

## RHP Internet Research Class

5-2 Keywords; Then Essence of Search      Name: \_\_\_\_\_



---

Your assignment is to read this document (yes, the WHOLE thing!) and answer the questions that can be found throughout (look for the 📄). Email me the answers at [rhp@denovodental.com](mailto:rhp@denovodental.com), and use the following subject **Compapp 5-2 lastname** (Last name is YOUR last name.)

---

Despite all the gobbledygook about things like ‘Boolean’ and query operators, the most difficult – and fundamental – aspect of a search are the keywords used in your query.

A search is inherently looking for information about a **topic**. This part describes how you can proceed from search concepts to identifying the specific keywords – or terms – that will give you the results you’re seeking. We begin by presenting an information problem which will be the basis for progressing through the tutorial’s remaining topics.

---

Jan is an office worker in downtown Minneapolis. While on lunch break one fine Spring day, Jan’s eye is caught by a flash in the sky above. Jan sees a bird about the size of a crow diving at high speed and catching in mid-air what appears to be a pigeon. The bird then swoops out of sight. Jan is captivated by the mostly gray and white bird, with the crooked black and yellow beak. Jan has never seen this bird before, and wonders what it is doing in the city. That night, Jan decides to find out more about this mystery bird on the Internet.

Where does Jan begin?

### ***Query Concepts: What, Where, When, How, Why***

Mastering the concepts behind a search is not as complicated as may seem at first. The first few searches are perhaps difficult, but, once done, the nuggets behind your information request start becoming clear. Like riding a bike for the first time, it does take some practice.

One of the bigger mistakes you can make in preparing a query is not providing enough keywords. On average, most users submit 1.5 keywords per query. This number is insufficient to accurately find the information you are seeking. Thus, a central task in query formulation is for you to identify a sufficient number of appropriate keywords.

If you are new to searching, the first task we recommend when formulating a search is writing down what information you are seeking. This is best done – go ahead, use some paper and a pen – in the form of some questions. Before doing a search, it is important to bound your topic as completely yet succinctly as possible. After experience is gained, you can skip writing things down and plunge right into it.

Formulating a query is akin to solving a mystery. Some pieces of information are available, but if sufficient information were available the answer would be known

## RHP Internet Research Class

5-2 Keywords; Then Essence of Search Name: \_\_\_\_\_



and there would be no need to seek more. This is the essence of a query: missing information. It is up to you, the searcher, to define your snare – the query sufficiently to trap that missing information and solve the mystery.

1. What is the average number of keywords most searchers use per query?

As any good detective would, it is useful to begin by listing what you do know according to these standard categories. Jan lists these for the mystery bird:

- **WHO / WHAT?** – gray and white bird, about the size of a crow; yellow and black beak
- **WHERE?** – downtown office buildings in the City of Minneapolis
- **WHEN?** – daylight in the Spring
- **HOW?** – fast flyer, hunting pigeons (?) as prey
- **WHY?** – hunting bird; why never seen before? blown off course? is it migrating?

### TIP

Always keep in mind the **who, what, where, when, how** and **why** in formulating your query.

Of course, not all of these five categories will apply to a given query, and the specifics will obviously vary for your desired topic. But it is useful to keep these five categories in mind – the what, where, when, how and why – when analyzing the major components.

### ***Breaking Down Your Query***

Let's take the five responses to the query tests (above) apart (yours will differ substantially, but the same ideas apply). First, there are many common words in these responses that are prepositions, conjunctions or common verbs. These include: **and, about, the, of, a, in, as, if, not, why, never, before, is** and **it**. These common words are referred to as "stoplist" words: they are essential to the connecting tissue in language, but they are filler in any search request. **All** search engines ignore them because they have minimal information value and are found commonly in all language. Search services include on the order of 600 of these common words in their "stoplists"; if you use them in a query they are ignored. Therefore, you should ignore them as well.

# RHP Internet Research Class

5-2 Keywords; Then Essence of Search

Name: \_\_\_\_\_



Okay, removing such words from our responses leaves these remaining words:

<b>gray</b>	<b>downtown</b>	<b>flyer</b>
<b>white</b>	<b>office</b>	<b>hunting</b>
<b>bird</b>	<b>buildings</b>	<b>pigeons</b>
<b>size</b>	<b>city</b>	<b>blown</b>
<b>crow</b>	<b>Minneapolis</b>	<b>off</b>
<b>yellow</b>	<b>daylight</b>	<b>course</b>
<b>black</b>	<b>Spring</b>	<b>migrating</b>
<b>beak</b>	<b>fast</b>	

## TIP

Never use articles, pronouns, conjunctions or prepositions – the connecting tissue in language – in your queries.

Now, let's further classify these terms into three categories, similar to diagramming a sentence (but made simpler for our purposes). Let's use the classifications of objects/nouns, actions/verbs and modifiers/qualifiers (adjectives, adverbs and predicate subjects). And, let's now re-list these words by these categories:

<u>Objects</u>	<u>Actions</u>	<u>Modifiers</u>
bird	blown	gray
buildings	migrating	white
city	not seen	size
Spring		crow
daylight		yellow
		black
		beak
		downtown
		office
		Minneapolis
		fast
		flyer
		hunting
		pigeons
		off
		course

Not all of these categories are equally useful in a query.

2. What are the five "W" words to keep in mind when formulating a query?
3. What four parts of speech are used for connecting words together and therefore should not be included in a query?

## ***Focus on Nouns and Objects***

## RHP Internet Research Class

5-2 Keywords; Then Essence of Search      Name: \_\_\_\_\_



Almost without exception, the central keywords in your queries will be nouns. Though sometimes adverbs and adjectives can help refine your search, the key pivot point is a noun, or series of nouns.

The most precise terms we have in language are for tangible, concrete “things” or objects. Actions and modifiers are very diverse, easily substitutable, and generally not universally applied in any given description. For, example, take the concept of “fast”. A thesaurus will give 75 or more different words for fast. Here are some counts from AltaVista for numbers of Web documents containing these terms:

fast 2,524,008  
speed 2,210,325  
quick 1,833,511  
rapid 787,344  
fleet 311,925  
swift 180,903  
breakneck 7,743

Or, alternatively, take a modifying concept like ‘color’. Again, here are the AltaVista document counts:

color 5,073,422  
red 3,683,578  
yellow 1,593,705  
blue 2,946,413  
gray 707,505  
grey 469,630  
slate 111,170  
white 3,925,525

### TIP

The keywords in your queries will most often be nouns – and then likely no more than 6 or 8 of them.

Note two aspects about these lists. First, a concept like speed or color can be described in lots of ways (most of which are not shown). Second, you generally don’t know how others would describe the same thing. In our example of Jan’s mystery hunting bird, would someone else describe it as “fast”, “quick” or “like a bolt from the sky”? Would someone else describe the bird as “gray”, “grey”, “slate-gray” or “smoky”?

The same kind of ambiguity and substitutability applies to actions or verbs. Does the bird “fly”, “soar”, “swoop” or “glide”, or any of the other dozens of ways the act of flying can be described?

As a general rule, try to avoid using action terms and mostly try to avoid using modifiers in your queries. Where exceptions to these guidelines may make



sense is when a modifier helps to precisely define your object, such as in “Limburger cheese.”

We’ve thus gone through a process that has led us to these possible objects as the focal points for constructing our query terms:

- bird
- buildings
- city
- Spring
- daylight

The obvious main subject is **bird**. The next few topics will concentrate on it; we’ll return to the other objects as we later refine our final query.

4. What parts of speech should you use in a query?

### ***Word Stemming and Use of Wildcards***

One of the first mistakes in query formulation is not using word stemming – or truncation – sufficiently. Let’s look at this question in regards to our subject, **bird**. Accounting for singular and plural cases of an object is easy to overlook; but, if done, can act to unduly restrict the universe of documents in which you will be conducting your search. Using AltaVista again, here are the document counts for



484,529

the single and plural versions of **bird**:



605,011

By using either only **bird** or **birds** as our subject, we would eliminate half or so of the potential documents that we’d like to use as our search basis. We could use both **bird** and **birds** as query terms, but that takes up valuable keyword slots. The better way to handle this problem is through truncation.

Truncation is applying a wildcard character after the first few letters in a term (the “stem”). The asterisk (\*) is the almost universally accepted truncation wildcard. Generally, you must also have a minimum of three characters at the beginning of the word as your stem basis. Once marked for truncation, then any matching

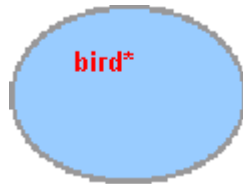
## RHP Internet Research Class

5-2 Keywords; Then Essence of Search Name:



characters after that will be picked up in the search query. Some search engines do stemming and truncation for you if you pick the right option on the search form. Some engines don't support stemming or truncation at all. In any case, using the asterisk wildcard will generally be ignored or you'll get a query format error if the search engine doesn't support it.

Remember, ANY words with characters after the stem will be matched to your query term if the search engine supports truncation. Thus, if we stem **bird\***, our search will match on the words **bird**, **birds**, **birding** and **birdbrain**. Posing **bird\*** to AltaVista we now get these document counts:



**1,076,900**

Note the document count is a bit lower than the total for the individual words **bird**, **birds**, **birding** and **birdbrain**. There are minor errors in how search engines retrieve word stems. But they are of a smaller magnitude than ignoring singular and plural cases altogether in the query, and seem to be a minor price to pay for being able to eliminate another keyword (**birds**, in addition to **bird**) from the search.

As you first begin to use truncation you need to be aware of unintended consequences. In the case of the stem **bird\*** there are relatively few unwanted words (**birdbrain**) picked up in the search. But let's look at another of the objects, **city**, in our mystery bird sample problem.

To stem and pick up the plural form of **city**, **cities**, we would need to specify **cit\***. But look at some of the words this stem specification would match:

citadel	cities	citric
citadels	citify	citriculture
citation	citizen	citrine
citations	citizenry	citrone
cite	citizens	citronella
cites	citizenship	citrus
cited	citrate	city

The **cit\*** stem clearly picks up way too many unwanted words.

### TIP

Truncation, or word stemming, keeps your keyword count down and makes for simpler queries.



5. What wild card character will match all characters?

Stemming tends to work best when the actual stem is longer, when plurals are represented by an added '-s' (as opposed to '-ies' or other forms), and the stem itself is not a root to many other common words. With just a little thought, however, truncation is easy and can pay useful dividends in properly scoping your query with a minimum of keywords. We highly recommend its use.

### ***Finding the Right Level***

Perhaps you've already noticed, but our query subject **bird\*** is contained on more than 1 million documents (in AltaVista alone). It would be a little difficult to review all of those documents at one sitting.

**THE MOST CRITICAL PROBLEM IN ALL QUERIES IS FINDING THE RIGHT LEVEL OF SPECIFICITY FOR THE SUBJECT QUERY TERM(S).** Too broad a keyword specification, and too many results are returned; too narrow a specification, and too few are returned.

All information is classifiable and amenable to structure. We are all familiar with dictionaries, which classify words alphabetically. However, an alphabetical structure is not of much use to query formulation. But there are many other classification schemes used for information which CAN help find the right level, or specificity, for your keywords. A few examples appropriate to our mystery bird search are presented in this topic.

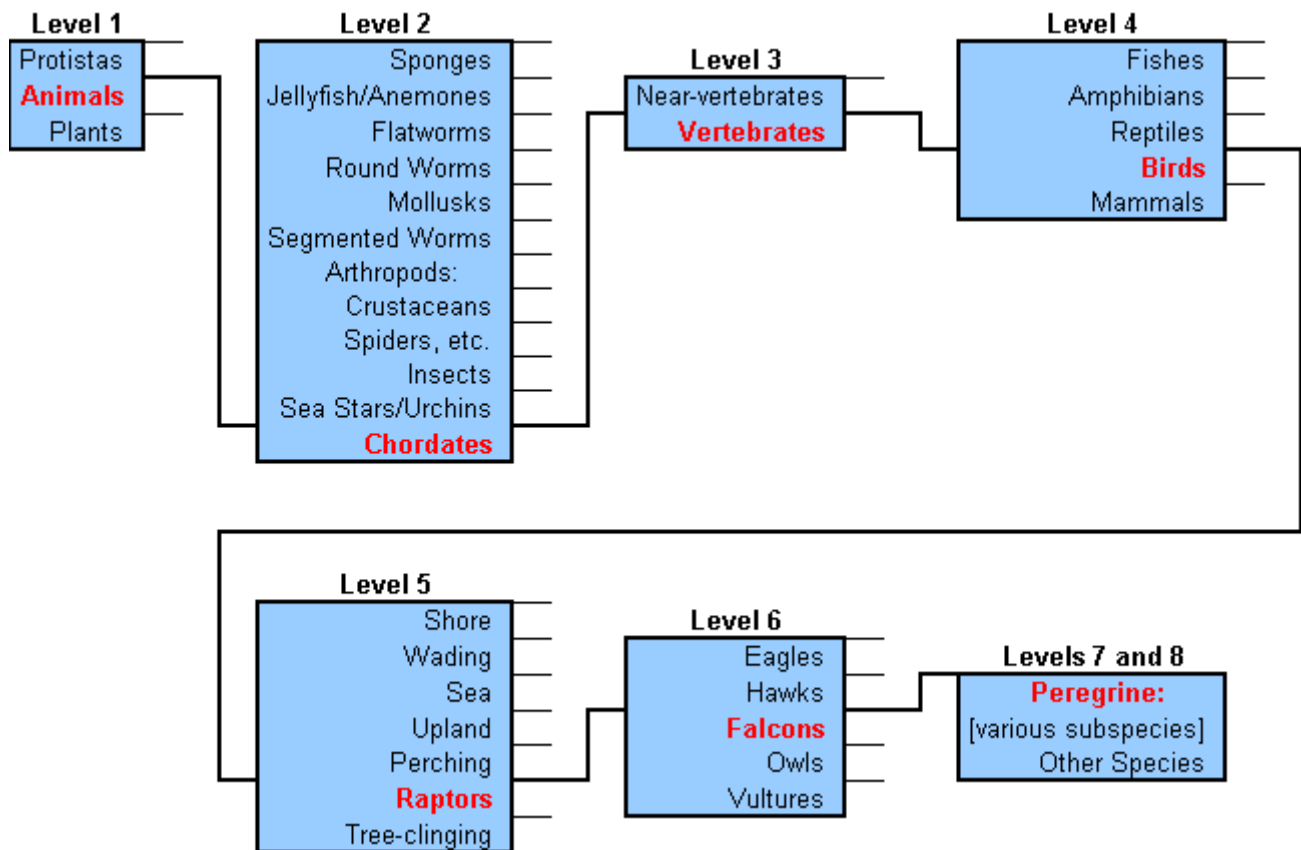
# RHP Internet Research Class

5-2 Keywords; Then Essence of Search

Name:



Our first example classification presents the structure of the animal kingdom:



'Level' Example Using the Kingdom of Life

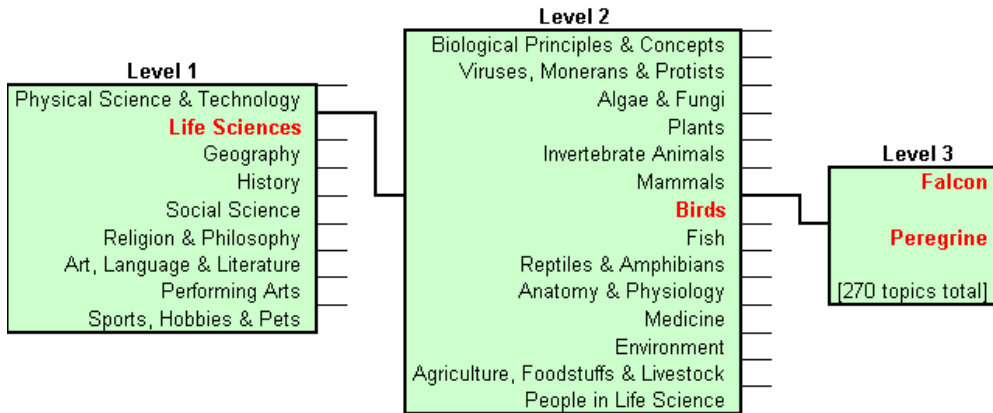
As we will see, our initial keyword term of **bird\*** is at least three levels off of where it should be. Using **bird\*** as is would lead to massive results sets from the search engines and virtually no likelihood that we will find the information we're looking for.

# RHP Internet Research Class

5-2 Keywords; Then Essence of Search Name:

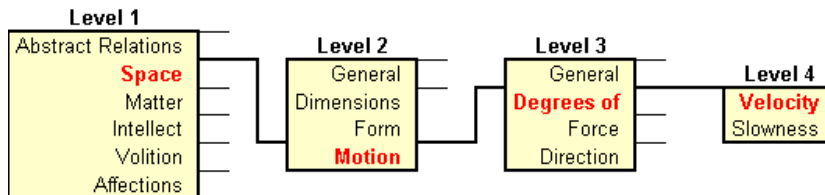


Another way to classify information is shown by the encyclopedia, with this example being drawn from Microsoft's Encarta 96 (the actual encyclopedia doesn't matter; we're only illustrating a point).



'Level' Example Using Encarta

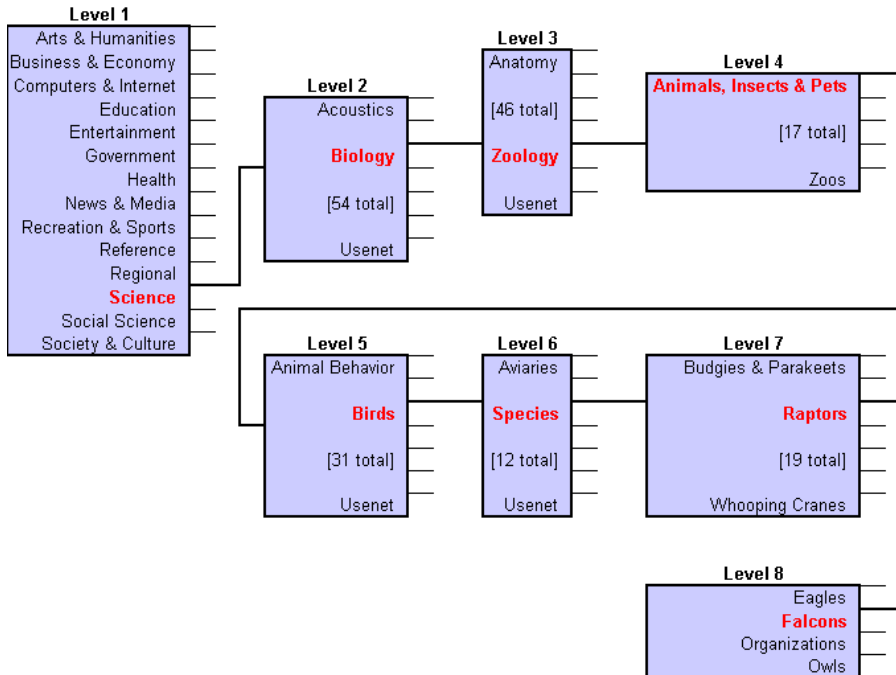
As a very different example, the chart below shows how the word "fast" is placed within the structure of a thesaurus:



'Level' Example Using Thesaurus



As noted, search 'directories' also apply a classification structure for how they organize and present Web sites. The structure for the largest and best known of these directories, Yahoo, with some 1400-odd individual categories, is shown below :



**'Level' Example Using Yahoo**

Like the first animal phylum example above, **bird\*** is about three or four levels off from where our subject keyword should be.

Finding the right level may involve your personal knowledge and experience, doing a preliminary search or consulting other references. In the case of Jan and the mystery bird, looking in a bird book was sufficient to match pictures with the bird seen as a **peregrine falcon**.

The time spent in finding how to characterize your subject at the proper level is definitely well spent, as these document counts from AltaVista illustrate:

**bird\*** 1,076,900  
**falcon\*** 235,635  
**peregrine falcon\*** 9,157

By identifying our mystery bird as a **peregrine falcon**, we've narrowed the search by 99%! Remember, at 30 seconds to 2.5 minutes per document reviewed, the effort spent in zeroing in on the bird of interest has saved us tremendous overall search time.



The critical point about finding the right “level” in your keywords is that words at levels higher than where you should be return way too many results; those at levels lower than where you should be return too few or no results. This phenomenon is due to the fact that “things” at lower levels tend to “rollup” and sum into “things” at higher levels.

Philosophers, epistemologists, taxonomists, linguists and others can argue for centuries about “proper” ways to classify information. That is not our concern. Rather, the point is that keyword objects can be placed into a structure at various levels. Always keeping forefront whether your query subject is at the right level or not in those structures can bring big benefits in faster, and more accurate searches.

6. What happens when too broad a keyword specification is used in a search?

### **Synonyms**

Let’s assume, however, that Jan was not able to match the bird book pictures with the mystery bird to identify it as a peregrine falcon.

One useful place to begin is with synonyms. Jan knows the mystery bird is a hunting bird. Jan lists other synonyms that come to mind for **hunting bird**. We provide AltaVista document counts for these synonyms:

**hunting bird\*** 1,448  
**bird\* of prey** 18,650

Jan, however, suspects neither of these terms is the “correct” synonym. Attacking this problem from another angle, Jan writes down specific kinds of birds of prey:

**hawk**  
**eagle**  
**owl**

Using these three keywords, Jan’s search immediately turns up a number of sites referring to raptors, the technical term for hunting birds. Jan finds a great site on raptors that also has pictures that positively identifies the mystery bird as a peregrine falcon. Jan also learns that vultures are raptors, too.

The best synonyms provide relatively complete coverage for the subject at hand and are “pitched” for the right informational objective. In Jan’s case, it was needing to identify a specific bird, and a more technical term like “raptor” fit the bill. Were Jan’s interest more oriented to references in novels, perhaps “hunting bird” or “bird of prey” would have been more appropriate.

An illustration of a good synonym with proper coverage is:



**Good Synonyms Provide  
Good Coverage**

**TIP**

You can use synonyms both to find the right "level" for your query subject and to ensure proper coverage.

Good coverage is not always possible. Where not possible, provide a couple of alternate terms (that is, synonyms). But, remember, always play the numbers game. Your query terms are limited so choose them carefully.

Having determined the mystery bird to be a **peregrine falcon**, Jan considers whether synonyms for this term are also worthwhile. Based on what Jan has learned, these are the possible synonyms and document counts from AltaVista:

**peregrine falcon\*** 9,157  
**Falco peregrinus** 1,836  
**duck hawk\*** 111  
 all three combined 9,964

Again, note the three synonym counts do not exactly sum due to indexing gaps by the search engines. This example is a good instance where multiple synonyms do not buy enough increased coverage to be warranted. **peregrine falcon** is the most used description of this bird; adding the other terms increases coverage less than 10%.

You need not get actual document counts from search engines in order to weigh such choices in your own queries. Simply use good judgment of what you're gaining – if anything – by adding more synonyms to your query subjects. Common sense should be a sufficient guide.

A thesaurus, a dictionary, personal knowledge or a preliminary Internet search can all be worthwhile places to find synonyms for the major subject(s) in your query. Generally, you should not waste the time thinking about synonyms for other terms in your queries, unless you know them to have very poor coverage.

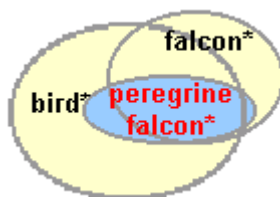
7. What four items can be used to find synonyms to complete your query?



## Use of Phrases

Your most powerful keyword term is the phrase. Phrases are combinations of words that must be found in the search documents in the EXACT order as shown. You denote phrases within closed quotes (“**peregrine falcon\***”). Some search services provide specific options for phrases, some do not allow them at all, but almost all will allow you to enter a phrase in quotes, ignoring the quotations if not supported.

Why phrases are powerful is illustrated below:



### Phrases Target Results

Again, using AltaVista document counts, the ability of phrases to zero in on desired results is clear:

**bird\*** 1,076,900  
**falcon\*** 235,635  
**peregrine falcon\*** 9,157

Phrases should be used where the constituent terms are naturally married. Other examples would be “rain in Spain”, “Gettysburg Address”, “solar system” or “big bad wolf”. Where two or more words are necessary to capture the subject, but may not always be next to one another in the same order, the **AND** or **NEAR** Boolean operators should be used.

#### TIP

Always look for natural phrases in your query concepts – they are one of the most powerful weapons available.

[When using phrases, it is important to consider nuances of the phrase that wouldn't normally be of concern. For example, the spaces between words are as important as any other character. If you include a double space between any two words in the query and the phrase typically has only one, the search will fail. Also, sometimes two dashes are used together on Web documents to approximate an en- or em- dash. If you include only one dash, the search engine may miss all those documents that use two. There is variability in the way certain search engines treat spaces, dashes, and the like. If you suspect there may be a problem, consider submitting your phrases in different ways to capture these variations.]

## RHP Internet Research Class

5-2 Keywords; Then Essence of Search Name:



In addition to “**peregrine falcon\***”, Jan also uses “**endangered species**” to help focus the search. Jan chose “**endangered species**” because information gained in identifying the mystery bird indicated that peregrine falcons were at risk of extinction in the 1970s due to DDT effects. Jan suspects that the answer to the **why** question of the search is the rarity of the bird and not migration or being blown off course. “**endangered species**” is a logical construct for a phrase because the terms are almost always used together to discuss organisms at risk of extinction.

- ☞ 8. What is considered the most powerful keyword term?
- ☞ 9. What special punctuation is used to specify a phrase?
- ☞ 10. When you specify a phrase in a query, what are you looking for?