## 1 GB Search Aid是什么（用来干什么的）？

GB, the Great Britain Architechure. GB是一种数据库架构，这个架构的作用是使不同的数据库都可以使用同一套搜索引擎代码。



如上图，Web of Science核心集，MEDLINE，CABI都是不同的数据库，里面存放不同的数据，但他们都遵守了GB架构，即里面存放的数据都是按照GB要求的格式存放的，所以他们可以使用同一套搜索引擎的代码，而不用每个数据库都搞出一个自己的搜索引擎。

SA, Search Aid. 搜索助手，这个助手是怎么帮助用户完成一次搜索的？

用户在搜索框中输入的内容越专业，搜索出来的结果也就越准确。所谓“专业”就是用户在搜索框中输入的词语是主体词表中的词语。

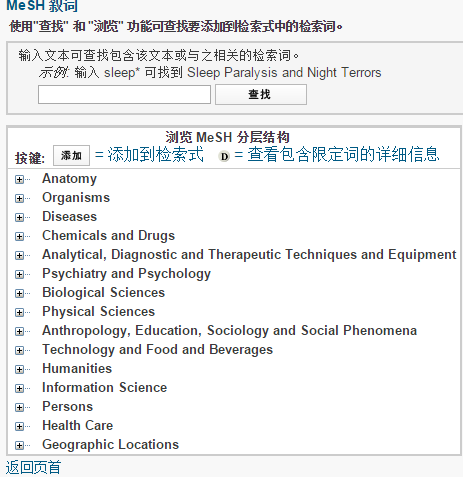
什么是主题词表？主题词表，又称叙词表、检索表或词库。是文献与情报检索中用以标引主题的一种检索工具。 它是一些规范化的、有组织的、体现主题内容的、已定义的名词术语的集合体（来自百度百科“主题词表”）。可是一个主题词表里的词语数以万计，用户怎么知道他要用哪些词语？这就要考Search Aid了，它帮助用户在主题词表中选出合适的词语

## 2如何使用Search Aid?

以在MEDLINE文库中搜索与Vitamin有关的文献举例，说明Search Aid如何帮助用户在主题词表中找到合适的主题词来进行搜索。



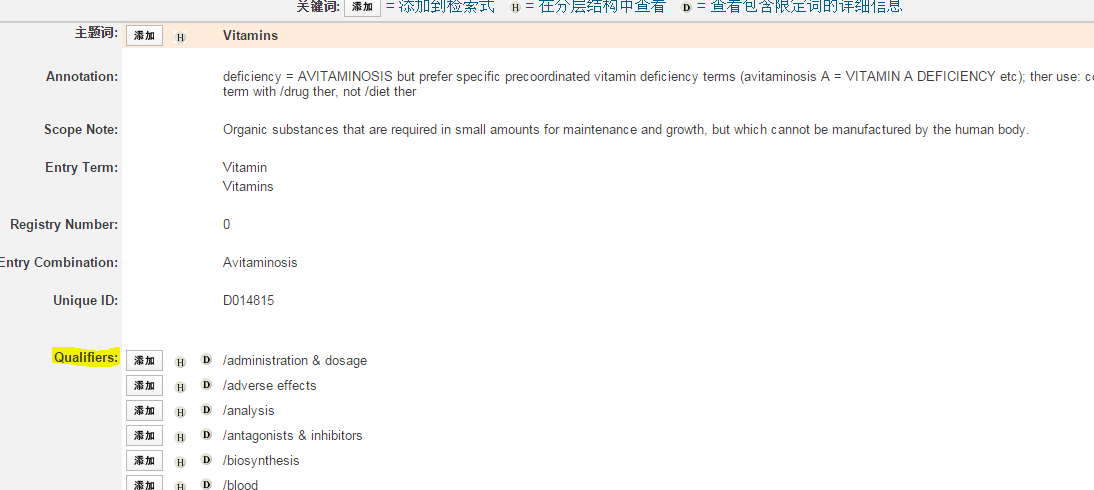
首先选择MeSH主题词，点击下面的“从叙词中选择”（MeSH, Medical Subject Heading医学主题词表，百度百科）。



MeSH中所有的主题词都被分到Anatomy, Organisms, Diseases…..等几个大类。这个分类来自于原始输入文件DescCategories.xml。在输入框中输入”Vitamins”，得到如下结果：



“Vitamin B Complex”和”Vitamins”就是词库中与Vitamin有关的两个叙词（主题词表里的每个词语叫叙词）。点击”H”可以参考该叙词在分类中的层次关系，点击”D”可以查看该叙词的详细信息。点击”D”查看叙词”Vitamins”的详细信息。



在主题词”Vitamins”的详细信息中有一项叫”Qualifiers”，它是副主题词(Subheadings)又称限定词(Qualifiers)，与主题词进行组配，对某一主题词的概念进行限定或复分，使主题词具有更高的专指性（来自百度百科“医学主题词表”）。如果我们选择限定词/metabolism（新陈代谢）（点击该词语前面的“添加”按钮），在搜索框中就会出现如下内容：



这时点击检索，搜索出来的都是关于Vitamin对于新陈代谢影响的文献。

**Ps:大学中有个冷门专业叫“图书馆学”，这个专业会学习如何使用“检索式”进行文献搜索。百度“检索式”，“主题词表”可以搜到关于文献搜索这类的信息。**

文档：

<http://seswiki.int.thomsonreuters.com:8080/docs/Universal_Script.html>

这个文档介绍了什么是GB。

<http://seswiki.int.thomsonreuters.com:8080/docs/mdelaure/sa/FAQ.pod.html>

想认识Search Aid就看这个文档。

例子Abnormalities, Radiation-Induced

See Also：同义词； Previous\_Indexing：以往主题词；

Classification VS Thesaurus

## 3 Search Aid如何工作的？

Search Aid工作过程分两块？第一步要先建立search aid database，第二步就是在search adi database中搜索。

#### 建立search aid database

Search aid filter; load; sort; deduped; index; link; gbdispatch.

1st ./medline\_desc\_qual\_thesaurus previdxrefs.xml desc2014.xml qual2014.xml DescCategories.xml

输入：输入的这四个xml文件一个都不能少，而且输入的顺序不能改变。

输出：desc\_qual\_2006\_entry.xml desc\_qual\_2006\_node.xml

在search aid给出的搜索结果中，一个主题词的详细信息就来自\*\_entry.xml文件，层次结构来自\*\_node.xml文件。

2nd ./gbsa\_load -verbose -name desc\_qual\_thesaurus desc\_qual\_2006\_entry.xml desc\_qual\_2006\_node.xml

输入：-verbose 在屏幕上打印信息；

-name desc\_qual\_thesaurus，在search aid目录下创建.sa目录时用的就是这个名字，config.xml中searchaid标签里的name属性也要是这个名字。

desc\_qual\_2006\_entry.xml

desc\_qual\_2006\_node.xml

输出：创建sa目录，在sa目录下依次创建.search\_aid\_version，data/entries，data/map，data/text，data/nodes文件

一个search aid database就是一个主题词库，database中的一条记录叫entry， 一个entry中记录着关于一个叙词的所有信息。

一个entry 由field, node两部分构成。在实际使用search aid中，一个叙词的详细信息全部来自field，一个叙词的层次结构的信息全部来自node。

Entry存到entries文件，field全部存到text文件，node全部存到nodes文件。Map文件中保存每次向entries文件时的偏移量。

3rd ./ gbsa\_sort -verbose -name desc\_qual\_thesaurus

输入：--verbose 在屏幕上打印信息；

--name 指定对哪个search aid database进行排序；

输出：创建了.sa/index目录，仅仅是创建，其他什么也没做；

生成map.sorted和nodes.sorted文件，里面的内容经过排序的，替换掉原来的map和node文件。

4th ./gbsa\_dedup -verbose -name desc\_qual\_thesaurus

输出：生成map.deduped, entries.deduped, text.deduped和nodes.deduped四个临时文件，用这四个新文件替换掉原来的四个文件。Dedup后database中不会出现有两个entry的key field text相同的情况。Key\_field就是在template.xml中有<key/>标签的fieldtemplate