CabinetOffice

e-Government Unit

Stockley House, 130 Wilton Road
London SW1V 1LQ

E-mail govtalk@cabinet-office.gsi.gov.uk
Web www.cabinet-office.gov.uk/egovernment
www.govtalk.gov.uk

Tomatoes are not the only fruit:
a rough guide to taxonomies, thesauri, ontologies and the like

April 2005

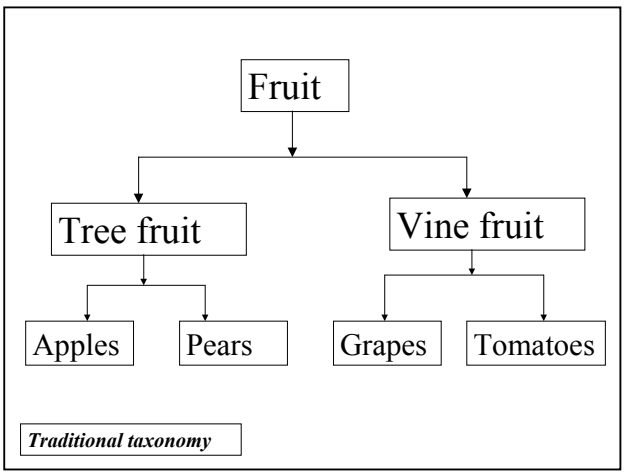
1. This is a brief introduction to the relationships between taxonomies, thesauri and ontologies, and similar 'things'. It doesn't contain definitive, scientific definitions, it is a personal interpretation of some fairly complex structures. It aims to give you a fairly clear what these 'things' are, so librarians or IT people can't blind you with science.

2. **Controlled vocabulary:** this is the only collective name for these 'things' that I can find. It's not exactly sharp and snappy, but it does the job. A controlled vocabulary is list of terms or headings, each one having an assigned meaning. It is designed for classifying, indexing and searching for information resources. Essentially, if everyone uses the same name for the same concept, things become much easier to find.

3. **List:** in this context, a straightforward list of words, in some logical order, usually alphabetical.

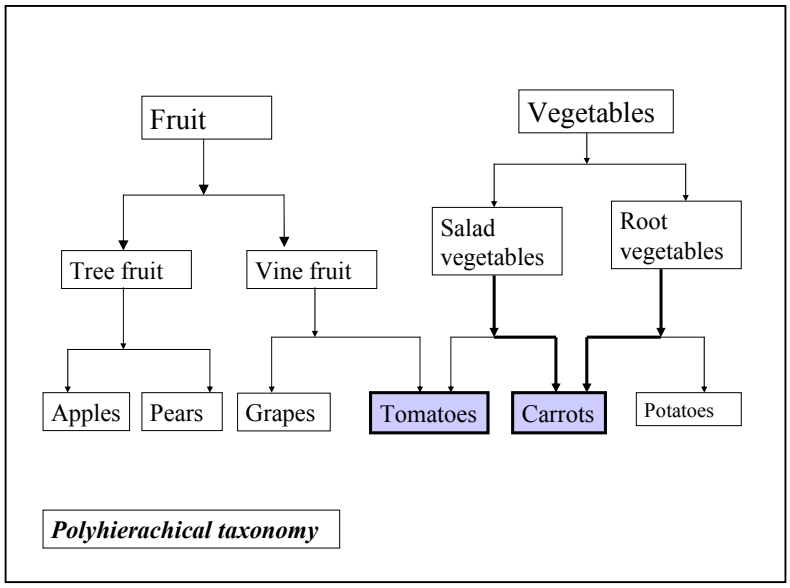
4. **Taxonomy:** a structured list, or 'tree', formed into a hierarchy with broader terms at the top. Taxonomies were once used by biologists to classify living things into species, genera, families, etc. Ideally, each item (or taxon) in a taxonomy should be mutually exclusive and unambiguous, so if 'bats' appears in one place referring to flying mammals, it shouldn't turn up in another place referring to sports equipment. For each term the broader and narrower terms are indicated, and related terms can also be included.

5. One of the problems with new technology is that its complexity stretches the use of language almost beyond recognition. How long has it been for many of us since the primary meaning of the word 'mouse' has been 'a small furry mammal that frightens elephants'? When web designers needed to map out the structure of a web site, they called these maps 'taxonomies'. However the pages on a site are interrelated in many ways, and can usually be reached from many other places, so the structure was more complex than a traditional taxonomy allowed. On top of this 'taxonomy' is now used to describe almost any type of controlled vocabulary. This is where much of the current confusion arises.



A traditional taxonomy – each item exists only once, and in one place. Relationships are vertical.

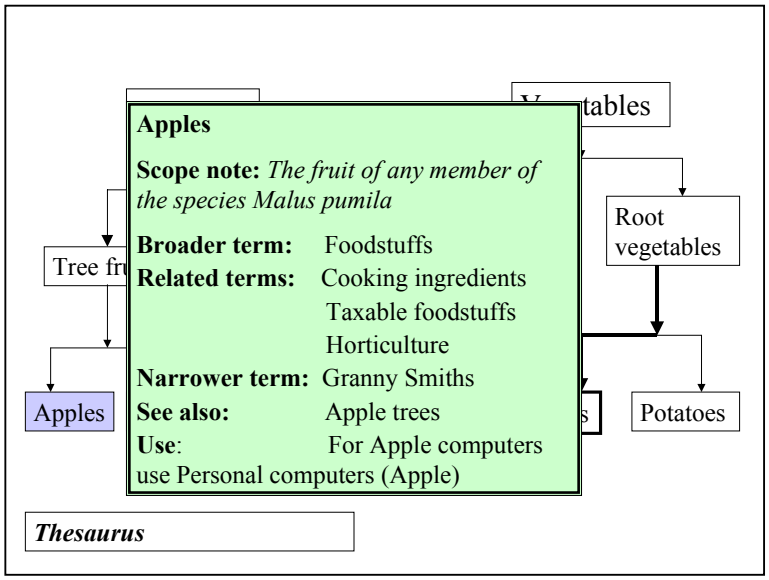
6. **Polyhierarchical taxonomy.** Who needs buzzwords when you can casually drop a 'polyhierarchical taxonomy' into a discussion? Half way between a traditional taxonomy and a fully-fledged thesaurus, it's a taxonomy with additional dimensions, in particular an item can appear in more than one place, so you can reach a given taxon by different routes (hence 'polyhierarchical').



Polyhierarchical taxonomy. 'Tomatoes' can be reached from 'fruit' for those who think like botanists, or by 'vegetables' for the rest of us. Carrots can be either salad or root vegetables.

7. **Thesaurus:** not wanting to be outdone by the techies, librarians borrowed the word 'thesaurus' from Roget and his ilk, and used it to refer to structured sets of the terms used to index information. However, the original concept, a set of synonyms used to make you look more literate, is still the most widely used meaning of 'thesaurus'. This does nothing to reduce the confusion surrounding controlled vocabularies.

8. A thesaurus is taxonomy with extras, it shows lateral connections ('related' and 'see also' terms), has an underlying index showing words that might spring to mind but which you shouldn't use ('non-preferred' terms), and tells you what you should use instead ('preferred' terms). It can be polyhierarchical, and usually contains scope notes to indicate exactly what the term means. A thesaurus lists concepts – the actual words used aren't of paramount importance. While a taxonomy is designed to classify things, a thesaurus is designed to help you find the right words or phrases to describe what you are ultimately looking for. There is an ISO standard for thesauri, and creating a proper one is a major undertaking. The standard (ISO-2788) is an excellent place to go if you want the 'official' definitions of terms such as 'taxonomy' 'thesaurus' 'paradigmatic relationship' and 'enumerated classification scheme'.

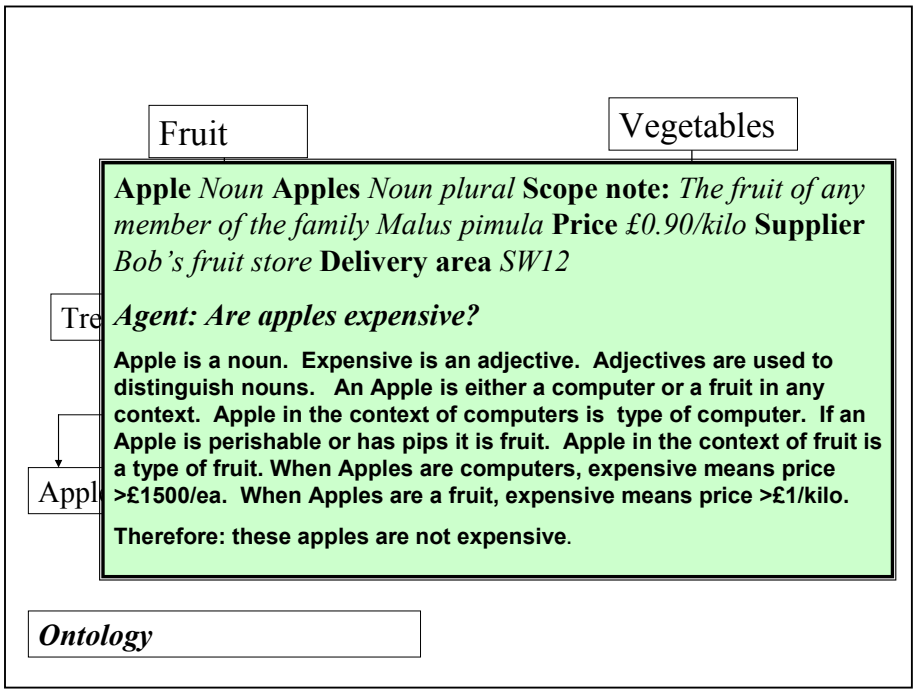


Thesaurus. Behind-the-scenes notes show some of the additional features of a thesaurus. It gives a variety of related terms, and shows that 'apples' can be fruit (as in this entry) or computers.

9. **Ontology:** the new kid on the block, 'ontology' comes from the study of philosophy, and came to its current use by artificial intelligence workers. There is comfort to be gained from knowing that there is no definition that everyone agrees on.

10. Ontologies are designed to allow computers to really interact with each other. For an electronic 'agent' to crawl around the web to find the best holiday, or get you a great deal on a computer, information has to be marked up in such a way that machines can understand it. Remember at this point that, although your PC can calculate *pi* to the *n*th degree, it can't tell the difference between a computer and a granny smith if they are both referred to as 'apples'.

11. An ontology is therefore a thesaurus gone mad. It is more specific in defining a concept or item and its relationships. A dog will be a noun, and an animal, and a mammal, and possibly a pet. Instead of having 'puppies' as a *narrower term* of 'dogs', it might have puppies as *offspring of* dogs. Ontologies often cover all elements of metadata, not just other subject terms. For example, a document could be connected to a person by the relationship 'created by', this person may then connect to an organisation by the relationship 'works for'; to an address by 'lives at' and other documents by 'author of'. This can lead directly from one document to others written by the same person, or by others working for the same organisation that the author works for. Ontologies can be machine generated from good metadata. The semantic web will be built on ontologies.



This shows some of the data an ontology might hold about 'apple' so that the agent can answer the simple questions about the cost. The agent needs a lot of details to follow its logic pattern. Try and imagine the extra details needed to find out if it's worth getting the apples delivered, payment methods, etc. And I specifically want British-grown organic granny smiths.

12. More information about the UK Government's policies on controlled vocabularies, including the Government Category List, can be found at http://www.govtalk.gov.uk/interoperability/gcl.asp

Links to more information about some of the subjects mentioned here can be found in the 'links' section of www.govtalk.gov.uk. For more details about the semantic web and ontologies, I particularly recommend the W3C site.