**	1	1	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0073	**E**	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0010	**E**	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0059	**E**	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0060	**E**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0006	**E**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0097	**E**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0005	**E**	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0028	**E**	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0042	**E**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0076	**E**	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0068	**E**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0071	**E**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0027	**E**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0058	**E**	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0002	**E**	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table F.3: User phrase-term-by-document matrix results – phrase-term pt34 to pt65

doc	User	pt34	pt35	pt36	pt37	pt38	pt39	pt40	pt41	pt42	pt43	pt44	pt45	pt46	pt47	pt48	pt49	pt50	pt51	pt52	pt53	pt54	pt55	pt56	pt57	pt58	pt59	pt60	pt61	pt62	pt63	pt64	pt65
d0052	**A**	1	1	1	1	1	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
d0030	**A**	1	1	1	1	1	0	0	0	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0001	**A**	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0
d0046	**A**	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0087	**A**	1	1	1	1	1	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
d0057	**A**	1	1	1	1	1	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0
d0007	**A**	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
d0063	**A**	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0
d0009	**A**	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0
d0036	**A**	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
d0033	**A**	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0
d0031	**A**	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
d0029	**A**	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0
d0040	**A**	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
d0048	**A**	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
d0066	**A**	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
d0053	**A**	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0092	**A**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0
d0100	**A**	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0
d0064	**A**	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0
d0088	**B**	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0024	**B**	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0034	**B**	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0019	**B**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0077	**B**	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0011	**B**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0012	**B**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0044	**B**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0016	**B**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0025	**B**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0014	**B**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0069	**B**	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0070	**B**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0079	**B**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0015	**B**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0072	**B**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0	1	0	0	0	0	0	0	0
d0013	**B**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0008	**B**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

doc	User	pt34	pt35	pt36	pt37	pt38	pt39	pt40	pt41	pt42	pt43	pt44	pt45	pt46	pt47	pt48	pt49	pt50	pt51	pt52	pt53	pt54	pt55	pt56	pt57	pt58	pt59	pt60	pt61	pt62	pt63	pt64	pt65
d0074	**B**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0075	**B**	0	0	0	0	0	0	0	0	1	0	0	1	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0094	**C**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0047	**C**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0039	**C**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0067	**C**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0041	**C**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0051	**C**	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0037	**C**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0035	**C**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0032	**C**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0065	**C**	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0062	**C**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0061	**C**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0003	**C**	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0050	**C**	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
d0056	**C**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0049	**C**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
d0043	**C**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0054	**C**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0055	**C**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0045	**C**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
d0038	**D**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0020	**D**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0018	**D**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
d0090	**D**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0089	**D**	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0086	**D**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0085	**D**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	1	0
d0084	**D**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
d0083	**D**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0
d0023	**D**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0082	**D**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0
d0080	**D**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0022	**D**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0096	**D**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0004	**D**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0095	**D**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0091	**D**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

doc	User	pt34	pt35	pt36	pt37	pt38	pt39	pt40	pt41	pt42	pt43	pt44	pt45	pt46	pt47	pt48	pt49	pt50	pt51	pt52	pt53	pt54	pt55	pt56	pt57	pt58	pt59	pt60	pt61	pt62	pt63	pt64	pt65
d0099	**D**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0098	**D**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0093	**D**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0017	**E**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0026	**E**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0081	**E**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0021	**E**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0078	**E**	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0073	**E**	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0010	**E**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0059	**E**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0060	**E**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0006	**E**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0097	**E**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0005	**E**	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0028	**E**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0042	**E**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0076	**E**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0068	**E**	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0071	**E**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0027	**E**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0058	**E**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0002	**E**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

APPENDIX G: IRS-H JUDGEMENT RESULTS

Table G.1: IRS-H information-need-by-document matrix results

doc	in01	in02	in03	in04	in05	in06	in07	in08	in09	in10
d0001	0	0	0	0	0	1	0	0	1	0
d0002	1	0	0	0	0	0	0	1	0	0
d0003	0	0	0	0	0	1	0	0	0	0
d0004	0	1	0	0	1	0	0	0	0	0
d0005	1	0	0	0	0	0	0	0	0	0
d0006	0	0	0	0	0	1	0	0	0	0
d0007	0	0	0	0	0	0	0	0	0	1
d0008	0	0	0	0	0	0	0	0	1	0
d0009	1	1	0	0	0	0	0	0	0	0
d0010	1	1	0	0	0	0	0	0	0	0
d0011	0	0	0	0	0	0	0	0	0	0
d0012	0	1	0	0	0	0	0	0	0	0
d0013	0	0	0	0	0	1	0	0	0	0
d0014	0	0	0	0	0	0	0	0	0	0
d0015	0	1	0	0	0	0	0	0	0	1
d0016	0	1	0	0	0	0	0	0	0	0
d0017	0	0	0	0	0	0	0	0	0	0
d0018	0	0	0	0	0	0	0	0	1	0
d0019	0	0	0	0	0	1	0	0	1	0
d0020	0	0	0	0	0	0	0	0	1	0
d0021	1	1	0	0	0	0	0	1	0	0
d0022	0	0	0	0	0	0	0	0	0	0
d0023	0	0	0	0	0	0	0	0	0	0
d0024	0	0	0	0	0	0	0	0	0	0
d0025	0	0	0	0	0	1	0	0	0	0
d0026	0	0	0	0	0	0	0	0	0	0
d0027	0	0	0	0	0	1	0	0	0	0
d0028	1	0	0	0	0	0	0	1	0	0
d0029	0	0	0	0	0	0	0	0	0	0
d0030	0	0	0	0	0	0	0	0	0	0
d0031	0	1	0	1	0	0	0	0	0	1
d0032	0	0	0	0	0	0	0	0	0	1
d0033	0	1	0	0	0	0	0	1	0	0
d0034	0	1	0	0	0	0	0	1	0	0
d0035	1	0	0	0	0	1	0	0	0	0
d0036	0	0	0	0	0	1	0	0	0	0
d0037	0	0	0	0	0	0	0	0	0	0
d0038	0	0	0	0	0	1	0	0	0	0
d0039	0	0	0	0	0	0	0	0	0	0
d0040	0	0	0	0	0	0	1	0	0	0
d0041	0	0	0	0	0	0	0	0	0	0
d0042	0	0	0	0	0	0	0	0	0	0
d0043	1	1	0	0	0	0	0	0	0	0
d0044	0	0	0	0	0	1	0	0	0	1
d0045	0	0	0	1	0	0	0	1	0	1
d0046	0	0	0	0	0	0	0	0	0	0
d0047	0	0	0	0	0	0	0	0	0	1
d0048	0	1	0	0	0	0	0	0	1	1
d0049	0	0	0	0	1	0	0	0	0	1
d0050	0	0	0	0	0	1	0	0	1	1
d0051	0	0	0	0	0	1	0	0	1	0
d0052	0	0	0	0	0	0	0	0	1	0

doc	in01	in02	in03	in04	in05	in06	in07	in08	in09	in10
d0053	0	1	0	0	0	0	0	1	0	0
d0054	0	0	0	0	0	0	0	0	0	0
d0055	0	0	0	0	0	0	0	0	0	0
d0056	0	1	0	0	0	0	0	0	0	0
d0057	1	0	0	0	0	0	0	0	0	0
d0058	1	0	0	0	0	1	0	1	0	0
d0059	1	0	0	0	0	0	0	1	0	0
d0060	1	0	0	0	0	0	0	0	0	0
d0061	1	1	0	0	0	1	0	1	0	0
d0062	1	1	0	0	0	0	0	1	0	0
d0063	1	1	0	0	0	1	0	1	0	0
d0064	1	1	1	0	0	1	0	1	1	0
d0065	1	0	0	0	0	0	0	1	0	0
d0066	0	0	0	0	0	1	0	0	0	0
d0067	0	0	0	0	0	0	0	0	0	0
d0068	1	0	0	0	0	1	0	1	0	0
d0069	1	0	0	0	0	0	0	1	0	0
d0070	0	0	0	0	0	0	0	0	1	0
d0071	0	0	0	0	0	0	0	0	1	0
d0072	0	0	0	0	0	1	0	0	1	0
d0073	1	0	0	0	0	1	0	1	0	0
d0074	1	0	0	0	0	0	0	1	0	0
d0075	1	0	1	0	0	0	0	1	0	0
d0076	1	0	0	0	0	0	0	1	0	0
d0077	1	0	0	0	0	0	0	1	0	0
d0078	1	1	0	0	0	0	0	1	0	0
d0079	1	0	0	0	0	1	0	1	0	0
d0080	0	1	0	0	0	0	0	0	0	0
d0081	0	0	0	0	0	0	0	0	1	0
d0082	0	0	0	0	0	1	0	0	0	1
d0083	0	0	0	1	0	0	0	0	1	1
d0084	0	0	1	0	0	0	0	0	1	1
d0085	0	0	1	1	0	0	0	0	1	1
d0086	0	0	0	0	0	0	0	0	0	0
d0087	0	0	0	0	0	0	0	0	0	0
d0088	1	1	1	0	0	0	1	0	0	0
d0089	0	0	0	0	0	0	0	0	0	0
d0090	1	0	0	0	0	0	0	0	0	0
d0091	0	1	0	0	0	1	0	0	0	0
d0092	0	0	0	0	0	1	0	0	0	0
d0093	0	0	1	0	0	1	0	0	0	0
d0094	0	0	0	0	0	1	0	0	0	0
d0095	0	0	0	0	0	0	0	0	0	0
d0096	1	0	0	0	0	0	0	0	0	0
d0097	0	0	0	0	0	0	0	0	0	0
d0098	0	1	1	0	1	1	1	0	1	0
d0099	0	0	1	0	1	0	0	0	1	0
d0100	0	1	0	0	0	0	0	0	0	0

Table G.2: IRS-H phrase-term-by-document matrix results – phrase-term pt01 to pt33

doc	pt01	pt02	pt03	pt04	pt05	pt06	pt07	pt08	pt09	pt10	pt11	pt12	pt13	pt14	pt15	pt16	pt17	pt18	pt19	pt20	pt21	pt22	pt23	pt24	pt25	pt26	pt27	pt28	pt29	pt30	pt31	pt32	pt33
d0001	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0002	28	1	19	0	0	17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0003	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0004	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0
d0005	3	0	1	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0006	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
d0007	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0008	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0009	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0010	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0011	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0012	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0013	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
d0014	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0015	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0016	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0017	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0018	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0019	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
d0020	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0021	1	0	0	0	0	20	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0022	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0023	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0024	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0025	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0026	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0027	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
d0028	17	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0029	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0030	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0031	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0
d0032	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0033	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

doc	pt01	pt02	pt03	pt04	pt05	pt06	pt07	pt08	pt09	pt10	pt11	pt12	pt13	pt14	pt15	pt16	pt17	pt18	pt19	pt20	pt21	pt22	pt23	pt24	pt25	pt26	pt27	pt28	pt29	pt30	pt31	pt32	pt33
d0034	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0035	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0036	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
d0037	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0038	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
d0039	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0040	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0041	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0042	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0043	0	0	0	0	0	3	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0044	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
d0045	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
d0046	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0047	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0048	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0049	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
d0050	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0051	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0052	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0053	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0054	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0055	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0056	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0057	0	0	0	0	0	38	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0058	2	0	0	0	0	28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
d0059	21	0	11	0	0	52	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0060	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0061	0	0	0	0	0	2	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
d0062	2	0	0	0	0	1	0	0	0	0	2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0063	29	2	3	0	3	20	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0064	14	1	5	0	0	2	0	0	0	1	8	0	1	1	0	0	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
d0065	22	0	15	0	0	62	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0066	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
d0067	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

doc	pt01	pt02	pt03	pt04	pt05	pt06	pt07	pt08	pt09	pt10	pt11	pt12	pt13	pt14	pt15	pt16	pt17	pt18	pt19	pt20	pt21	pt22	pt23	pt24	pt25	pt26	pt27	pt28	pt29	pt30	pt31	pt32	pt33
d0068	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
d0069	2	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0070	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0071	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0072	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
d0073	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0074	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0075	76	0	28	0	0	24	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0076	22	0	11	0	0	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0077	26	1	18	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0078	71	4	56	0	0	16	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0079	39	1	25	0	0	17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
d0080	0	0	0	0	0	0	0	0	0	0	3	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0081	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0082	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0083	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0
d0084	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0085	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0
d0086	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0087	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0088	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0089	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0090	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0091	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
d0092	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0093	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
d0094	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0095	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0096	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0097	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0098	0	0	0	0	0	0	0	0	1	0	9	0	0	2	0	0	0	3	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
d0099	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
d0100	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table G.3: IRS-H phrase-term-by-document matrix results – phrase-term pt34 to pt65

doc	pt34	pt35	pt36	pt37	pt38	pt39	pt40	pt41	pt42	pt43	pt44	pt45	pt46	pt47	pt48	pt49	pt50	pt51	pt52	pt53	pt54	pt55	pt56	pt57	pt58	pt59	pt60	pt61	pt62	pt63	pt64	pt65
d0001	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	68	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0
d0002	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0003	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0004	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0005	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0006	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0007	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0
d0008	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0009	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0010	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0011	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0012	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0013	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0014	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0015	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
d0016	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0017	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0018	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	43	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0019	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	45	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0020	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0021	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0022	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0023	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0024	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0025	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0026	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0027	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0028	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0029	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0030	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0031	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	5	0	0	0	0
d0032	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
d0033	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

doc	pt34	pt35	pt36	pt37	pt38	pt39	pt40	pt41	pt42	pt43	pt44	pt45	pt46	pt47	pt48	pt49	pt50	pt51	pt52	pt53	pt54	pt55	pt56	pt57	pt58	pt59	pt60	pt61	pt62	pt63	pt64	pt65
d0034	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0035	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0036	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0037	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0038	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0039	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0040	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0041	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0042	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0043	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0044	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	1
d0045	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	6	0	0	6	5
d0046	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0047	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
d0048	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	3	0	0	0	2	0
d0049	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0
d0050	0	0	0	1	4	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
d0051	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	134	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0052	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0053	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0054	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0055	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0056	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0057	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0058	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	53	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0059	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0060	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0061	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0062	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	70	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0063	1	0	0	0	0	0	0	0	8	0	0	0	0	0	0	85	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0064	2	0	0	1	11	0	0	0	0	1	1	0	0	0	0	10	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0065	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0066	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0067	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

doc	pt34	pt35	pt36	pt37	pt38	pt39	pt40	pt41	pt42	pt43	pt44	pt45	pt46	pt47	pt48	pt49	pt50	pt51	pt52	pt53	pt54	pt55	pt56	pt57	pt58	pt59	pt60	pt61	pt62	pt63	pt64	pt65
d0068	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0069	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0070	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	34	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0071	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0072	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	236	0	5	3	0	0	0	0	3	0	0	0	0	0	0	0
d0073	0	0	0	90	35	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0074	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0075	0	0	0	0	0	0	0	0	4	3	0	0	1	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0076	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0077	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0078	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0079	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0080	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0081	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0082	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	4	0	0	0	0
d0083	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	11	4	0	0	0	0
d0084	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	0	0	0	0	0	0	0	0	0	13	2	0	0	0	0
d0085	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	3	3	0	0	3	5
d0086	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0087	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0088	0	0	0	0	0	27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0089	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0090	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0091	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0092	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0093	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0094	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0095	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0096	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0097	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0098	0	0	0	6	0	0	1	0	0	0	0	0	0	0	0	0	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0099	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

APPENDIX H: IRS-I JUDGEMENT RESULTS

Table H.1: IRS-I information-need-by-document matrix results

doc	in01	in02	in03	in04	in05	in06	in07	in08	in09	in10
d0001	1	1	1	1	1	1	1	1	1	1
d0002	1	1	1	1	1	1	1	1	1	1
d0003	1	1	1	1	1	1	1	1	1	1
d0004	1	1	1	1	1	1	1	1	1	1
d0005	1	1	1	1	1	1	1	1	1	1
d0006	1	1	1	1	1	1	1	1	1	1
d0007	1	1	1	1	1	1	1	1	1	1
d0008	1	1	1	1	1	1	1	1	1	1
d0009	1	1	1	1	1	1	1	1	1	1
d0010	1	1	1	1	1	1	1	1	1	1
d0011	1	0	0	1	0	1	1	1	0	0
d0012	1	1	1	1	1	1	1	1	1	1
d0013	1	1	1	1	1	1	1	1	1	1
d0014	1	1	1	1	1	1	1	1	1	1
d0015	1	1	1	1	1	1	1	1	1	1
d0016	1	1	1	1	1	1	1	1	1	1
d0017	1	1	1	1	0	1	1	1	1	0
d0018	1	1	1	1	1	1	1	1	1	0
d0019	1	1	1	1	1	1	1	1	1	1
d0020	1	1	1	1	1	1	1	1	1	1
d0021	1	1	1	1	1	1	1	1	1	1
d0022	1	1	1	1	1	1	1	1	1	1
d0023	1	1	1	1	1	1	1	1	1	0
d0024	1	1	1	1	1	1	1	1	1	1
d0025	1	1	1	1	1	1	1	1	1	1
d0026	1	1	1	1	0	1	1	1	1	1
d0027	1	1	1	1	1	1	1	1	1	1
d0028	1	1	1	1	1	1	1	1	1	1
d0029	1	1	1	1	1	1	1	1	1	1
d0030	1	1	1	1	1	1	1	1	1	1
d0031	1	1	1	1	1	1	1	1	1	1
d0032	1	1	1	1	1	1	1	1	1	1
d0033	1	1	1	1	1	1	1	1	1	1
d0034	1	1	1	1	1	1	1	1	1	1
d0035	1	1	1	1	1	1	1	1	1	1
d0036	1	1	1	1	1	1	1	1	1	1
d0037	1	1	1	1	1	1	1	1	1	1
d0038	1	1	1	1	1	1	1	1	1	1
d0039	1	1	1	1	1	1	1	1	1	1
d0040	1	1	1	1	1	1	1	1	1	1
d0041	1	1	1	1	1	1	1	1	1	1
d0042	1	1	1	1	1	1	1	1	1	1
d0043	1	1	1	1	1	1	1	1	1	1
d0044	1	1	1	1	1	1	1	1	1	1
d0045	1	1	1	1	1	1	1	1	1	1
d0046	1	1	1	1	1	1	1	1	1	1
d0047	1	1	1	1	1	1	1	1	1	1
d0048	1	1	1	1	1	1	1	1	1	1
d0049	1	1	1	1	1	1	1	1	1	1
d0050	1	1	1	1	1	1	1	1	1	1
d0051	1	1	1	1	1	1	1	1	1	1

doc	in01	in02	in03	in04	in05	in06	in07	in08	in09	in10
d0052	1	1	1	1	0	1	1	1	1	1
d0053	1	1	1	1	1	1	1	1	1	1
d0054	1	1	1	1	1	1	1	1	1	1
d0055	1	1	1	1	0	1	1	1	1	0
d0056	1	1	1	1	1	1	1	1	1	1
d0057	1	1	1	1	1	1	1	1	1	1
d0058	1	1	1	1	1	1	1	1	1	1
d0059	1	1	1	1	1	1	1	1	1	1
d0060	1	1	1	1	1	1	1	1	1	1
d0061	1	1	1	1	1	1	1	1	1	1
d0062	1	1	1	1	1	1	1	1	1	1
d0063	1	1	1	1	1	1	1	1	1	1
d0064	1	1	1	1	1	1	1	1	1	1
d0065	1	1	1	1	1	1	1	1	1	1
d0066	1	1	1	1	1	1	1	1	1	1
d0067	1	1	1	1	1	1	1	1	1	1
d0068	1	1	1	1	1	1	1	1	1	1
d0069	1	1	1	1	1	1	1	1	1	1
d0070	1	1	1	1	1	1	1	1	1	1
d0071	1	1	1	1	0	1	0	1	1	1
d0072	1	1	1	1	1	1	1	1	1	1
d0073	1	1	1	1	1	1	1	1	1	1
d0074	1	1	1	1	1	1	1	1	1	1
d0075	1	1	1	1	1	1	1	1	1	1
d0076	1	1	1	1	1	1	1	1	1	1
d0077	1	1	1	1	1	1	1	1	1	1
d0078	1	1	1	1	1	1	1	1	1	1
d0079	1	1	1	1	1	1	1	1	1	1
d0080	1	1	1	1	1	1	1	1	1	1
d0081	1	1	1	1	1	1	1	1	1	1
d0082	1	1	1	1	1	1	1	1	1	1
d0083	1	1	1	1	1	1	1	1	1	1
d0084	1	1	1	1	1	1	1	1	1	1
d0085	1	1	1	1	1	1	1	1	1	1
d0086	1	1	1	1	1	1	1	1	1	1
d0087	1	1	1	1	1	1	1	1	1	1
d0088	1	1	1	1	1	1	1	1	1	1
d0089	1	1	1	1	1	1	1	1	1	1
d0090	1	1	1	1	1	1	1	1	1	1
d0091	1	1	1	1	1	1	1	1	1	1
d0092	1	1	1	1	1	1	1	1	1	1
d0093	1	1	1	1	1	1	1	1	1	1
d0094	1	1	1	1	1	1	1	1	1	1
d0095	1	1	1	1	1	1	1	1	1	1
d0096	1	1	1	1	1	1	1	1	1	1
d0097	1	1	1	1	1	1	1	1	1	1
d0098	1	1	1	1	1	1	1	1	1	1
d0099	1	1	1	1	1	1	1	1	1	1
d0100	1	1	1	1	0	1	1	1	1	0

Table H.2: IRS-I term-by-document matrix results – term t01 to t33

doc	t01	t02	t03	t04	t05	t06	t07	t08	t09	t10	t11	t12	t13	t14	t15	t16	t17	t18	t19	t20	t21	t22	t23	t24	t25	t26	t27	t28	t29	t30	t31	t32	t33
d0001	5	0	0	0	1	6	210	4	7	2	0	0	124	1	0	0	1	0	135	9	16	1	3	9	29	10	6	0	0	0	0	0	0
d0002	16	2	0	0	4	5	60	141	12	0	0	0	235	13	0	1	5	20	306	27	28	6	62	15	33	16	2	9	5	0	0	18	0
d0003	4	0	0	0	1	2	67	1	3	4	0	0	103	18	2	1	2	0	108	38	0	0	1	0	5	0	5	0	0	0	0	4	0
d0004	18	2	1	3	3	3	13	4	15	2	0	0	139	79	0	1	5	2	168	9	12	0	8	8	2	1	0	2	0	0	0	2	0
d0005	3	0	0	0	1	0	6	24	7	0	0	0	67	23	1	0	0	1	130	24	9	0	0	1	6	1	2	0	0	0	0	15	0
d0006	2	0	0	0	0	2	4	2	26	0	0	3	52	9	0	0	1	0	129	5	5	1	0	5	0	0	0	1	1	0	0	5	0
d0007	0	19	12	0	0	0	8	0	9	9	0	1	48	1	0	1	18	42	69	6	7	0	0	0	0	0	0	0	0	40	0	0	0
d0008	2	0	0	0	0	0	18	0	12	1	0	10	58	4	0	0	0	0	113	7	2	0	0	1	5	0	1	0	0	0	0	16	0
d0009	10	0	0	0	5	1	3	8	12	1	0	0	54	2	0	0	0	0	244	13	1	0	0	0	13	3	0	0	0	0	1	15	0
d0010	10	0	0	0	4	1	3	7	13	1	0	0	53	2	0	0	0	0	239	9	1	0	0	0	13	2	0	0	0	0	1	15	0
d0011	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d0012	4	126	0	0	0	0	4	15	15	0	1	0	82	1	0	0	0	1	144	9	2	0	1	2	0	0	0	0	0	2	0	4	0
d0013	2	2	0	0	0	1	0	1	4	1	0	1	67	2	1	0	0	2	217	2	4	0	1	2	0	2	0	3	1	0	1	20	0
d0014	8	0	0	0	0	1	3	2	4	0	1	1	56	2	1	0	0	0	113	7	0	0	8	3	1	0	0	0	1	0	0	3	0
d0015	12	0	1	0	0	9	0	2	19	1	0	0	69	2	0	0	0	2	253	5	3	0	1	5	0	3	0	4	0	0	2	26	0
d0016	4	0	0	0	0	5	0	3	7	0	0	0	124	4	2	0	0	1	241	7	0	7	0	1	1	5	2	15	4	0	1	9	0
d0017	1	0	0	0	0	0	4	0	0	0	0	0	3	0	0	0	0	0	6	4	0	0	0	0	2	0	0	0	0	0	0	0	0
d0018	1	0	0	0	0	0	110	0	0	0	0	0	42	0	0	0	1	0	36	4	0	1	0	0	2	2	2	0	0	0	0	0	0
d0019	5	4	0	0	3	3	265	9	47	0	0	0	129	4	1	0	0	12	225	12	0	1	1	1	6	6	1	3	0	1	0	2	0
d0020	31	0	0	1	0	0	196	9	1	0	0	1	90	2	0	0	1	0	95	14	0	0	2	0	35	24	20	2	0	0	0	1	0
d0021	10	1	0	0	0	0	0	272	18	0	1	1	80	21	2	0	0	0	170	170	1	2	6	10	2	2	1	47	23	0	1	12	0
d0022	3	0	0	0	0	0	3	1	6	0	1	0	29	29	0	0	1	0	92	3	4	6	11	4	6	1	0	56	23	0	1	0	0
d0023	3	0	0	0	0	0	0	0	0	0	0	0	10	2	0	0	0	0	15	0	5	0	0	6	2	4	0	15	9	0	0	0	0
d0024	0	0	0	0	0	0	0	0	2	0	0	0	2	0	0	0	0	0	4	0	0	0	0	0	5	0	0	0	0	0	0	0	0
d0025	43	0	0	0	1	1	21	10	8	2	0	0	161	2	0	1	4	0	265	53	2	37	85	2	10	12	0	1	0	0	0	3	0
d0026	5	2	0	0	0	0	0	0	1	0	0	0	34	1	0	0	0	2	56	0	0	0	0	0	0	0	0	1	0	0	0	1	0
d0027	13	0	0	0	0	5	0	5	6	0	5	3	98	11	0	0	0	3	212	0	0	1	0	0	0	0	0	4	2	0	17	3	0
d0028	5	39	0	0	0	0	11	56	6	1	0	1	90	0	2	0	4	18	148	6	0	1	1	6	7	2	0	0	0	26	0	6	0
d0029	4	0	0	0	0	3	15	7	1	1	2	0	79	4	0	0	0	0	157	2	1	1	1	2	1	0	0	0	0	0	0	6	0
d0030	13	0	0	0	2	0	5	2	2	0	0	0	24	0	0	0	3	0	76	4	33	0	4	32	0	0	0	0	0	0	0	5	0
d0031	13	53	37	0	0	0	27	7	17	24	1	28	72	8	0	0	4	24	161	11	3	0	1	4	4	1	1	0	0	35	0	33	0
d0032	0	0	0	0	1	0	1	10	8	1	0	0	19	0	0	0	0	0	68	1	2	0	1	0	2	0	0	1	0	1	1	1	0
d0033	34	0	0	0	1	1	8	28	22	2	0	0	44	13	0	0	0	0	170	17	1	1	3	5	5	10	1	0	0	0	0	9	0

doc	t01	t02	t03	t04	t05	t06	t07	t08	t09	t10	t11	t12	t13	t14	t15	t16	t17	t18	t19	t20	t21	t22	t23	t24	t25	t26	t27	t28	t29	t30	t31	t32	t33
d0034	34	0	0	0	1	1	8	28	22	2	0	0	44	13	0	0	0	0	170	17	1	1	3	5	5	10	1	0	0	0	0	9	0
d0035	6	1	0	0	0	6	25	41	20	2	0	0	145	3	0	0	0	1	339	77	0	0	2	3	20	6	15	3	0	0	0	8	0
d0036	12	79	9	0	0	6	3	6	15	0	0	36	271	13	1	0	1	596	313	10	1	0	4	18	19	19	0	1	0	6	2	32	0
d0037	4	0	0	0	1	1	0	0	22	1	0	0	51	1	0	0	0	1	131	0	1	0	0	0	5	0	0	1	0	0	0	1	0
d0038	0	14	0	0	0	2	3	0	56	0	0	4	87	4	0	0	0	142	73	0	0	0	0	4	12	4	0	0	0	0	0	2	0
d0039	12	5	0	0	0	0	15	0	7	0	0	10	69	2	0	1	0	69	216	0	0	1	0	3	3	1	0	1	0	0	2	2	1
d0040	0	0	1	0	0	1	1	0	3	0	10	0	65	14	0	0	5	35	101	0	0	1	0	2	0	1	0	0	0	0	0	4	0
d0041	1	2	4	0	0	0	1	1	5	0	0	0	20	8	2	0	1	4	94	8	1	0	0	0	2	1	0	1	0	0	0	66	0
d0042	5	2	0	0	0	0	7	9	15	1	3	1	55	0	0	0	0	0	195	36	0	0	2	3	3	7	4	0	0	0	0	3	0
d0043	5	1	0	0	0	1	11	27	9	0	0	0	61	1	1	0	0	0	156	4	3	2	8	8	1	2	0	1	1	0	0	18	0
d0044	6	73	3	0	65	1	4	16	18	5	0	0	77	13	1	0	0	71	177	0	0	0	2	2	3	2	0	0	2	27	0	0	0
d0045	11	117	47	0	27	4	28	68	53	43	1	8	278	4	1	0	13	126	736	15	9	0	2	13	6	12	3	2	0	66	0	36	0
d0046	4	1	0	0	5	9	0	13	18	0	3	0	73	1	0	0	0	0	223	6	0	0	0	6	2	1	4	1	0	0	6	123	0
d0047	0	1	3	0	0	0	1	0	0	2	0	0	8	0	0	0	0	4	25	0	0	0	0	0	1	0	0	0	0	0	0	0	0
d0048	23	38	40	0	8	0	28	18	30	8	0	0	94	2	2	1	11	158	138	60	0	0	1	25	12	2	9	0	4	11	0	12	0
d0049	13	29	32	1	4	0	8	12	7	7	0	0	69	1	0	0	0	76	164	28	0	2	1	4	0	0	0	0	0	13	0	8	0
d0050	13	26	56	0	2	15	463	6	31	2	0	0	175	5	1	0	0	33	263	14	3	2	0	15	19	44	4	1	0	21	0	8	0
d0051	1	0	0	0	1	6	234	1	16	0	0	0	103	2	0	0	0	0	99	18	1	3	1	2	5	1	0	0	0	0	0	0	0
d0052	1	0	10	0	0	0	41	0	1	0	0	0	6	0	0	0	0	0	13	0	0	0	0	7	0	0	0	0	0	0	0	0	0
d0053	11	0	0	0	1	1	13	29	7	10	0	0	157	11	0	3	8	0	202	25	46	0	8	5	1	0	0	0	0	0	0	3	0
d0054	11	0	0	14	0	0	5	26	9	0	0	0	74	15	29	0	1	0	149	0	0	1	4	3	3	10	0	1	0	0	0	28	0
d0055	0	1	0	0	0	1	0	27	0	0	0	0	7	0	0	0	0	0	33	0	0	1	1	0	0	0	0	0	1	0	2	2	0
d0056	5	0	0	0	0	0	4	67	7	1	1	0	30	0	0	0	0	0	136	0	1	0	2	6	1	0	0	0	0	0	0	2	0
d0057	8	0	0	0	1	2	1	105	11	5	0	8	81	5	0	0	0	0	169	9	3	0	0	12	1	4	1	3	5	2	3	4	0
d0058	6	0	0	0	2	8	1	327	69	0	0	3	204	3	1	0	0	0	308	0	21	0	0	4	2	2	3	0	1	0	23	4	1
d0059	4	0	0	0	0	1	1	125	8	0	0	0	81	0	1	4	3	0	140	3	2	0	1	7	0	3	0	22	0	0	4	13	78
d0060	2	0	0	0	0	0	1	21	6	0	1	0	18	0	0	0	0	0	40	0	5	0	0	1	0	1	0	0	0	0	16	7	0
d0061	9	0	0	0	1	5	10	98	5	0	0	3	65	4	0	0	5	0	63	26	6	0	2	18	6	13	15	0	0	0	0	0	0
d0062	13	0	0	0	0	16	12	178	16	1	0	1	94	3	0	0	0	0	155	3	1	0	1	6	8	2	0	0	0	0	4	8	2
d0063	11	0	0	0	2	5	10	355	11	3	1	0	233	16	3	0	8	0	369	52	18	6	9	18	19	12	1	3	2	0	18	16	0
d0064	305	12	2	0	29	116	256	258	170	7	7	16	922	54	12	0	3	11	2292	260	89	22	54	46	205	70	4	21	12	58	6	103	0
d0065	2	0	0	0	0	5	4	128	13	0	0	0	51	6	0	0	0	0	144	3	0	0	1	0	0	6	0	3	0	0	1	72	8
d0066	2	17	1	0	0	1	72	5	20	0	0	1	239	2	0	0	1	118	516	2	5	1	9	0	0	0	0	0	0	17	0	2	0
d0067	0	1	0	0	0	0	3	0	0	0	0	0	3	0	0	0	0	5	7	0	0	0	1	0	0	0	0	0	0	0	0	0	0
d0068	3	0	0	0	0	1	13	141	20	1	1	0	47	2	0	0	0	0	131	33	1	1	0	5	7	1	0	1	1	0	0	8	0

doc	t01	t02	t03	t04	t05	t06	t07	t08	t09	t10	t11	t12	t13	t14	t15	t16	t17	t18	t19	t20	t21	t22	t23	t24	t25	t26	t27	t28	t29	t30	t31	t32	t33
d0069	3	1	0	4	0	0	7	137	12	1	0	0	94	2	3	0	4	0	170	5	4	0	2	26	5	4	0	4	0	0	0	6	0
d0070	10	0	0	0	0	0	66	5	23	0	0	1	117	7	0	0	0	2	177	27	3	0	1	0	38	3	1	0	0	0	0	2	0
d0071	0	1	3	0	0	0	15	1	0	0	0	0	1	0	0	0	0	1	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0
d0072	35	1	0	0	53	8	695	19	12	2	0	0	259	22	0	0	22	9	501	59	6	157	175	12	33	9	3	1	0	0	0	1	0
d0073	17	2	1	0	0	185	103	25	28	1	0	0	151	65	63	0	15	3	340	35	10	2	6	14	142	142	0	5	0	0	0	78	0
d0074	0	0	0	0	0	0	1	31	5	0	0	0	7	0	0	0	2	0	18	0	2	0	0	5	3	1	0	0	0	0	0	0	0
d0075	25	0	0	0	0	1	5	254	2	7	0	0	60	33	3	0	0	0	159	14	13	0	8	23	32	0	1	2	1	0	3	2	0
d0076	6	0	0	0	0	0	9	99	18	1	0	0	81	45	1	1	3	0	148	6	8	3	3	20	12	4	0	1	0	0	0	2	0
d0077	7	1	0	0	0	0	0	95	24	0	1	0	84	59	1	1	0	0	163	6	42	1	6	52	2	1	0	3	0	0	0	5	0
d0078	2	0	0	0	1	1	4	154	11	1	0	0	75	26	22	2	4	0	153	13	3	0	0	9	1	2	0	12	0	0	16	5	6
d0079	21	4	0	0	3	6	61	170	15	0	0	0	221	14	0	1	5	23	314	24	31	7	60	15	36	15	2	9	5	0	0	19	0
d0080	6	7	0	0	0	0	12	8	8	1	1	0	92	4	0	1	2	32	178	71	3	24	13	19	48	10	0	0	0	1	0	2	0
d0081	21	21	0	0	0	1	77	3	6	21	0	0	91	1	0	0	0	150	221	15	0	0	0	0	0	1	1	0	0	6	0	36	0
d0082	4	73	4	0	0	13	68	2	14	18	0	3	70	1	0	0	5	89	57	4	0	0	0	1	39	2	0	0	0	37	0	0	0
d0083	10	50	59	0	0	3	148	3	10	30	0	4	124	4	0	1	16	242	125	8	0	0	0	11	24	2	0	0	0	42	0	7	0
d0084	16	16	8	0	0	1	22	7	14	59	0	0	51	9	0	2	1	37	118	12	2	0	0	6	26	3	0	0	0	12	0	2	0
d0085	12	118	26	0	0	0	95	1	37	44	0	1	70	3	0	0	5	95	236	16	10	0	1	15	8	0	0	0	0	84	0	10	0
d0086	9	0	0	0	0	0	9	0	5	1	0	0	44	0	0	0	0	1	93	1	0	0	0	2	6	0	0	0	0	0	0	0	0
d0087	106	0	0	0	1	0	49	0	54	0	0	0	227	0	0	0	0	2	420	3	12	0	0	13	9	7	0	0	0	0	0	0	0
d0088	4	2	3	0	10	1	7	10	18	0	92	3	123	6	1	0	3	19	193	2	7	0	3	13	0	0	0	0	0	3	1	18	0
d0089	0	2	0	0	0	0	0	39	13	0	7	0	49	5	2	7	2	1	95	0	0	0	1	0	1	1	0	0	0	0	1	1	0
d0090	3	0	0	0	0	0	6	34	2	2	0	0	75	1	0	0	4	1	117	4	5	0	1	2	2	1	0	0	0	0	0	6	0
d0091	9	0	0	0	0	20	40	10	21	2	0	1	111	10	5	0	1	3	183	7	19	3	3	45	10	3	0	1	0	0	0	6	0
d0092	14	0	0	0	0	4	10	1	3	5	0	0	35	0	0	0	1	0	75	31	4	0	0	2	48	4	0	0	1	0	0	1	0
d0093	133	5	0	0	1	2	151	43	212	32	0	11	494	268	60	1	4	48	762	22	154	2	37	100	97	25	2	0	1	2	1	26	0
d0094	11	0	0	0	0	1	5	11	24	1	0	0	62	4	0	0	0	1	229	10	2	0	3	2	2	6	1	0	0	0	0	8	0
d0095	3	0	0	0	0	0	7	3	19	0	0	0	31	2	0	0	1	0	84	0	3	0	1	2	11	15	1	1	0	0	0	0	0
d0096	37	1	0	0	1	1	7	9	29	6	0	0	168	3	0	0	1	2	589	29	6	0	2	7	2	2	0	1	3	1	5	31	0
d0097	3	0	0	0	0	0	4	0	3	1	0	0	23	0	0	0	0	0	121	53	0	0	1	1	27	6	0	0	0	0	0	0	0
d0098	29	0	0	2	1	8	84	7	15	1	8	1	181	16	2	0	0	1	692	96	8	3	17	20	28	12	7	0	1	0	20	2	1
d0099	46	0	0	2	7	0	755	18	14	1	0	0	347	12	11	2	9	0	627	17	1	0	4	5	5	7	3	0	1	0	0	0	0
d0100	9	0	0	0	0	0	33	1	0	0	0	1	17	0	1	0	0	0	88	0	0	0	0	26	0	0	0	1	0	0	0	0	0

Table H.3: IRS-I term-by-document matrix results – term t34 through to t49

doc	t34	t35	t36	t37	t38	t39	t40	t41	t42	t43	t44	t45	t46	t47	t48	t49
d0001	0	1	1	0	100	6	1	0	6	2	0	0	0	472	0	5
d0002	0	9	2	0	6	2	0	0	323	45	15	2	0	772	28	1
d0003	2	2	0	0	1	0	0	0	5	0	0	6	0	397	4	1
d0004	1	2	11	0	4	2	0	0	397	3	1	0	0	479	36	6
d0005	0	0	1	0	0	0	0	0	36	8	0	0	0	353	0	7
d0006	0	4	2	0	5	1	0	0	18	3	0	8	0	406	10	21
d0007	0	2	0	0	5	0	24	5	3	0	0	1	0	257	0	0
d0008	0	0	0	0	4	0	0	2	15	3	0	1	0	273	1	8
d0009	1	0	1	0	0	0	0	0	19	14	0	0	0	520	74	2
d0010	1	0	1	0	0	0	0	0	15	13	0	0	0	497	73	1
d0011	0	0	0	0	0	0	0	0	0	0	0	0	0	8	8	0
d0012	0	3	1	0	17	0	1	0	19	2	1	1	0	449	13	0
d0013	0	0	0	0	0	0	0	0	1	26	3	0	0	467	54	0
d0014	1	0	0	0	0	0	0	0	3	3	1	1	0	359	10	0
d0015	1	1	1	0	0	0	0	2	40	13	1	0	0	527	99	3
d0016	0	9	1	0	0	0	0	0	6	11	3	0	0	677	55	4
d0017	0	0	0	0	0	0	0	0	2	0	0	0	0	9	0	0
d0018	1	0	0	1	52	0	0	0	5	0	0	1	0	146	0	0
d0019	0	0	2	0	224	1	3	1	8	1	1	0	0	658	0	2
d0020	0	1	4	0	9	0	1	0	6	1	1	2	0	321	0	3
d0021	0	1	1	1	1	0	0	0	120	14	3	6	0	408	23	1
d0022	4	0	15	0	0	15	0	0	75	1	0	0	0	393	12	2
d0023	3	0	0	0	0	0	0	0	7	3	3	0	0	72	7	0
d0024	0	0	0	0	0	0	0	0	2	0	0	0	0	15	4	1
d0025	0	1	0	0	6	1	0	0	16	3	0	1	0	769	2	0
d0026	0	0	0	0	2	1	0	0	1	1	0	0	0	176	0	0
d0027	7	5	0	0	4	1	1	0	7	4	1	0	0	510	19	1
d0028	1	0	0	0	4	0	6	3	38	19	1	5	0	534	2	0
d0029	0	0	0	4	22	0	0	0	33	1	0	2	0	367	1	2
d0030	0	0	2	0	0	1	0	0	103	1	0	1	0	191	4	2
d0031	23	1	9	0	17	0	7	13	20	0	1	0	0	555	8	1
d0032	0	0	0	0	0	0	0	1	1	0	0	0	0	138	0	0
d0033	0	0	2	1	0	0	0	0	13	16	1	0	0	261	4	1
d0034	0	0	2	1	0	0	0	0	13	16	1	0	0	261	4	1
d0035	6	0	3	0	4	0	0	0	100	10	2	8	0	984	4	0
d0036	16	7	1	0	13	0	0	0	41	9	2	23	0	516	8	7
d0037	1	0	0	0	2	0	0	0	5	2	1	1	0	361	1	0
d0038	0	0	0	0	7	0	0	0	12	7	3	4	0	106	2	1
d0039	6	1	1	0	1	1	0	0	73	12	2	8	0	491	0	0
d0040	1	3	0	0	1	0	1	0	85	8	0	1	0	222	0	0
d0041	0	0	1	1	1	1	0	0	22	0	0	2	0	201	9	0
d0042	1	2	0	0	2	0	0	0	96	3	7	1	0	367	39	2
d0043	7	0	1	1	1	0	1	1	44	1	0	1	0	321	3	2
d0044	0	2	0	0	1	1	6	1	47	5	5	1	0	319	2	1
d0045	9	3	6	0	9	1	54	43	92	2	0	4	2	1208	7	4
d0046	0	19	0	0	2	0	0	0	9	6	21	0	0	574	30	0
d0047	0	0	0	0	2	1	1	0	4	1	3	1	0	13	0	0
d0048	4	20	4	0	10	2	7	3	5	11	6	3	0	310	8	1
d0049	1	13	0	0	4	2	4	7	1	6	6	2	0	336	4	0
d0050	18	3	0	0	18	0	20	28	8	2	3	4	0	542	1	3
d0051	0	1	0	0	168	0	0	0	1	0	0	0	0	337	0	2
d0052	0	0	0	0	26	1	0	0	0	1	0	0	0	45	0	0
d0053	0	5	5	0	1	2	1	0	60	1	6	0	0	640	15	1
d0054	0	5	1	0	0	1	0	0	30	1	1	0	0	334	21	8

doc	t34	t35	t36	t37	t38	t39	t40	t41	t42	t43	t44	t45	t46	t47	t48	t49
d0055	0	0	0	1	0	0	0	0	3	0	0	0	0	67	0	0
d0056	0	1	1	0	0	0	0	0	38	4	2	1	0	275	4	0
d0057	0	0	0	0	4	0	0	0	99	4	0	0	0	512	1	8
d0058	3	5	0	0	0	0	0	0	37	8	1	0	0	828	78	2
d0059	0	2	0	0	1	0	0	0	100	27	3	0	0	190	17	12
d0060	0	0	0	0	1	0	0	0	1	25	2	0	0	96	0	0
d0061	0	0	2	6	2	0	0	0	27	4	0	181	0	328	6	0
d0062	0	7	2	0	0	0	0	0	28	6	2	0	0	342	121	7
d0063	3	47	2	0	0	1	0	0	121	65	21	0	0	924	283	4
d0064	9	20	34	15	55	24	10	4	541	63	23	9	0	5841	462	100
d0065	2	2	1	0	0	0	0	0	113	27	1	0	0	216	25	3
d0066	57	0	1	0	46	7	1	6	16	1	0	94	0	932	0	1
d0067	1	0	0	0	2	1	0	0	2	0	0	10	0	25	0	0
d0068	0	3	0	0	4	0	0	0	23	8	0	59	2	310	2	2
d0069	0	6	0	0	0	0	0	0	13	9	1	156	0	440	3	0
d0070	1	0	0	0	180	0	0	0	41	6	0	1	0	535	32	6
d0071	0	0	0	0	4	0	0	0	0	0	8	0	0	8	0	0
d0072	1	7	7	3	536	9	24	6	17	3	1	7	0	1458	8	48
d0073	0	14	2	0	437	3	0	0	153	10	0	2	0	569	10	7
d0074	0	2	0	0	0	0	0	0	2	18	0	24	0	49	0	0
d0075	0	2	5	0	1	5	0	0	91	91	4	20	0	432	14	1
d0076	2	0	0	0	23	1	0	1	52	29	1	1	0	383	11	0
d0077	1	1	2	1	5	2	0	0	54	31	1	0	0	376	11	0
d0078	3	0	2	0	2	2	0	0	148	99	17	0	0	400	31	9
d0079	0	9	2	0	8	2	0	0	317	57	14	2	0	800	33	1
d0080	3	11	6	0	115	1	1	1	14	0	0	2	0	419	0	0
d0081	5	0	0	0	10	0	2	1	12	4	0	0	0	948	2	0
d0082	26	0	0	0	5	0	19	13	1	0	0	12	0	213	0	0
d0083	4	0	0	0	18	0	27	19	19	6	4	4	0	327	0	6
d0084	0	1	0	0	39	2	19	5	43	3	2	12	0	184	3	1
d0085	16	1	2	0	39	2	50	51	28	0	1	6	0	470	0	5
d0086	0	0	0	0	0	0	0	0	2	1	0	1	0	290	0	0
d0087	0	0	0	0	3	0	0	0	11	2	0	0	0	1211	0	2
d0088	3	5	5	0	1	7	0	0	218	6	4	0	0	399	5	1
d0089	0	5	0	2	7	0	0	0	10	1	0	3	0	256	0	1
d0090	5	35	0	0	2	0	2	1	35	2	1	2	0	233	6	0
d0091	8	0	24	0	4	22	0	0	35	2	1	4	0	512	10	7
d0092	3	0	0	0	37	1	0	0	20	1	0	3	0	400	0	0
d0093	83	6	62	0	24	64	0	22	113	10	0	38	0	1397	7	19
d0094	0	3	2	0	10	0	0	0	50	2	0	0	0	462	6	4
d0095	3	0	0	0	3	0	0	0	11	4	0	40	0	208	0	1
d0096	5	6	2	0	0	3	2	0	247	18	5	1	0	974	277	20
d0097	0	0	0	0	2	0	0	0	19	14	0	2	0	167	1	1
d0098	1	6	17	0	14	4	0	0	161	2	2	28	0	2293	2	8
d0099	6	3	20	0	8	20	3	4	60	1	1	4	0	1280	3	15
d0100	1	0	2	0	1	0	0	0	22	4	0	0	0	301	3	0

APPENDIX I: PERFORMANCE MEASUREMENTS

Table I.1: IRS-I: Performance measurement results per query

Query	tp	fp	fn	tn	tpfp	fntn	tpfn	fptn	tpfpfntn	P	R	Fo	F	Sn	S	Nf	A
q01	23	49	2	26	72	28	25	75	100	0.32	0.92	0.65	0.47	0.08	0.35	0.68	0.49
q02	9	41	6	44	50	50	15	85	100	0.18	0.60	0.48	0.28	0.40	0.52	0.82	0.53
q03	23	49	2	26	72	28	25	75	100	0.32	0.92	0.65	0.47	0.08	0.35	0.68	0.49
q04	11	42	6	41	53	47	17	83	100	0.21	0.65	0.51	0.32	0.35	0.49	0.79	0.52
q05	14	46	4	36	60	40	18	82	100	0.23	0.78	0.56	0.36	0.22	0.44	0.77	0.50
q06	26	55	1	18	81	19	27	73	100	0.32	0.96	0.75	0.48	0.04	0.25	0.68	0.44
q07	6	31	8	55	37	63	14	86	100	0.16	0.43	0.36	0.23	0.57	0.64	0.84	0.61
q08	4	19	10	67	23	77	14	86	100	0.17	0.29	0.22	0.21	0.71	0.78	0.83	0.71
q09	11	30	5	54	41	59	16	84	100	0.27	0.69	0.36	0.39	0.31	0.64	0.73	0.65
q10	11	39	3	47	50	50	14	86	100	0.22	0.79	0.45	0.34	0.21	0.55	0.78	0.58
q11	12	38	3	47	50	50	15	85	100	0.24	0.80	0.45	0.37	0.20	0.55	0.76	0.59
q12	8	41	3	48	49	51	11	89	100	0.16	0.73	0.46	0.26	0.27	0.54	0.84	0.56
q13	6	35	5	54	41	59	11	89	100	0.15	0.55	0.39	0.24	0.45	0.61	0.85	0.60
q14	7	37	4	52	44	56	11	89	100	0.16	0.64	0.42	0.26	0.36	0.58	0.84	0.59
q15	6	35	5	54	41	59	11	89	100	0.15	0.55	0.39	0.24	0.45	0.61	0.85	0.60
q16	10	20	5	65	30	70	15	85	100	0.33	0.67	0.24	0.44	0.33	0.76	0.67	0.75
q17	8	31	4	57	39	61	12	88	100	0.21	0.67	0.35	0.32	0.33	0.65	0.79	0.65
q18	9	31	4	56	40	60	13	87	100	0.23	0.69	0.36	0.35	0.31	0.64	0.78	0.65
q19	7	29	4	60	36	64	11	89	100	0.19	0.64	0.33	0.29	0.36	0.67	0.81	0.67
q20	6	24	5	65	30	70	11	89	100	0.20	0.55	0.27	0.29	0.45	0.73	0.80	0.71
q21	6	28	5	61	34	66	11	89	100	0.18	0.55	0.31	0.27	0.45	0.69	0.82	0.67
q22	6	25	5	64	31	69	11	89	100	0.19	0.55	0.28	0.28	0.45	0.72	0.81	0.70
q23	2	3	5	90	5	95	7	93	100	0.40	0.29	0.03	0.34	0.71	0.97	0.60	0.92
q24	7	9	1	83	16	84	8	92	100	0.44	0.88	0.10	0.59	0.13	0.90	0.56	0.90
q25	4	7	3	86	11	89	7	93	100	0.36	0.57	0.08	0.44	0.43	0.92	0.64	0.90
q26	3	5	4	88	8	92	7	93	100	0.38	0.43	0.05	0.40	0.57	0.95	0.63	0.91
q27	2	2	5	91	4	96	7	93	100	0.50	0.29	0.02	0.37	0.71	0.98	0.50	0.93
q28	0	1	7	92	1	99	7	93	100	0.00	0.00	0.01	0.00	1.00	0.99	1.00	0.92
q29	0	4	16	80	4	96	16	84	100	0.00	0.00	0.05	0.00	1.00	0.95	1.00	0.80
q30	0	3	16	81	3	97	16	84	100	0.00	0.00	0.04	0.00	1.00	0.96	1.00	0.81
q31	0	3	16	81	3	97	16	84	100	0.00	0.00	0.04	0.00	1.00	0.96	1.00	0.81
q32	0	2	16	82	2	98	16	84	100	0.00	0.00	0.02	0.00	1.00	0.98	1.00	0.82
q33	11	42	6	41	53	47	17	83	100	0.21	0.65	0.51	0.32	0.35	0.49	0.79	0.52
q34	3	18	13	66	21	79	16	84	100	0.14	0.19	0.21	0.16	0.81	0.79	0.86	0.69
q35	10	43	6	41	53	47	16	84	100	0.19	0.63	0.51	0.29	0.38	0.49	0.81	0.51
q36	3	18	13	66	21	79	16	84	100	0.14	0.19	0.21	0.16	0.81	0.79	0.86	0.69
q37	12	33	9	46	45	55	21	79	100	0.27	0.57	0.42	0.37	0.43	0.58	0.73	0.58
q38	9	36	9	46	45	55	18	82	100	0.20	0.50	0.44	0.29	0.50	0.56	0.80	0.55
q39	7	14	6	73	21	79	13	87	100	0.33	0.54	0.16	0.41	0.46	0.84	0.67	0.80
q40	8	13	6	73	21	79	14	86	100	0.38	0.57	0.15	0.46	0.43	0.85	0.62	0.81
q41	7	10	7	76	17	83	14	86	100	0.41	0.50	0.12	0.45	0.50	0.88	0.59	0.83
q42	14	46	6	34	60	40	20	80	100	0.23	0.70	0.58	0.35	0.30	0.43	0.77	0.48

Query	tp	fp	fn	tn	tpfp	fntn	tpfn	fptn	tpfpfntn	P	R	Fo	F	Sn	S	Nf	A
q43	13	55	6	26	68	32	19	81	100	0.19	0.68	0.68	0.30	0.32	0.32	0.81	0.39
q44	14	57	5	24	71	29	19	81	100	0.20	0.74	0.70	0.31	0.26	0.30	0.80	0.38
q45	12	56	6	26	68	32	18	82	100	0.18	0.67	0.68	0.28	0.33	0.32	0.82	0.38
q46	11	49	6	34	60	40	17	83	100	0.18	0.65	0.59	0.28	0.35	0.41	0.82	0.45
q47	7	20	11	62	27	73	18	82	100	0.26	0.39	0.24	0.31	0.61	0.76	0.74	0.69
q48	0	6	19	75	6	94	19	81	100	0.00	0.00	0.07	0.00	1.00	0.93	1.00	0.75
q49	16	48	5	31	64	36	21	79	100	0.25	0.76	0.61	0.38	0.24	0.39	0.75	0.47
q50	22	44	3	31	66	34	25	75	100	0.33	0.88	0.59	0.48	0.12	0.41	0.67	0.53
q51	4	5	14	77	9	91	18	82	100	0.44	0.22	0.06	0.29	0.78	0.94	0.56	0.81
q52	8	35	10	47	43	57	18	82	100	0.19	0.44	0.43	0.27	0.56	0.57	0.81	0.55
q53	6	19	12	63	25	75	18	82	100	0.24	0.33	0.23	0.28	0.67	0.77	0.76	0.69
q54	10	43	8	39	53	47	18	82	100	0.19	0.56	0.52	0.28	0.44	0.48	0.81	0.49
q55	9	40	8	43	49	51	17	83	100	0.18	0.53	0.48	0.27	0.47	0.52	0.82	0.52
q56	6	26	11	57	32	68	17	83	100	0.19	0.35	0.31	0.25	0.65	0.69	0.81	0.63
q57	6	27	11	56	33	67	17	83	100	0.18	0.35	0.33	0.24	0.65	0.67	0.82	0.62
q58	11	42	7	40	53	47	18	82	100	0.21	0.61	0.51	0.31	0.39	0.49	0.79	0.51
q59	4	18	13	65	22	78	17	83	100	0.18	0.24	0.22	0.21	0.76	0.78	0.82	0.69
q60	7	12	5	76	19	81	12	88	100	0.37	0.58	0.14	0.45	0.42	0.86	0.63	0.83
q61	7	12	5	76	19	81	12	88	100	0.37	0.58	0.14	0.45	0.42	0.86	0.63	0.83
q62	6	17	3	74	23	77	9	91	100	0.26	0.67	0.19	0.37	0.33	0.81	0.74	0.80
q63	4	17	4	75	21	79	8	92	100	0.19	0.50	0.18	0.28	0.50	0.82	0.81	0.79
q64	6	10	5	79	16	84	11	89	100	0.38	0.55	0.11	0.45	0.45	0.89	0.63	0.85
q65	5	12	5	78	17	83	10	90	100	0.29	0.50	0.13	0.37	0.50	0.87	0.71	0.83
q66	28	72	0	0	100	0	28	72	100	0.28	1.00	1.00	0.44	0.00	0.00	0.72	0.28
q67	12	87	0	1	99	1	12	88	100	0.12	1.00	0.99	0.21	0.00	0.01	0.88	0.13
q68	11	88	0	1	99	1	11	89	100	0.11	1.00	0.99	0.20	0.00	0.01	0.89	0.12
q69	9	91	0	0	100	0	9	91	100	0.09	1.00	1.00	0.17	0.00	0.00	0.91	0.09
q70	15	78	2	5	93	7	17	83	100	0.16	0.88	0.94	0.27	0.12	0.06	0.84	0.20
q71	21	79	0	0	100	0	21	79	100	0.21	1.00	1.00	0.35	0.00	0.00	0.79	0.21
q72	15	84	0	1	99	1	15	85	100	0.15	1.00	0.99	0.26	0.00	0.01	0.85	0.16
q73	29	71	0	0	100	0	29	71	100	0.29	1.00	1.00	0.45	0.00	0.00	0.71	0.29
q74	23	76	0	1	99	1	23	77	100	0.23	1.00	0.99	0.37	0.00	0.01	0.77	0.24
q75	12	82	0	6	94	6	12	88	100	0.13	1.00	0.93	0.23	0.00	0.07	0.87	0.18

Table I.2: IRS-H: Performance measurement results per query

Query	tp	fp	fn	tn	tpfp	fntn	tpfn	fptn	tpfpfntn	P	R	Fo	F	Sn	S	Nf	A
q01	13	9	12	66	22	78	25	75	100	0.59	0.52	0.12	0.55	0.48	0.88	0.41	0.79
q02	3	4	12	81	7	93	15	85	100	0.43	0.20	0.05	0.27	0.80	0.95	0.57	0.84
q03	13	1	12	74	14	86	25	75	100	0.93	0.52	0.01	0.67	0.48	0.99	0.07	0.87
q04	0	0	17	83	0	100	17	83	100	0.00	0.00	0.00	0.00	1.00	1.00	0.00	0.83
q05	1	0	17	82	1	99	18	82	100	1.00	0.06	0.00	0.11	0.94	1.00	0.00	0.83
q06	16	5	11	68	21	79	27	73	100	0.76	0.59	0.07	0.66	0.41	0.93	0.24	0.84
q07	0	0	14	86	0	100	14	86	100	0.00	0.00	0.00	0.00	1.00	1.00	0.00	0.86
q08	0	0	14	86	0	100	14	86	100	0.00	0.00	0.00	0.00	1.00	1.00	0.00	0.86
q09	0	1	16	83	1	99	16	84	100	0.00	0.00	0.01	0.00	1.00	0.99	1.00	0.83
q10	2	3	12	83	5	95	14	86	100	0.40	0.14	0.03	0.21	0.86	0.97	0.60	0.85
q11	8	13	7	72	21	79	15	85	100	0.38	0.53	0.15	0.44	0.47	0.85	0.62	0.80
q12	0	1	11	88	1	99	11	89	100	0.00	0.00	0.01	0.00	1.00	0.99	1.00	0.88
q13	0	1	11	88	1	99	11	89	100	0.00	0.00	0.01	0.00	1.00	0.99	1.00	0.88
q14	0	3	11	86	3	97	11	89	100	0.00	0.00	0.03	0.00	1.00	0.97	1.00	0.86
q15	0	0	11	89	0	100	11	89	100	0.00	0.00	0.00	0.00	1.00	1.00	0.00	0.89
q16	0	0	15	85	0	100	15	85	100	0.00	0.00	0.00	0.00	1.00	1.00	0.00	0.85
q17	2	2	10	86	4	96	12	88	100	0.50	0.17	0.02	0.25	0.83	0.98	0.50	0.88
q18	3	2	10	85	5	95	13	87	100	0.60	0.23	0.02	0.33	0.77	0.98	0.40	0.88
q19	0	0	11	89	0	100	11	89	100	0.00	0.00	0.00	0.00	1.00	1.00	0.00	0.89
q20	0	0	11	89	0	100	11	89	100	0.00	0.00	0.00	0.00	1.00	1.00	0.00	0.89
q21	0	0	11	89	0	100	11	89	100	0.00	0.00	0.00	0.00	1.00	1.00	0.00	0.89
q22	0	0	11	89	0	100	11	89	100	0.00	0.00	0.00	0.00	1.00	1.00	0.00	0.89
q23	1	0	6	93	1	99	7	93	100	1.00	0.14	0.00	0.25	0.86	1.00	0.00	0.94
q24	2	2	6	90	4	96	8	92	100	0.50	0.25	0.02	0.33	0.75	0.98	0.50	0.92
q25	0	0	7	93	0	100	7	93	100	0.00	0.00	0.00	0.00	1.00	1.00	0.00	0.93
q26	0	0	7	93	0	100	7	93	100	0.00	0.00	0.00	0.00	1.00	1.00	0.00	0.93
q27	0	0	7	93	0	100	7	93	100	0.00	0.00	0.00	0.00	1.00	1.00	0.00	0.93
q28	0	0	7	93	0	100	7	93	100	0.00	0.00	0.00	0.00	1.00	1.00	0.00	0.93
q29	0	4	16	80	4	96	16	84	100	0.00	0.00	0.05	0.00	1.00	0.95	1.00	0.80
q30	0	0	16	84	0	100	16	84	100	0.00	0.00	0.00	0.00	1.00	1.00	0.00	0.84
q31	0	0	16	84	0	100	16	84	100	0.00	0.00	0.00	0.00	1.00	1.00	0.00	0.84
q32	0	0	16	84	0	100	16	84	100	0.00	0.00	0.00	0.00	1.00	1.00	0.00	0.84
q33	4	12	13	71	16	84	17	83	100	0.25	0.24	0.14	0.24	0.76	0.86	0.75	0.75
q34	2	1	14	83	3	97	16	84	100	0.67	0.13	0.01	0.22	0.88	0.99	0.33	0.85
q35	0	0	16	84	0	100	16	84	100	0.00	0.00	0.00	0.00	1.00	1.00	0.00	0.84
q36	0	0	16	84	0	100	16	84	100	0.00	0.00	0.00	0.00	1.00	1.00	0.00	0.84
q37	5	7	16	72	12	88	21	79	100	0.42	0.24	0.09	0.31	0.76	0.91	0.58	0.77
q38	3	2	15	80	5	95	18	82	100	0.60	0.17	0.02	0.26	0.83	0.98	0.40	0.83
q39	2	0	11	87	2	98	13	87	100	1.00	0.15	0.00	0.26	0.85	1.00	0.00	0.89
q40	0	1	14	85	1	99	14	86	100	0.00	0.00	0.01	0.00	1.00	0.99	1.00	0.85
q41	0	0	14	86	0	100	14	86	100	0.00	0.00	0.00	0.00	1.00	1.00	0.00	0.86
q42	4	2	16	78	6	94	20	80	100	0.67	0.20	0.03	0.31	0.80	0.98	0.33	0.82
q43	4	6	15	75	10	90	19	81	100	0.40	0.21	0.07	0.28	0.79	0.93	0.60	0.79
q44	3	4	16	77	7	93	19	81	100	0.43	0.16	0.05	0.23	0.84	0.95	0.57	0.80
q45	0	0	18	82	0	100	18	82	100	0.00	0.00	0.00	0.00	1.00	1.00	0.00	0.82
q46	0	1	17	82	1	99	17	83	100	0.00	0.00	0.01	0.00	1.00	0.99	1.00	0.82

Query	tp	fp	fn	tn	tpfp	fntn	tpfn	fptn	tpfpfntn	P	R	Fo	F	Sn	S	Nf	A
q47	0	0	18	82	0	100	18	82	100	0.00	0.00	0.00	0.00	1.00	1.00	0.00	0.82
q48	0	0	19	81	0	100	19	81	100	0.00	0.00	0.00	0.00	1.00	1.00	0.00	0.81
q49	7	8	14	71	15	85	21	79	100	0.47	0.33	0.10	0.39	0.67	0.90	0.53	0.78
q50	11	8	14	67	19	81	25	75	100	0.58	0.44	0.11	0.50	0.56	0.89	0.42	0.78
q51	0	0	18	82	0	100	18	82	100	0.00	0.00	0.00	0.00	1.00	1.00	0.00	0.82
q52	1	0	17	82	1	99	18	82	100	1.00	0.06	0.00	0.11	0.94	1.00	0.00	0.83
q53	1	0	17	82	1	99	18	82	100	1.00	0.06	0.00	0.11	0.94	1.00	0.00	0.83
q54	1	0	17	82	1	99	18	82	100	1.00	0.06	0.00	0.11	0.94	1.00	0.00	0.83
q55	0	0	17	83	0	100	17	83	100	0.00	0.00	0.00	0.00	1.00	1.00	0.00	0.83
q56	0	0	17	83	0	100	17	83	100	0.00	0.00	0.00	0.00	1.00	1.00	0.00	0.83
q57	0	0	17	83	0	100	17	83	100	0.00	0.00	0.00	0.00	1.00	1.00	0.00	0.83
q58	1	0	17	82	1	99	18	82	100	1.00	0.06	0.00	0.11	0.94	1.00	0.00	0.83
q59	0	0	17	83	0	100	17	83	100	0.00	0.00	0.00	0.00	1.00	1.00	0.00	0.83
q60	7	2	5	86	9	91	12	88	100	0.78	0.58	0.02	0.67	0.42	0.98	0.22	0.93
q61	5	3	7	85	8	92	12	88	100	0.63	0.42	0.03	0.50	0.58	0.97	0.38	0.90
q62	0	0	9	91	0	100	9	91	100	0.00	0.00	0.00	0.00	1.00	1.00	0.00	0.91
q63	0	0	8	92	0	100	8	92	100	0.00	0.00	0.00	0.00	1.00	1.00	0.00	0.92
q64	4	1	7	88	5	95	11	89	100	0.80	0.36	0.01	0.50	0.64	0.99	0.20	0.92
q65	1	3	9	87	4	96	10	90	100	0.25	0.10	0.03	0.14	0.90	0.97	0.75	0.88
q66	16	13	12	59	29	71	28	72	100	0.55	0.57	0.18	0.56	0.43	0.82	0.45	0.75
q67	7	17	5	71	24	76	12	88	100	0.29	0.58	0.19	0.39	0.42	0.81	0.71	0.78
q68	1	7	10	82	8	92	11	89	100	0.13	0.09	0.08	0.11	0.91	0.92	0.88	0.83
q69	3	1	6	90	4	96	9	91	100	0.75	0.33	0.01	0.46	0.67	0.99	0.25	0.93
q70	0	4	17	79	4	96	17	83	100	0.00	0.00	0.05	0.00	1.00	0.95	1.00	0.79
q71	7	21	14	58	28	72	21	79	100	0.25	0.33	0.27	0.28	0.67	0.73	0.75	0.65
q72	2	1	13	84	3	97	15	85	100	0.67	0.13	0.01	0.22	0.87	0.99	0.33	0.86
q73	14	9	15	62	23	77	29	71	100	0.61	0.48	0.13	0.54	0.52	0.87	0.39	0.76
q74	9	10	14	67	19	81	23	77	100	0.47	0.39	0.13	0.43	0.61	0.87	0.53	0.76
q75	6	8	6	80	14	86	12	88	100	0.43	0.50	0.09	0.46	0.50	0.91	0.57	0.86

Table I.3: IRS-H phrase-term frequencies

pt	Phrase-term	doc	ptf
pt01	design science	d0002	28
pt01	design science	d0005	3
pt01	design science	d0009	1
pt01	design science	d0010	1
pt01	design science	d0021	1
pt01	design science	d0028	17
pt01	design science	d0035	1
pt01	design science	d0058	2
pt01	design science	d0059	21
pt01	design science	d0060	1
pt01	design science	d0062	2
pt01	design science	d0063	29
pt01	design science	d0064	14
pt01	design science	d0065	22
pt01	design science	d0069	2
pt01	design science	d0074	1
pt01	design science	d0075	76
pt01	design science	d0076	22
pt01	design science	d0077	26
pt01	design science	d0078	71
pt01	design science	d0079	39
pt01	design science	d0096	1
pt02	design sciences	d0002	1
pt02	design sciences	d0035	1
pt02	design sciences	d0063	2
pt02	design sciences	d0064	1
pt02	design sciences	d0077	1
pt02	design sciences	d0078	4
pt02	design sciences	d0079	1
pt03	design science research	d0002	19
pt03	design science research	d0005	1
pt03	design science research	d0028	4
pt03	design science research	d0059	11
pt03	design science research	d0063	3
pt03	design science research	d0064	5
pt03	design science research	d0065	15
pt03	design science research	d0069	2
pt03	design science research	d0074	1
pt03	design science research	d0075	28
pt03	design science research	d0076	11
pt03	design science research	d0077	18
pt03	design science research	d0078	56
pt03	design science research	d0079	25
pt05	the design method	d0063	3
pt06	design research	d0002	17
pt06	design research	d0005	2

pt	Phrase-term	doc	ptf
pt06	design research	d0021	20
pt06	design research	d0043	3
pt06	design research	d0057	38
pt06	design research	d0058	28
pt06	design research	d0059	52
pt06	design research	d0061	2
pt06	design research	d0062	1
pt06	design research	d0063	20
pt06	design research	d0064	2
pt06	design research	d0065	62
pt06	design research	d0068	2
pt06	design research	d0073	2
pt06	design research	d0075	24
pt06	design research	d0076	11
pt06	design research	d0077	5
pt06	design research	d0078	16
pt06	design research	d0079	17
pt06	design research	d0088	1
pt06	design research	d0090	2
pt09	qualitative method	d0098	1
pt10	qualitative analysis	d0033	1
pt10	qualitative analysis	d0034	1
pt10	qualitative analysis	d0048	2
pt10	qualitative analysis	d0061	1
pt10	qualitative analysis	d0064	1
pt11	qualitative research	d0004	2
pt11	qualitative research	d0009	1
pt11	qualitative research	d0010	1
pt11	qualitative research	d0012	1
pt11	qualitative research	d0015	1
pt11	qualitative research	d0016	1
pt11	qualitative research	d0021	1
pt11	qualitative research	d0031	2
pt11	qualitative research	d0043	1
pt11	qualitative research	d0053	2
pt11	qualitative research	d0056	1
pt11	qualitative research	d0061	1
pt11	qualitative research	d0062	2
pt11	qualitative research	d0063	2
pt11	qualitative research	d0064	8
pt11	qualitative research	d0078	1
pt11	qualitative research	d0080	3
pt11	qualitative research	d0088	1
pt11	qualitative research	d0091	3
pt11	qualitative research	d0098	9
pt11	qualitative research	d0100	1
pt12	qualitative research design	d0080	2

pt	Phrase-term	doc	ptf
pt13	qualitative research method	d0064	1
pt14	qualitative research methods	d0062	1
pt14	qualitative research methods	d0064	1
pt14	qualitative research methods	d0098	2
pt17	quantitative analysis	d0064	1
pt17	quantitative analysis	d0075	1
pt17	quantitative analysis	d0085	1
pt17	quantitative analysis	d0093	1
pt18	quantitative research	d0064	2
pt18	quantitative research	d0084	1
pt18	quantitative research	d0088	2
pt18	quantitative research	d0098	3
pt18	quantitative research	d0099	1
pt23	clinical guideline	d0031	1
pt24	clinical guidelines	d0031	1
pt24	clinical guidelines	d0045	1
pt24	clinical guidelines	d0083	2
pt24	clinical guidelines	d0085	3
pt29	cloud computing	d0004	3
pt29	cloud computing	d0049	1
pt29	cloud computing	d0098	1
pt29	cloud computing	d0099	1
pt33	conceptual framework	d0006	1
pt33	conceptual framework	d0013	1
pt33	conceptual framework	d0019	1
pt33	conceptual framework	d0027	1
pt33	conceptual framework	d0036	4
pt33	conceptual framework	d0038	1
pt33	conceptual framework	d0044	1
pt33	conceptual framework	d0058	1
pt33	conceptual framework	d0061	1
pt33	conceptual framework	d0064	2
pt33	conceptual framework	d0066	1
pt33	conceptual framework	d0068	1
pt33	conceptual framework	d0072	1
pt33	conceptual framework	d0079	1
pt33	conceptual framework	d0091	1
pt33	conceptual framework	d0093	1
pt34	conceptual frameworks	d0063	1
pt34	conceptual frameworks	d0064	2
pt34	conceptual frameworks	d0091	2
pt37	conceptual model	d0003	1
pt37	conceptual model	d0025	1
pt37	conceptual model	d0035	2
pt37	conceptual model	d0038	1
pt37	conceptual model	d0050	1
pt37	conceptual model	d0051	3

pt	Phrase-term	doc	ptf
pt37	conceptual model	d0064	1
pt37	conceptual model	d0073	90
pt37	conceptual model	d0082	1
pt37	conceptual model	d0092	2
pt37	conceptual model	d0093	1
pt37	conceptual model	d0098	6
pt38	conceptual models	d0001	2
pt38	conceptual models	d0050	4
pt38	conceptual models	d0064	11
pt38	conceptual models	d0073	35
pt38	conceptual models	d0094	1
pt39	research ethics	d0040	3
pt39	research ethics	d0088	27
pt40	ethics in research	d0098	1
pt42	design method	d0053	1
pt42	design method	d0063	8
pt42	design method	d0075	4
pt42	design method	d0077	3
pt42	design method	d0078	1
pt42	design method	d0079	1
pt43	design methods	d0021	4
pt43	design methods	d0033	2
pt43	design methods	d0034	2
pt43	design methods	d0061	3
pt43	design methods	d0064	1
pt43	design methods	d0068	1
pt43	design methods	d0069	10
pt43	design methods	d0073	1
pt43	design methods	d0074	3
pt43	design methods	d0075	3
pt44	design practice	d0045	1
pt44	design practice	d0058	1
pt44	design practice	d0059	3
pt44	design practice	d0062	1
pt44	design practice	d0064	1
pt44	design practice	d0065	3
pt44	design practice	d0068	1
pt46	design research method	d0075	1
pt49	design theory	d0002	7
pt49	design theory	d0028	1
pt49	design theory	d0033	1
pt49	design theory	d0034	1
pt49	design theory	d0058	53
pt49	design theory	d0059	7
pt49	design theory	d0062	70
pt49	design theory	d0063	85
pt49	design theory	d0064	10

pt	Phrase-term	doc	ptf
pt49	design theory	d0065	7
pt49	design theory	d0075	8
pt49	design theory	d0076	4
pt49	design theory	d0077	8
pt49	design theory	d0078	9
pt49	design theory	d0079	7
pt50	data quality	d0001	68
pt50	data quality	d0008	2
pt50	data quality	d0018	43
pt50	data quality	d0019	45
pt50	data quality	d0020	5
pt50	data quality	d0048	2
pt50	data quality	d0050	3
pt50	data quality	d0051	134
pt50	data quality	d0052	17
pt50	data quality	d0064	2
pt50	data quality	d0070	34
pt50	data quality	d0071	5
pt50	data quality	d0072	236
pt50	data quality	d0081	3
pt50	data quality	d0083	1
pt50	data quality	d0084	7
pt50	data quality	d0085	2
pt50	data quality	d0098	12
pt50	data quality	d0099	2
pt52	data quality methodology	d0072	5
pt53	data quality methodologies	d0072	3
pt54	data quality model	d0001	3
pt58	data quality framework	d0072	3
pt60	electronic health record	d0007	2
pt60	electronic health record	d0031	2
pt60	electronic health record	d0045	3
pt60	electronic health record	d0047	1
pt60	electronic health record	d0048	3
pt60	electronic health record	d0082	2
pt60	electronic health record	d0083	11
pt60	electronic health record	d0084	13
pt60	electronic health record	d0085	3
pt61	electronic health records	d0015	1
pt61	electronic health records	d0031	5
pt61	electronic health records	d0045	6
pt61	electronic health records	d0049	2
pt61	electronic health records	d0082	4
pt61	electronic health records	d0083	4
pt61	electronic health records	d0084	2
pt61	electronic health records	d0085	3
pt64	electronic patient record	d0044	3

pt	Phrase-term	doc	ptf
pt64	electronic patient record	d0045	6
pt64	electronic patient record	d0048	2
pt64	electronic patient record	d0050	1
pt64	electronic patient record	d0085	3
pt65	electronic patient records	d0032	1
pt65	electronic patient records	d0044	1
pt65	electronic patient records	d0045	5
pt65	electronic patient records	d0085	5

Table I.4: IRS-H phrase-term document and collection frequencies

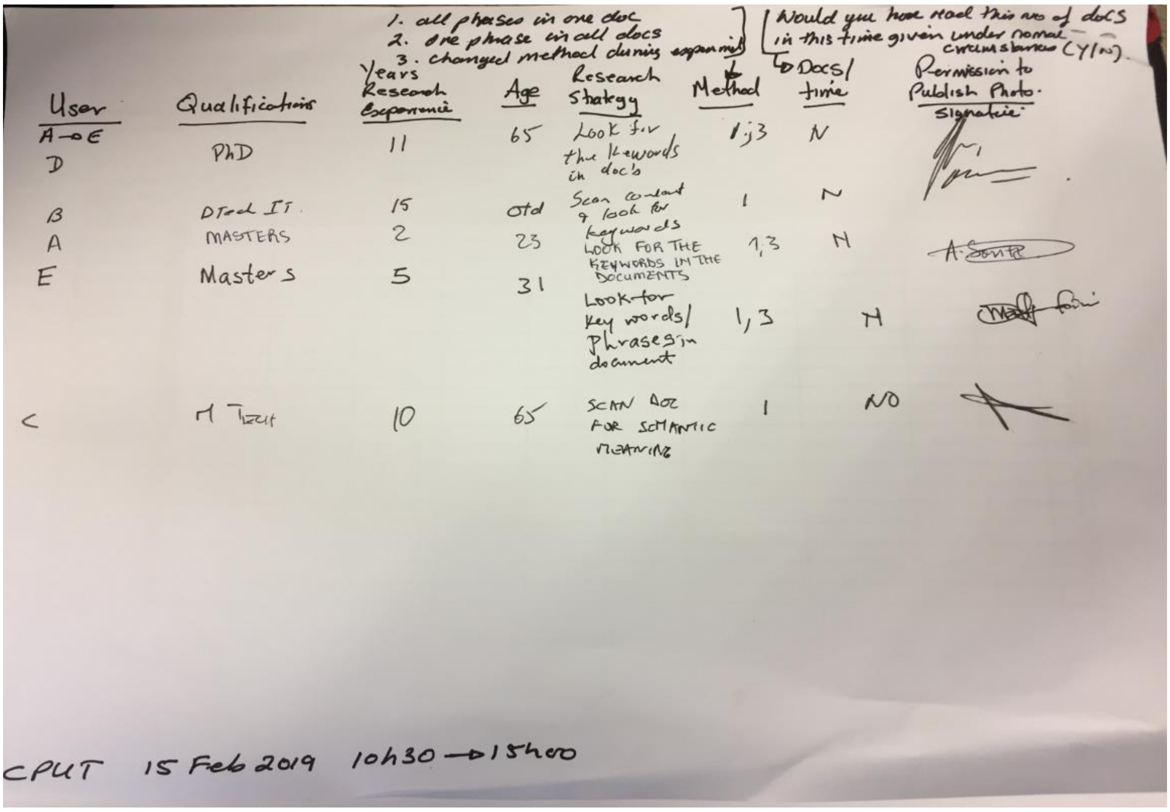
pt	Phrase-term	df	cf
pt01	design science	22	381
pt02	design sciences	7	11
pt03	design science research	14	199
pt05	the design method	1	3
pt06	design research	21	327
pt09	qualitative method	1	1
pt10	qualitative analysis	5	6
pt11	qualitative research	21	45
pt12	qualitative research design	1	2
pt13	qualitative research method	1	1
pt14	qualitative research methods	3	4
pt17	quantitative analysis	4	4
pt18	quantitative research	5	9
pt23	clinical guideline	1	1
pt24	clinical guidelines	4	7
pt29	cloud computing	4	6
pt33	conceptual framework	16	20
pt34	conceptual frameworks	3	5
pt37	conceptual model	12	110
pt38	conceptual models	5	53
pt39	research ethics	2	30
pt40	ethics in research	1	1
pt42	design method	6	18
pt43	design methods	10	30
pt44	design practice	7	11
pt46	design research method	1	1
pt49	design theory	15	278
pt50	data quality	19	623
pt52	data quality methodology	1	5
pt53	data quality methodologies	1	3
pt54	data quality model	1	3
pt58	data quality framework	1	3
pt60	electronic health record	9	40
pt61	electronic health records	8	27
pt64	electronic patient record	5	15
pt65	electronic patient records	4	12

Table I.5: IRS-I term based document and collection frequencies

t	Word	df	cf
t01	analysis	88	1438
t02	care	44	975
t03	clinical	23	363
t04	cloud	7	27
t05	computing	36	255
t06	conceptual	56	520
t07	data	86	4907
t08	design	82	3966
t09	e	91	1701
t10	electronic	54	392
t11	ethics	21	149
t12	family	30	167
t13	for	99	10029
t14	framework	78	1090
t15	frameworks	33	242
t16	guideline	19	33
t17	guidelines	46	220
t18	health	50	2342
t19	in	99	20499
t20	management	78	1778
t21	method	65	718
t22	methodologies	35	312
t23	methodology	62	668
t24	methods	77	843
t25	model	79	1266
t26	models	71	632
t27	operations	34	130
t28	paradigm	43	271
t29	paradigms	25	111
t30	patient	24	514
t31	philosophy	28	160
t32	practice	77	1114
t33	pragmatism	7	97
t34	primary	49	371
t35	principles	53	336
t36	qualitative	50	287
t37	qualities	13	38
t38	quality	73	2409
t39	quantitative	41	227
t40	record	29	299
t41	records	26	244
t42	research	97	5279
t43	science	84	993
t44	sciences	55	229
t45	service	59	831
t46	stroke	2	4
t47	the	100	50204
t48	theory	71	2186
t49	types	62	410

APPENDIX J: EXPERIMENT AND DEMOGRAPHIC DATA

Figure J.1: Photograph of five users participating in experiment on 15 February 2019



1. all phrases in one doc
2. one phrase in all docs
3. changed method during experiment

Would you have read this no of docs in this time given under normal circumstances (Y/N).

User A → E	Qualifications	Years Research Experience	Age	Research Strategy	Method	Docs/time	Permission to Publish Photo. Signature
D	PhD	11	65	Look for the keywords in doc's	1;3	N	
B	DTech IT.	15	old	Scan content & look for keywords	1	N	
A	MASTERS	2	23	LOOK FOR THE KEYWORDS IN THE DOCUMENTS	1,3	N	A. Sonte
E	Masters	5	31	Look for key words/ phrases in document	1,3	N	
C	M Tech	10	65	SCAN DOC FOR SEMANTIC MEANING	1	NO	

CPUT 15 Feb 2019 10h30 → 15h00

Figure J.2: Demographic data of the users with authorised signatures

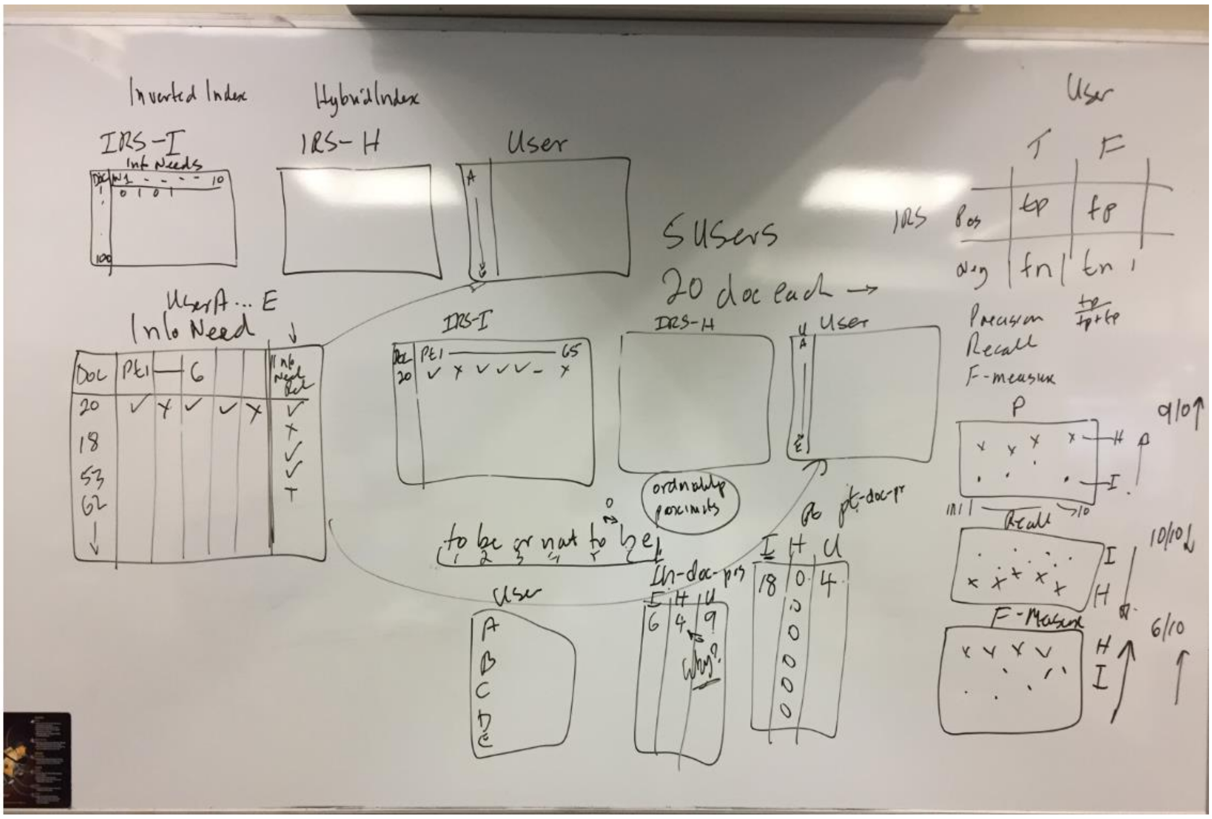
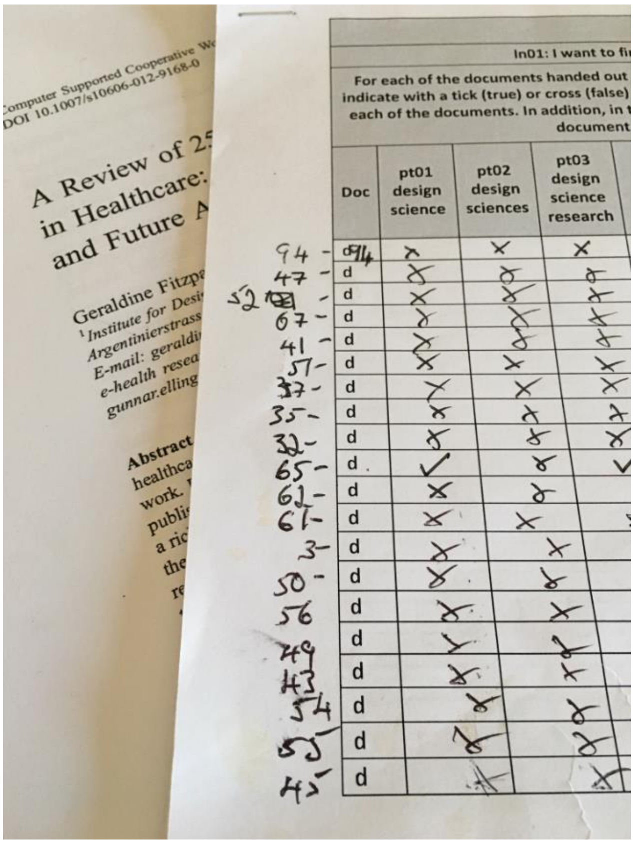


Figure J.3: Conceptualisation of performance measurements



Computer Supported Cooperative Wo
DOI 10.1007/s10606-012-9168-0

A Review of 25
in Healthcare:
and Future A

Geraldine Fitzpa
¹ Institute for Desi
Argentinierstrass
E-mail: geraldi
e-health resea
gunnar.elling

Abstract
healthca
work.
publi
a ric
the
re

In01: I want to fi

For each of the documents handed out
indicate with a tick (true) or cross (false)
each of the documents. In addition, in t
document

	Doc	pt01 design science	pt02 design sciences	pt03 design science research
94 -	d94	✗	✗	✗
47 -	d	✗	✗	✗
52 -	d	✗	✗	✗
67 -	d	✗	✗	✗
41 -	d	✗	✗	✗
57 -	d	✗	✗	✗
37 -	d	✗	✗	✗
35 -	d	✗	✗	✗
32 -	d	✗	✗	✗
65 -	d	✓	✗	✓
62 -	d	✗	✗	
61 -	d	✗	✗	
3 -	d	✗	✗	
50 -	d	✗	✗	
56	d	✗	✗	
49	d	✗	✗	
43	d	✗	✗	
54	d	✗	✗	
55	d	✗	✗	
45	d	✗	✗	

Figure J.4: Sample of completed questionnaire

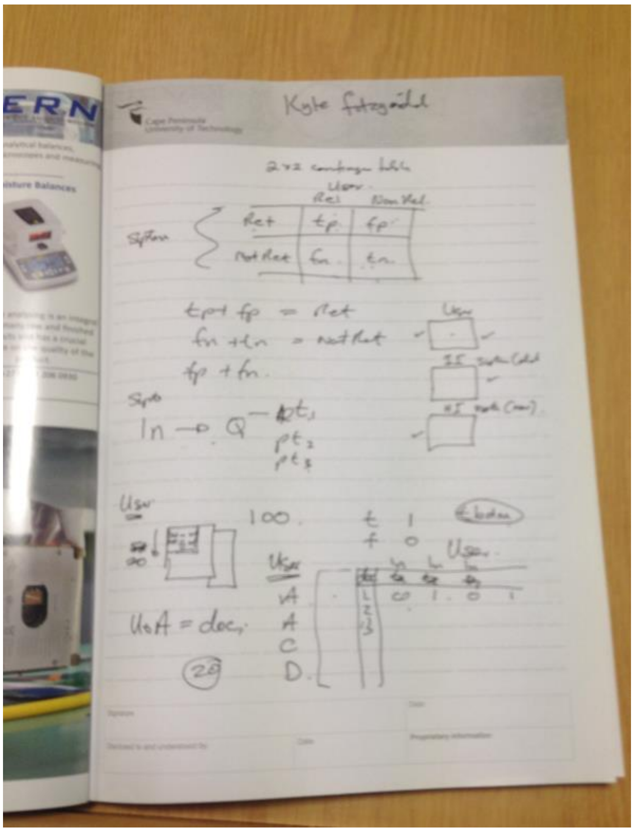


Figure J.5: Conceptualisation of UoA and UoO

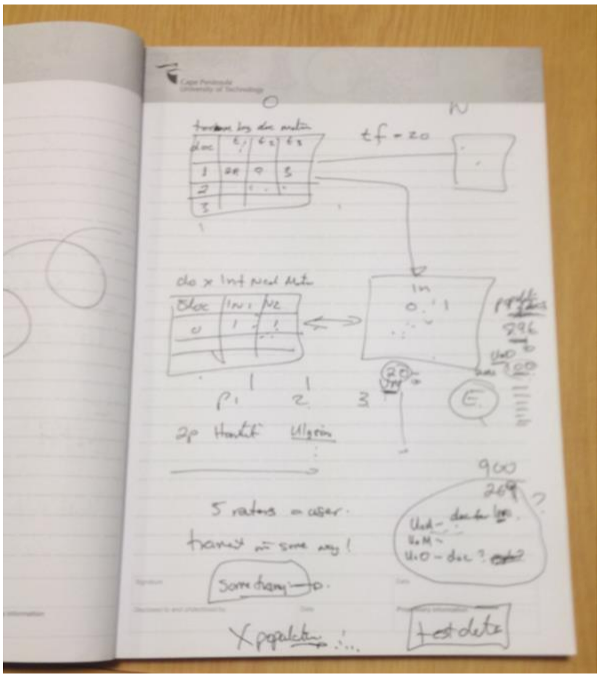
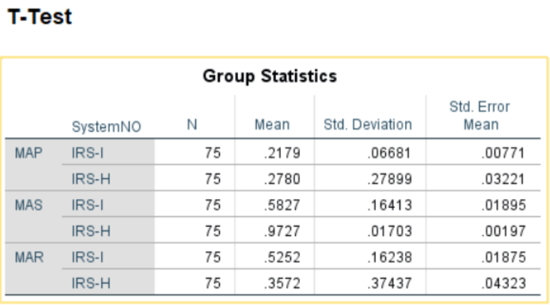


Figure J.6: Conceptualisation of questionnaire

APPENDIX K: SPSS RESULTS

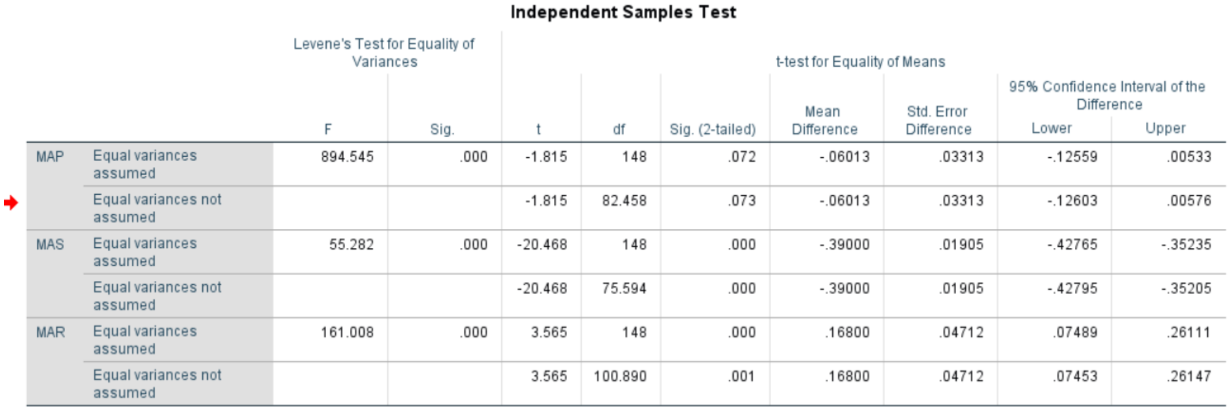
The statistical analysis results produced by SPSS are presented below:



T-Test

Group Statistics

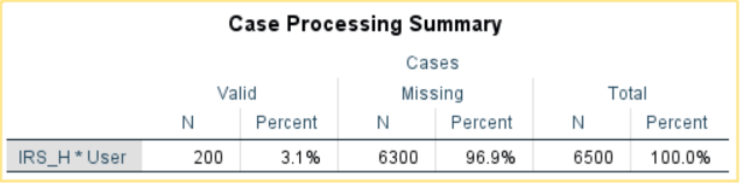
	SystemNO	N	Mean	Std. Deviation	Std. Error Mean
MAP	IRS-I	75	.2179	.06681	.00771
	IRS-H	75	.2780	.27899	.03221
MAS	IRS-I	75	.5827	.16413	.01895
	IRS-H	75	.9727	.01703	.00197
MAR	IRS-I	75	.5252	.16238	.01875
	IRS-H	75	.3572	.37437	.04323



Independent Samples Test

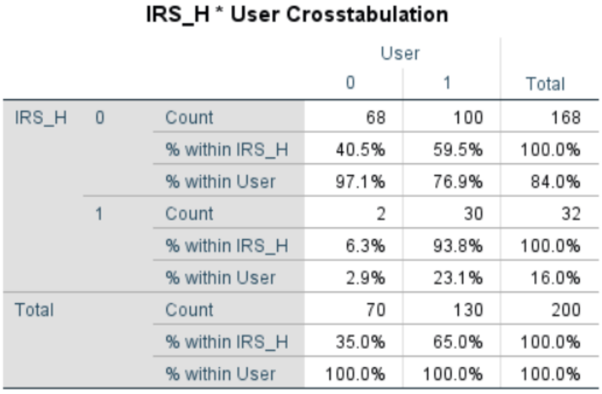
		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
MAP	Equal variances assumed	894.545	.000	-1.815	148	.072	-.06013	.03313	-.12559	.00533
	Equal variances not assumed			-1.815	82.458	.073	-.06013	.03313	-.12603	.00576
MAS	Equal variances assumed	55.282	.000	-20.468	148	.000	-.39000	.01905	-.42765	-.35235
	Equal variances not assumed			-20.468	75.594	.000	-.39000	.01905	-.42795	-.35205
MAR	Equal variances assumed	161.008	.000	3.565	148	.000	.16800	.04712	.07489	.26111
	Equal variances not assumed			3.565	100.890	.001	.16800	.04712	.07453	.26147

Kappa – Information need: User-A



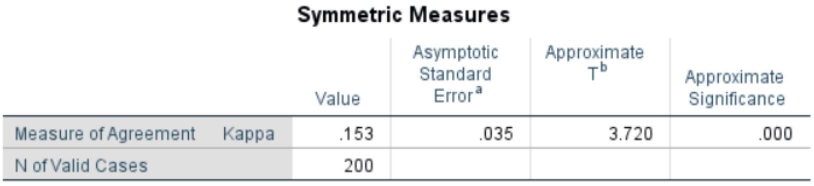
Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
IRS_H * User	200	3.1%	6300	96.9%	6500	100.0%



IRS_H * User Crosstabulation

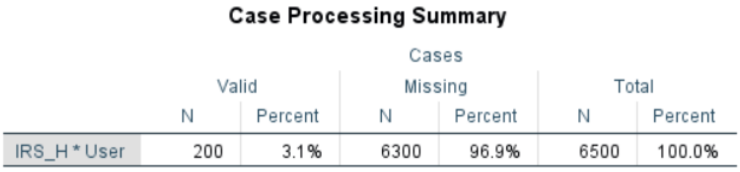
			User		
			0	1	Total
IRS_H	0	Count	68	100	168
		% within IRS_H	40.5%	59.5%	100.0%
		% within User	97.1%	76.9%	84.0%
	1	Count	2	30	32
		% within IRS_H	6.3%	93.8%	100.0%
		% within User	2.9%	23.1%	16.0%
Total		Count	70	130	200
		% within IRS_H	35.0%	65.0%	100.0%
		% within User	100.0%	100.0%	100.0%



Symmetric Measures

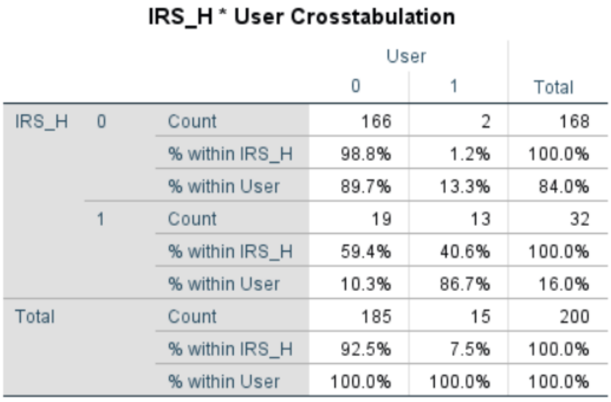
		Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Measure of Agreement	Kappa	.153	.035	3.720	.000
N of Valid Cases		200			

Kappa – Information need: User-B



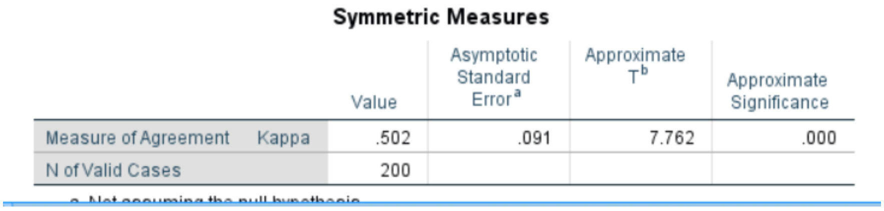
Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
IRS_H * User	200	3.1%	6300	96.9%	6500	100.0%



IRS_H * User Crosstabulation

			User		
			0	1	Total
IRS_H	0	Count	166	2	168
		% within IRS_H	98.8%	1.2%	100.0%
		% within User	89.7%	13.3%	84.0%
	1	Count	19	13	32
		% within IRS_H	59.4%	40.6%	100.0%
		% within User	10.3%	86.7%	16.0%
Total		Count	185	15	200
		% within IRS_H	92.5%	7.5%	100.0%
		% within User	100.0%	100.0%	100.0%

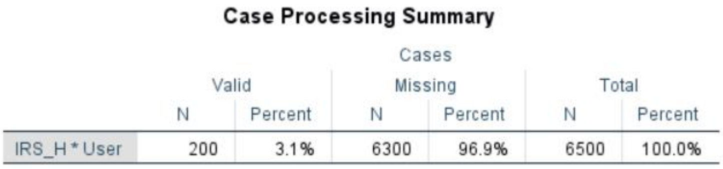


Symmetric Measures

		Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Measure of Agreement	Kappa	.502	.091	7.762	.000
N of Valid Cases		200			

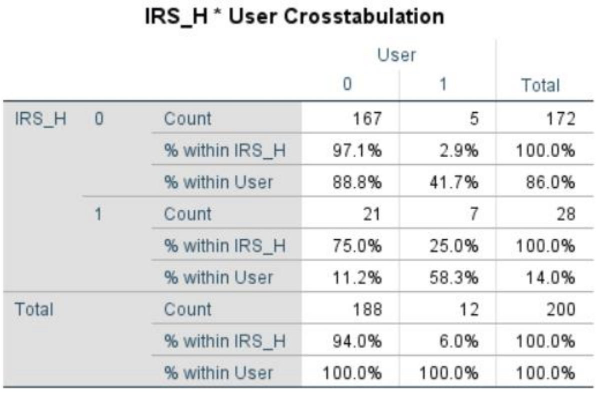
a. Not assuming the null hypothesis.

Kappa – Information need: User-C



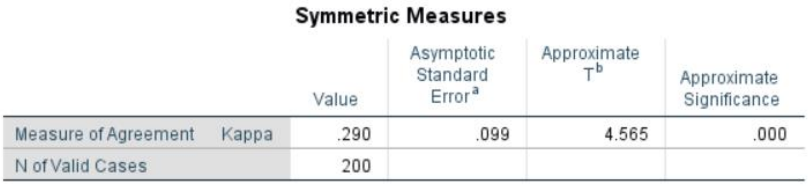
Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
IRS_H * User	200	3.1%	6300	96.9%	6500	100.0%



IRS_H * User Crosstabulation

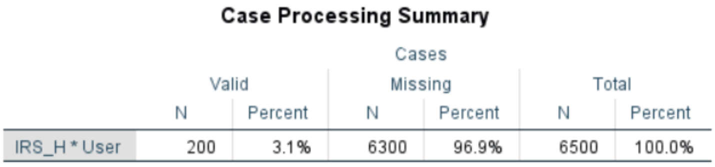
			User		
			0	1	Total
IRS_H	0	Count	167	5	172
		% within IRS_H	97.1%	2.9%	100.0%
		% within User	88.8%	41.7%	86.0%
	1	Count	21	7	28
		% within IRS_H	75.0%	25.0%	100.0%
		% within User	11.2%	58.3%	14.0%
Total		Count	188	12	200
		% within IRS_H	94.0%	6.0%	100.0%
		% within User	100.0%	100.0%	100.0%



Symmetric Measures

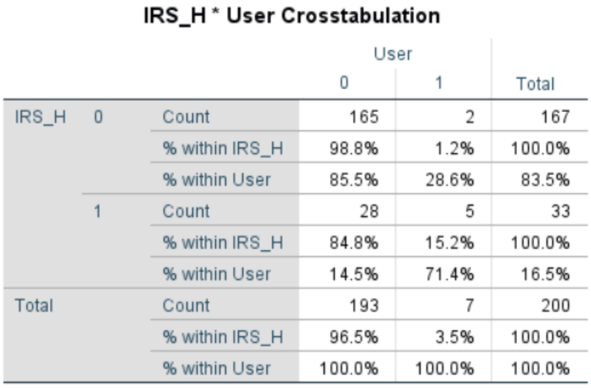
		Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Measure of Agreement	Kappa	.290	.099	4.565	.000
N of Valid Cases		200			

Kappa – Information need: User-D



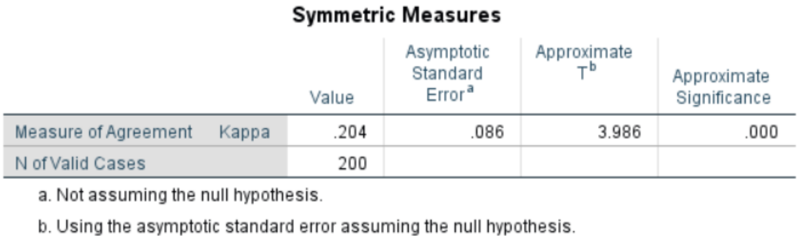
Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
IRS_H * User	200	3.1%	6300	96.9%	6500	100.0%



IRS_H * User Crosstabulation

			User		
			0	1	Total
IRS_H	0	Count	165	2	167
		% within IRS_H	98.8%	1.2%	100.0%
		% within User	85.5%	28.6%	83.5%
	1	Count	28	5	33
		% within IRS_H	84.8%	15.2%	100.0%
		% within User	14.5%	71.4%	16.5%
Total		Count	193	7	200
		% within IRS_H	96.5%	3.5%	100.0%
		% within User	100.0%	100.0%	100.0%



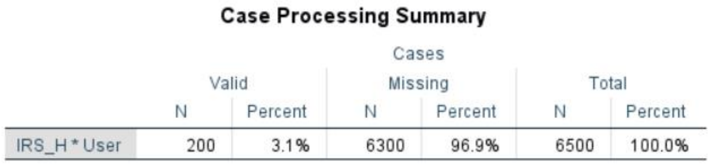
Symmetric Measures

		Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Measure of Agreement	Kappa	.204	.086	3.986	.000
N of Valid Cases		200			

a. Not assuming the null hypothesis.

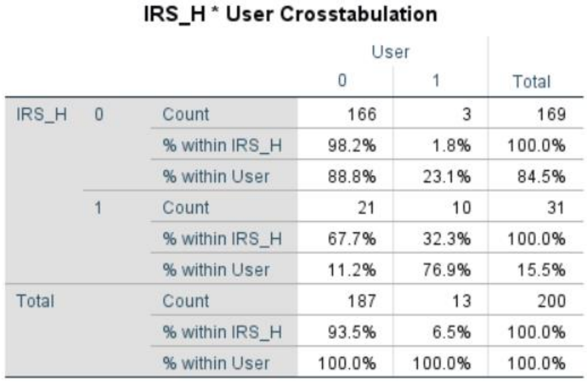
b. Using the asymptotic standard error assuming the null hypothesis.

Kappa – Information need: User-E



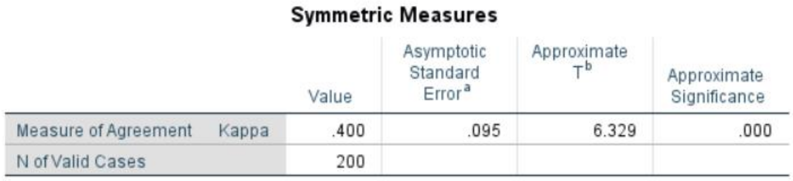
Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
IRS_H * User	200	3.1%	6300	96.9%	6500	100.0%



IRS_H * User Crosstabulation

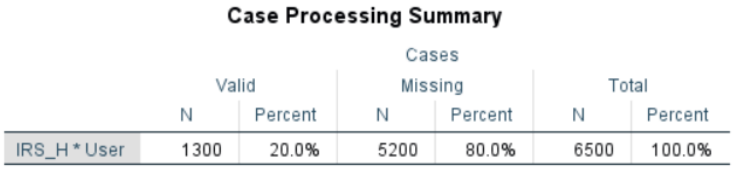
			User		
			0	1	Total
IRS_H	0	Count	166	3	169
		% within IRS_H	98.2%	1.8%	100.0%
		% within User	88.8%	23.1%	84.5%
	1	Count	21	10	31
		% within IRS_H	67.7%	32.3%	100.0%
		% within User	11.2%	76.9%	15.5%
Total		Count	187	13	200
		% within IRS_H	93.5%	6.5%	100.0%
		% within User	100.0%	100.0%	100.0%



Symmetric Measures

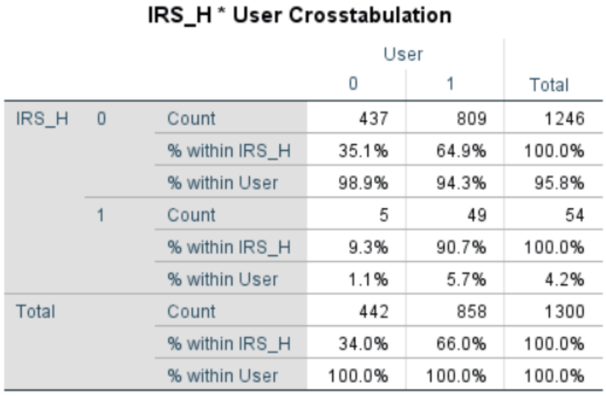
		Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Measure of Agreement	Kappa	.400	.095	6.329	.000
N of Valid Cases		200			

Kappa – 65 queries: User-A



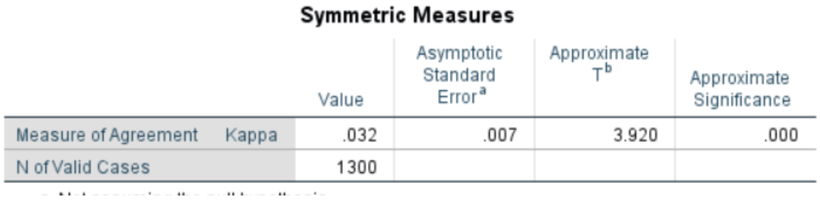
Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
IRS_H * User	1300	20.0%	5200	80.0%	6500	100.0%



IRS_H * User Crosstabulation

			User		
			0	1	Total
IRS_H	0	Count	437	809	1246
		% within IRS_H	35.1%	64.9%	100.0%
		% within User	98.9%	94.3%	95.8%
	1	Count	5	49	54
		% within IRS_H	9.3%	90.7%	100.0%
		% within User	1.1%	5.7%	4.2%
Total		Count	442	858	1300
		% within IRS_H	34.0%	66.0%	100.0%
		% within User	100.0%	100.0%	100.0%

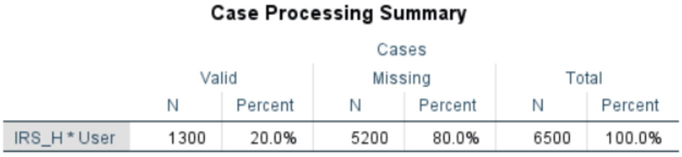


Symmetric Measures

		Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Measure of Agreement	Kappa	.032	.007	3.920	.000
N of Valid Cases		1300			

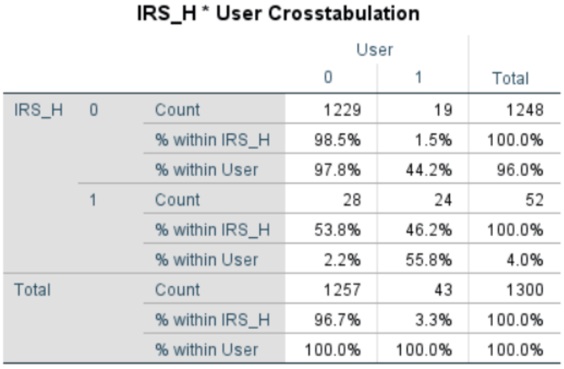
a. Not assuming the null hypothesis.

Kappa – 65 queries: User-B



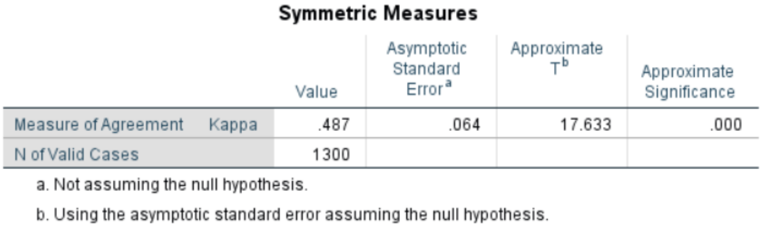
Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
IRS_H * User	1300	20.0%	5200	80.0%	6500	100.0%



IRS_H * User Crosstabulation

			User		
			0	1	Total
IRS_H	0	Count	1229	19	1248
		% within IRS_H	98.5%	1.5%	100.0%
		% within User	97.8%	44.2%	96.0%
	1	Count	28	24	52
		% within IRS_H	53.8%	46.2%	100.0%
		% within User	2.2%	55.8%	4.0%
Total		Count	1257	43	1300
		% within IRS_H	96.7%	3.3%	100.0%
		% within User	100.0%	100.0%	100.0%



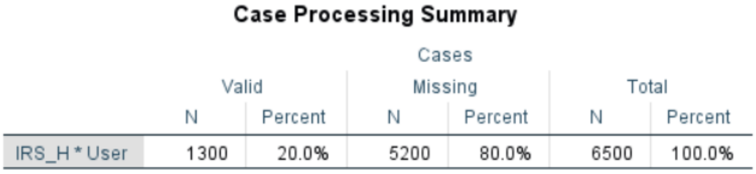
Symmetric Measures

		Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Measure of Agreement	Kappa	.487	.064	17.633	.000
N of Valid Cases		1300			

a. Not assuming the null hypothesis.

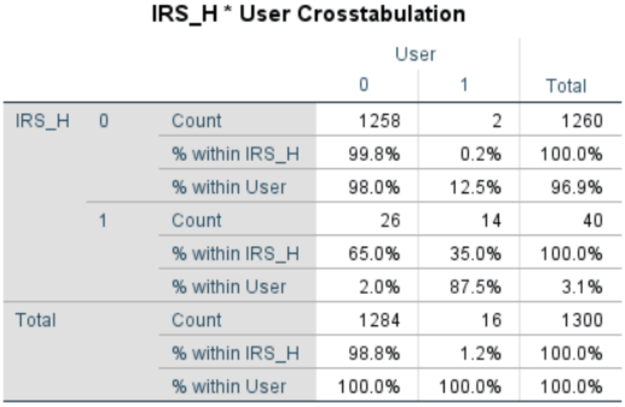
b. Using the asymptotic standard error assuming the null hypothesis.

Kappa – 65 queries: User-C



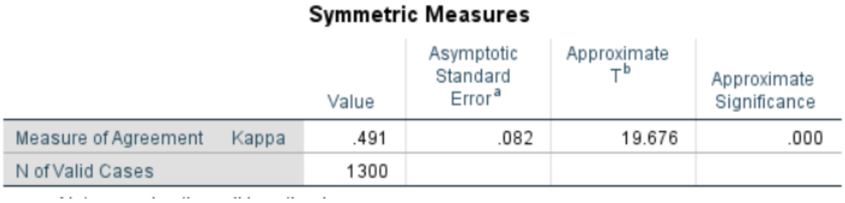
Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
IRS_H * User	1300	20.0%	5200	80.0%	6500	100.0%



IRS_H * User Crosstabulation

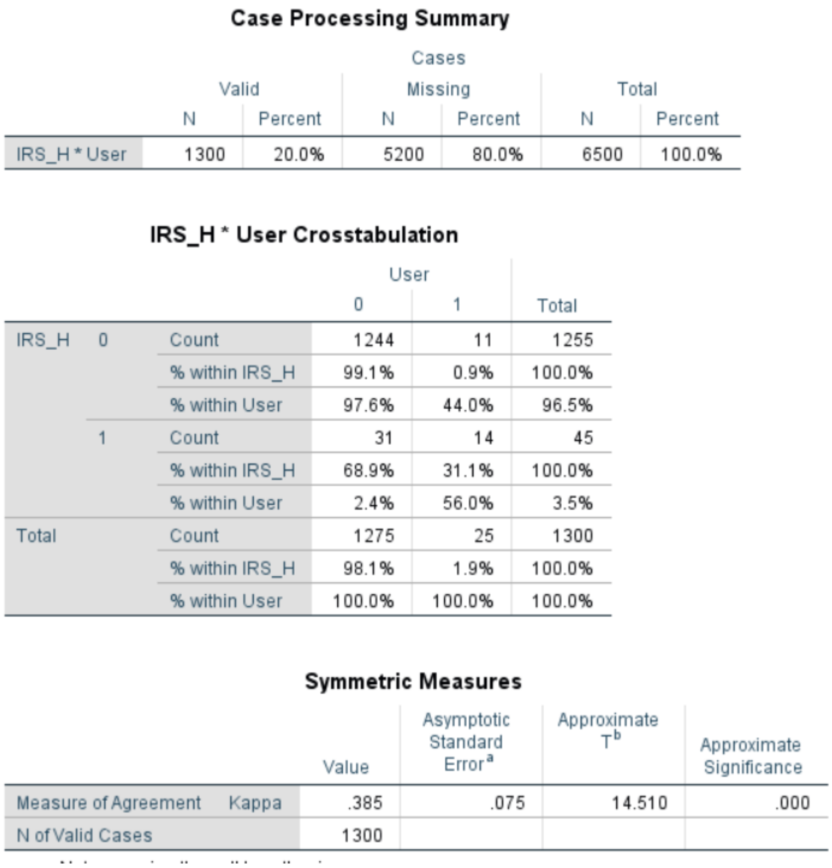
			User		
			0	1	Total
IRS_H	0	Count	1258	2	1260
		% within IRS_H	99.8%	0.2%	100.0%
		% within User	98.0%	12.5%	96.9%
	1	Count	26	14	40
		% within IRS_H	65.0%	35.0%	100.0%
		% within User	2.0%	87.5%	3.1%
Total		Count	1284	16	1300
		% within IRS_H	98.8%	1.2%	100.0%
		% within User	100.0%	100.0%	100.0%



Symmetric Measures

		Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Measure of Agreement	Kappa	.491	.082	19.676	.000
N of Valid Cases		1300			

Kappa – 65 queries: User-D



Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
IRS_H * User	1300	20.0%	5200	80.0%	6500	100.0%

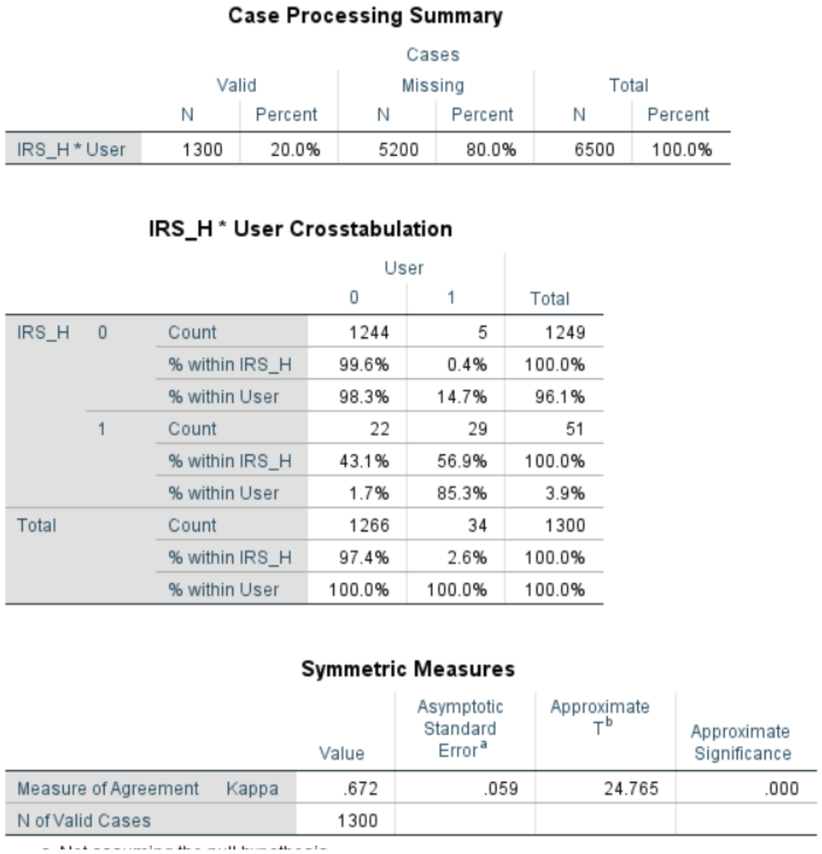
IRS_H * User Crosstabulation

			User		
			0	1	Total
IRS_H	0	Count	1244	11	1255
		% within IRS_H	99.1%	0.9%	100.0%
		% within User	97.6%	44.0%	96.5%
	1	Count	31	14	45
		% within IRS_H	68.9%	31.1%	100.0%
		% within User	2.4%	56.0%	3.5%
Total		Count	1275	25	1300
		% within IRS_H	98.1%	1.9%	100.0%
		% within User	100.0%	100.0%	100.0%

Symmetric Measures

		Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Measure of Agreement	Kappa	.385	.075	14.510	.000
N of Valid Cases		1300			

Kappa – 65 queries: User-E



Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
IRS_H * User	1300	20.0%	5200	80.0%	6500	100.0%

IRS_H * User Crosstabulation

			User		
			0	1	Total
IRS_H	0	Count	1244	5	1249
		% within IRS_H	99.6%	0.4%	100.0%
		% within User	98.3%	14.7%	96.1%
	1	Count	22	29	51
		% within IRS_H	43.1%	56.9%	100.0%
		% within User	1.7%	85.3%	3.9%
Total		Count	1266	34	1300
		% within IRS_H	97.4%	2.6%	100.0%
		% within User	100.0%	100.0%	100.0%

Symmetric Measures

		Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Measure of Agreement	Kappa	.672	.059	24.765	.000
N of Valid Cases		1300			

Table K.1: Table of critical values (Dougherty, 2019)

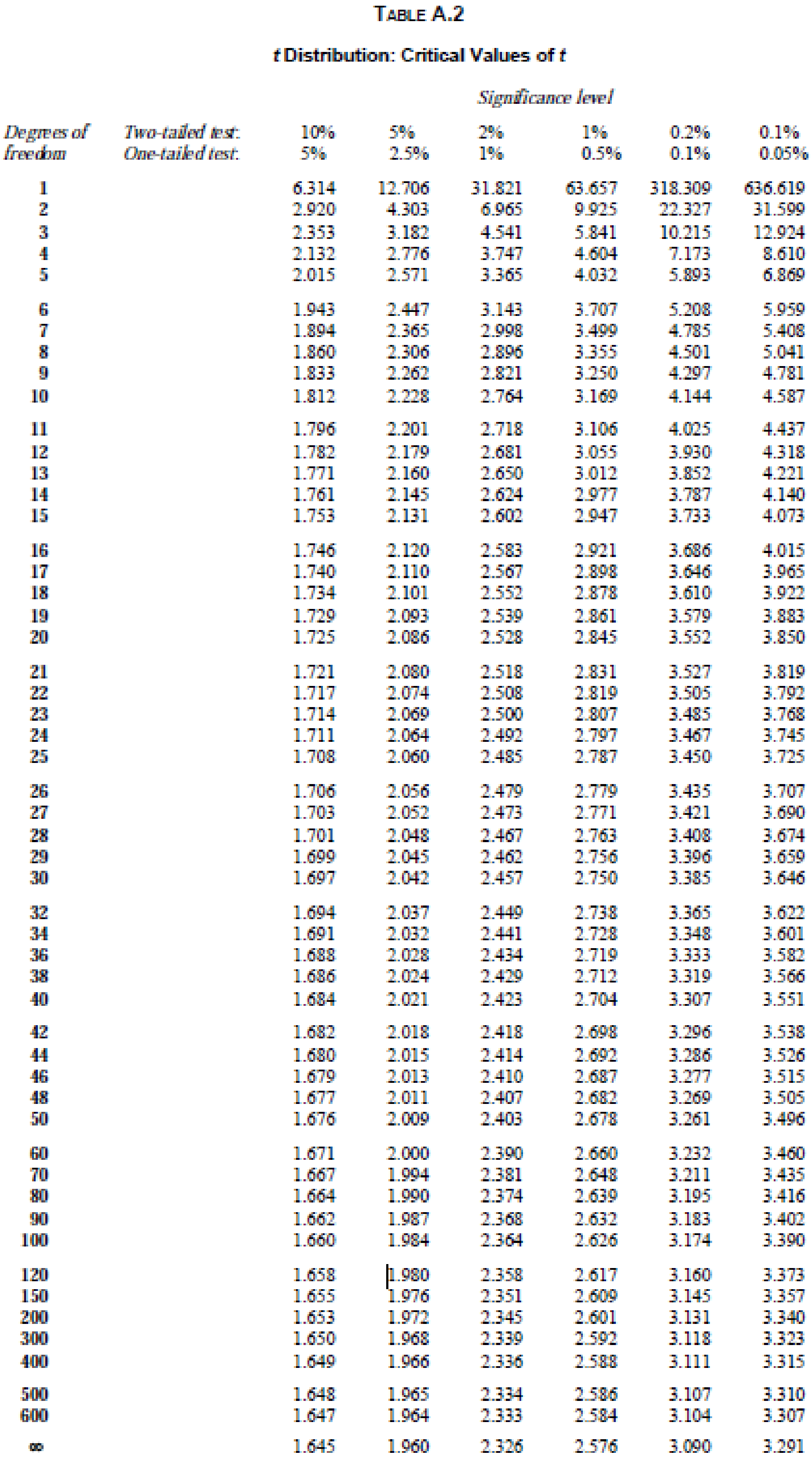


TABLE A.2

t Distribution: Critical Values of *t*

		Significance level					
Degrees of	*Two-tailed test:*	10%	5%	2%	1%	0.2%	0.1%
freedom	*One-tailed test:*	5%	2.5%	1%	0.5%	0.1%	0.05%
1		6.314	12.706	31.821	63.657	318.309	636.619
2		2.920	4.303	6.965	9.925	22.327	31.599
3		2.353	3.182	4.541	5.841	10.215	12.924
4		2.132	2.776	3.747	4.604	7.173	8.610
5		2.015	2.571	3.365	4.032	5.893	6.869
6		1.943	2.447	3.143	3.707	5.208	5.959
7		1.894	2.365	2.998	3.499	4.785	5.408
8		1.860	2.306	2.896	3.355	4.501	5.041
9		1.833	2.262	2.821	3.250	4.297	4.781
10		1.812	2.228	2.764	3.169	4.144	4.587
11		1.796	2.201	2.718	3.106	4.025	4.437
12		1.782	2.179	2.681	3.055	3.930	4.318
13		1.771	2.160	2.650	3.012	3.852	4.221
14		1.761	2.145	2.624	2.977	3.787	4.140
15		1.753	2.131	2.602	2.947	3.733	4.073
16		1.746	2.120	2.583	2.921	3.686	4.015
17		1.740	2.110	2.567	2.898	3.646	3.965
18		1.734	2.101	2.552	2.878	3.610	3.922
19		1.729	2.093	2.539	2.861	3.579	3.883
20		1.725	2.086	2.528	2.845	3.552	3.850
21		1.721	2.080	2.518	2.831	3.527	3.819
22		1.717	2.074	2.508	2.819	3.505	3.792
23		1.714	2.069	2.500	2.807	3.485	3.768
24		1.711	2.064	2.492	2.797	3.467	3.745
25		1.708	2.060	2.485	2.787	3.450	3.725
26		1.706	2.056	2.479	2.779	3.435	3.707
27		1.703	2.052	2.473	2.771	3.421	3.690
28		1.701	2.048	2.467	2.763	3.408	3.674
29		1.699	2.045	2.462	2.756	3.396	3.659
30		1.697	2.042	2.457	2.750	3.385	3.646
32		1.694	2.037	2.449	2.738	3.365	3.622
34		1.691	2.032	2.441	2.728	3.348	3.601
36		1.688	2.028	2.434	2.719	3.333	3.582
38		1.686	2.024	2.429	2.712	3.319	3.566
40		1.684	2.021	2.423	2.704	3.307	3.551
42		1.682	2.018	2.418	2.698	3.296	3.538
44		1.680	2.015	2.414	2.692	3.286	3.526
46		1.679	2.013	2.410	2.687	3.277	3.515
48		1.677	2.011	2.407	2.682	3.269	3.505
50		1.676	2.009	2.403	2.678	3.261	3.496
60		1.671	2.000	2.390	2.660	3.232	3.460
70		1.667	1.994	2.381	2.648	3.211	3.435
80		1.664	1.990	2.374	2.639	3.195	3.416
90		1.662	1.987	2.368	2.632	3.183	3.402
100		1.660	1.984	2.364	2.626	3.174	3.390
120		1.658	1.980	2.358	2.617	3.160	3.373
150		1.655	1.976	2.351	2.609	3.145	3.357
200		1.653	1.972	2.345	2.601	3.131	3.340
300		1.650	1.968	2.339	2.592	3.118	3.323
400		1.649	1.966	2.336	2.588	3.111	3.315
500		1.648	1.965	2.334	2.586	3.107	3.310
600		1.647	1.964	2.333	2.584	3.104	3.307
∞		1.645	1.960	2.326	2.576	3.090	3.291