efficient indexing for advanced document management

Vervenne D.

BIKIT, Gent

Some milestones…

Clay tablets 3500 BC

Printing on paper: Gütenberg 1454

Typewriter by Remington, 1874

windows-PC at Xerox Parc, 1973

Wwweb at Cern/Ncsa, 1993

What is indexing ?

How can we index ?

Thesaurus-based indexing ?

document indexing: what ?

Indexing is the process of creating **meta-data** about the document(s)

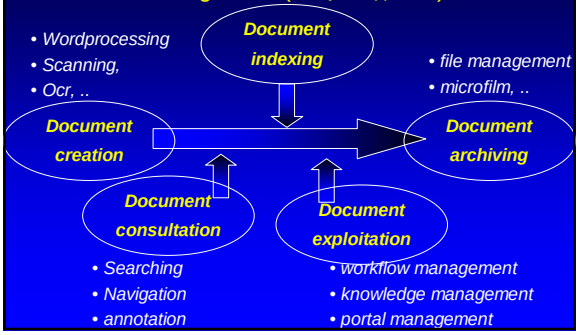
these meta-data represent knowledge about the documents, in order to classify and retrieve them (cf search engine).

document indexing: history ?

When printing was invented, documents had no title.

Titles have been added to books for indexing purpose in order to **classify** them by librarians.

Indexing as the key to efficient document management (life cycle approach)



document indexing: problem ?

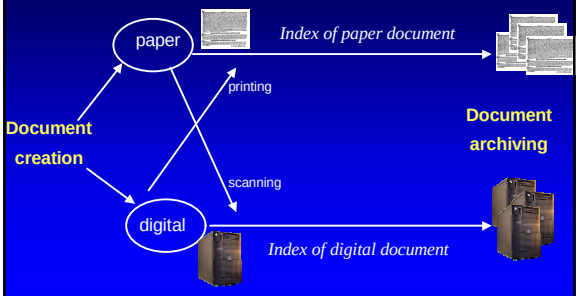
How to classify documents such that they are retrievable in the most efficient way ?

- using authority files with reference terms or codes ?

- linking all relevant semantic associations ?

Vervenne ©

document indexing: synchronizing and version management ?



Vervenne ©

What is indexing ?

How can we index ?

- manual

- automatic

thesaurus-based indexing ?

1: manual document indexing

Authors of documents can add meta-data, e.g.:
- give a **relevant** name to the file
- add **keywords** in pre-defined fields
- add headers or footers with relevant meta-data

(specialized) Indexers can add meta-data, e.g.:
- fill in external file forms
- fill in separate databases with links to the doc.

Vervenne ©

manual document indexing: exampl.

Filenames can contain meta-data about their content:

e.g.:

- *heldere_fazantensoep_met_sherry.txt*
- *final-report-Octopus-version3.pdf*

some names have an internal logic (and are mostly cryptic for non-experts)

e.g.:

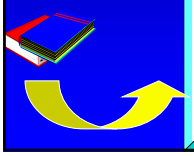
- *1996sep10_eindtermen.doc*
- *MOC-051299-DVF123a.doc*

Vervenne ©

(manual) document indexing

Vervenne ©

Indexing is traditionally a (formal) activity done by the expert **librarian**:
s/he wanted to classify books and journals in order to retrieve them easily (by paper forms)



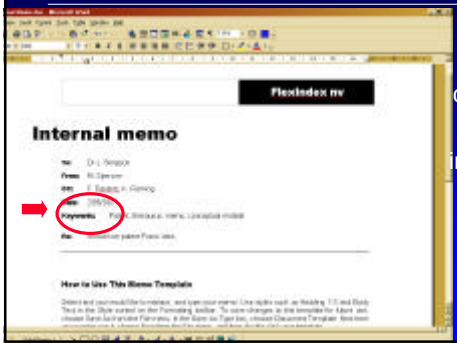
Cohen M. **BCS-78435**

An introduction to Logic

1972

Routledge & Kegan, London
ISBN O 7100 1197 9 **pp465**

Manual Indexing info in the document

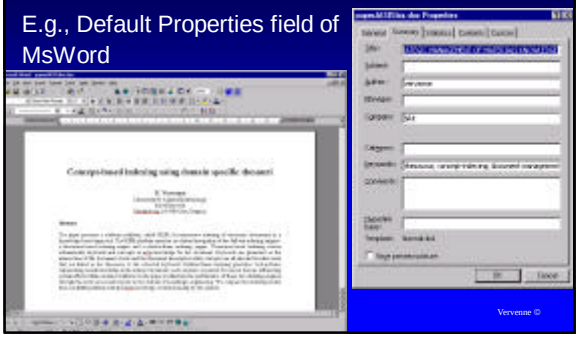


E.g., through customized fields in the word-processor

Vervenne ©

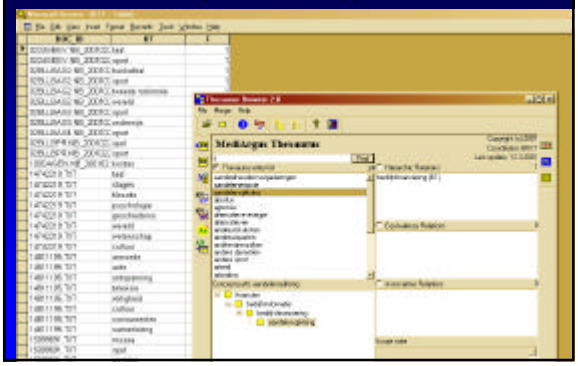
Manual Indexing info in the document

E.g., Default Properties field of MsWord



Vervenne ©

field indexing with manual support of thesaurus



2: Automatic document indexing

How can we extract meta-data automatically ?

- Statistical approach ?
- Neural network approach ?
- Thesaurus-based approach ?

Vervenne ©

statistical indexing

The **frequency** of all individual words in the document is counted and stored in an alphabetic index.

This approach is used for 'full text' search e.g. Internet search.

Our cat

Our cat is very lazy. She sleeps all the time. The mouse and the dog, living next door, are not afraid of our cat. The result is a very friendly neighbor-hood.

Stopword list

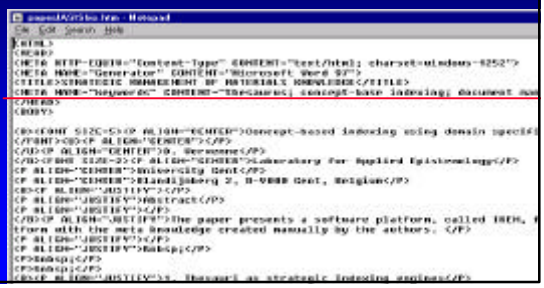
Full index

Afraid 1
cat 3
dog 1
is 3
lazy 1
living 1
neighborhood 1
next 1
our 3
result 1
she 1
sleeps 1
time 1
the 4

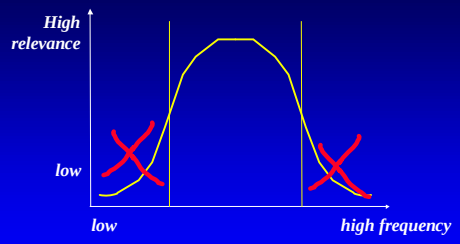
Vervenne ©

Manual Indexing info in the document

Meta Tags in HTML for Web Search Engines



Theory of Luhn, 1958



"The relevancy of a word to index a document is related to the word frequency"

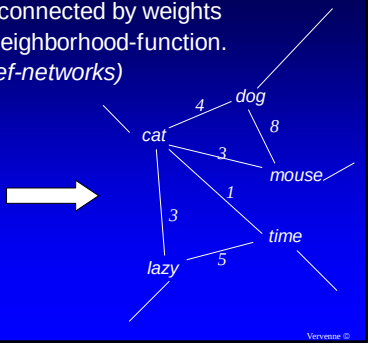
Vervenne ©

neural network indexing

All words are interconnected by weights on the basis of a neighborhood-function.
(cf Bayesion belief-networks)

Our cat

Our cat is very lazy. She sleeps all the time. The mouse and the dog, living next door, are not afraid of our cat. The result is a very friendly neighbor-hood.



Vervenne ©

Neural networks: usage

Neural networks are efficient for very large collections of documents …

… But the indexing process operates as a black box …

Cf. Autonomy

What is indexing ?

How can we index ?

Thesaurus-based indexing ?

What is a thesaurus ?

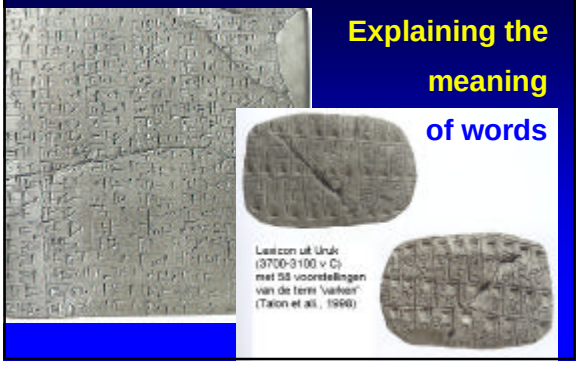
"The thesaurus (...) is a vocabulary of controlled indexing language,

formally organized so that a priori relationships between concepts are made explicit,

to be used in information retrieval systems, whether these are databases, or printed indexes or catalogues"

Aitchison et ali., 1997

What is a thesaurus ?



Explaining the meaning of words

What is a thesaurus ?: more than synonyms

Webster's Dictionary refers to Greek word 'thesauros' (Gr. θησαυρος) to be translated as "treasure room" :

L., treasure, store, collection, Gk thesauros: A book containing a store of words or of information about a particular field or set of concepts.

Webster Third New International Dictionary, 1971, Vol. III , p. 2374.

(MS-Word does not support a thesaurus !!)

First thesaurus-idea from John Wilkins, 1668 *'Essay towards a Real Character and a Philosophical Language'*.

Encyclopedia Britannica Macropaedia, Vol. 5, 1974, p. 719.

What is a thesaurus ?: 2 ISO norms

International Organisation for Standardization. *ISO 5964: Guidelines for the Establishement and Development of Multilingual Thesauri*. 2nd ed. - Geneva: ISO, 1985.

International Organisation for Standardization. *ISO 2788: Guidelines for the Establishement and Development of Monolingual Thesauri*. 2nd ed. Geneva: ISO, 1986.

What is a thesaurus ? ..ISO-approach

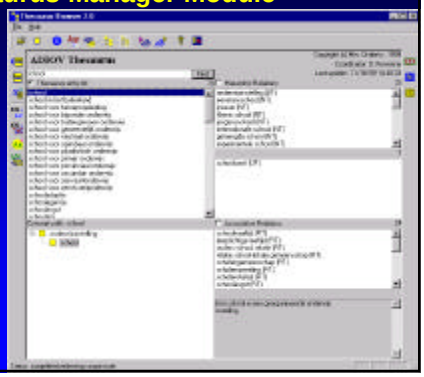
A semantic network of words with standardised relations:

- hierarchic relations (broader/narrow terms)
  - e.g., *Europe --> broader term --> Belgium*
- equivalent relations (use/use for relations)
  - e.g., *VRT -->use for --> Vlaamse Radio Televisie*
- related terms (rt relations)
  - e.g. *stripverhaal --> rt --> ontspanning*
- *scope notes (example)*
  - e.g. "*an AKB is a monitor module for automatic adjustment of the phosphor radiation"*

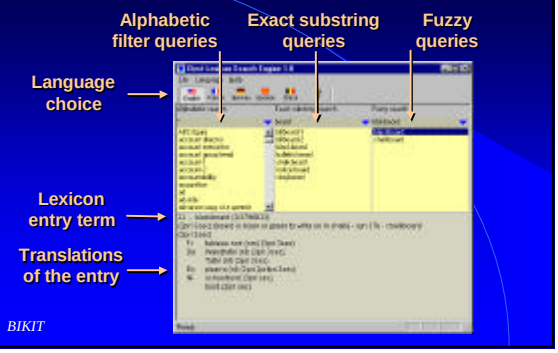
IKEM Thesaurus Manager Module

ISO-compatible database format

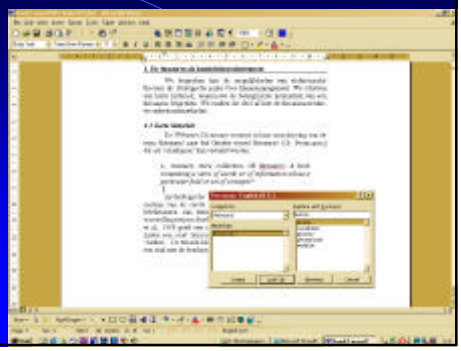
(mono-lingual)

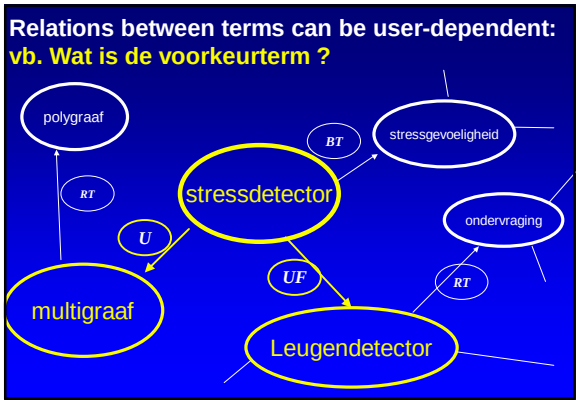


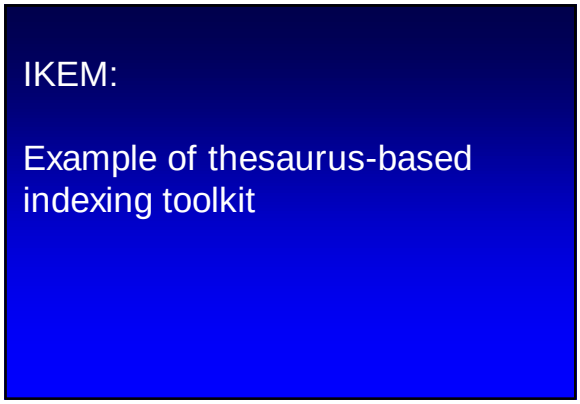
A lexicon is Not a thesaurus !!

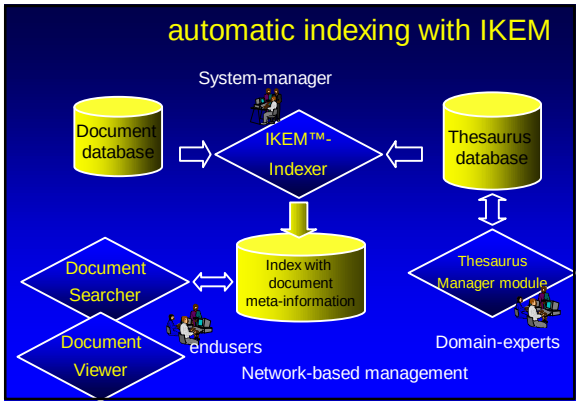


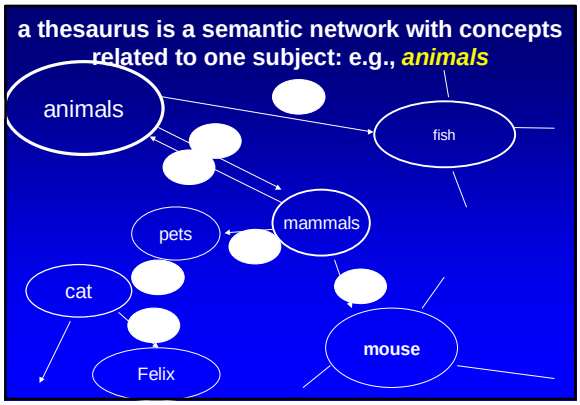
A thesaurus in MS.Word is Not a thesaurus !!

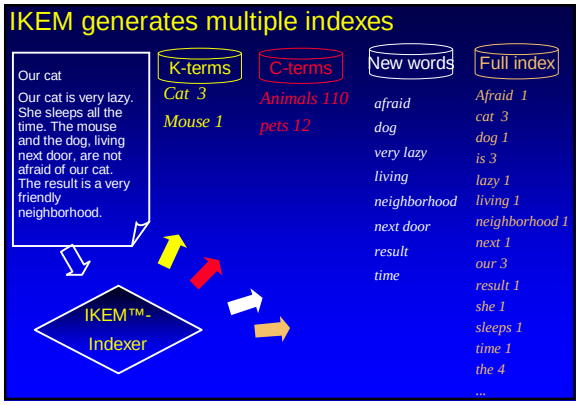


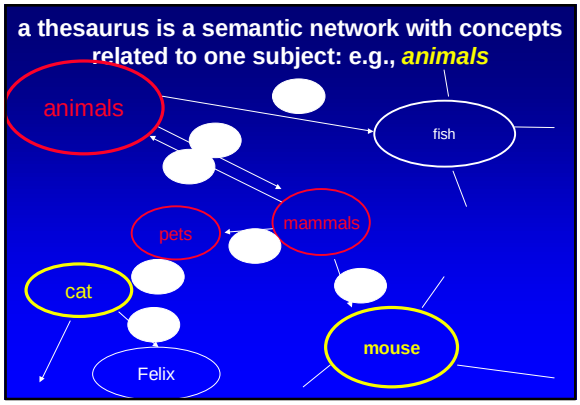

Relations between terms can be user-dependent:
vb. Wat is de voorkeurterm ?
polygraaf
BT
stressgevoeligheid
RT
stressdetector
ondervraging
U
UF
RT
multigraaf
Leugendetector

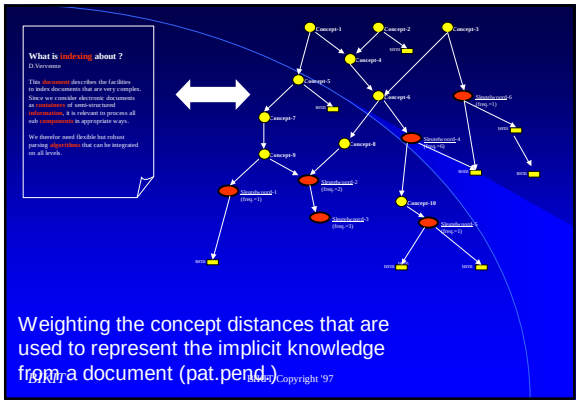

IKEM:
Example of thesaurus-based indexing toolkit

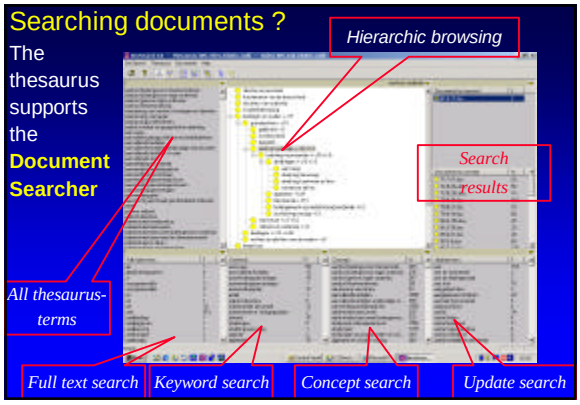

automatic indexing with IKEM
System-manager
Document database
IKEM™-Indexer
Thesaurus database
Index with document meta-information
Document Searcher
Document Viewer
Thesaurus Manager module
endusers
Domain-experts
Network-based management

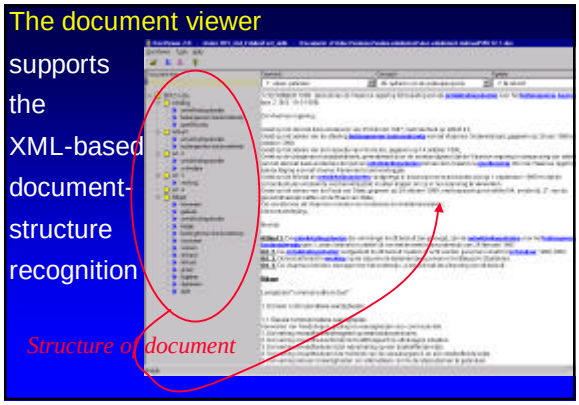

a thesaurus is a semantic network with concepts related to one subject: e.g., animals
animals
fish
pets
mammals
cat
Felix
mouse

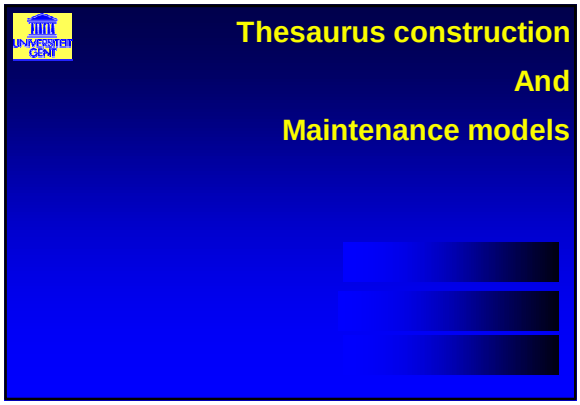

IKEM generates multiple indexes
Our cat
Our cat is very lazy. She sleeps all the time. The mouse and the dog, living next door, are not afraid of our cat. The result is a very friendly neighborhood.
K-terms
Cat 3
Mouse 1
C-terms
Animals 110
pets 12
New words
afraid
dog
very lazy
living
neighborhood
next door
result
time
Full index
Afraid 1
cat 3
dog 1
is 3
lazy 1
living 1
neighborhood 1
next 1
our 3
result 1
she 1
sleeps 1
time 1
the 4
...
IKEM™-Indexer

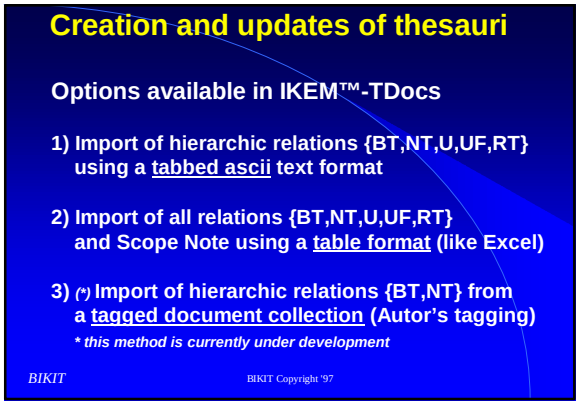

a thesaurus is a semantic network with concepts related to one subject: e.g., animals
animals
fish
pets
mammals
cat
Felix
mouse

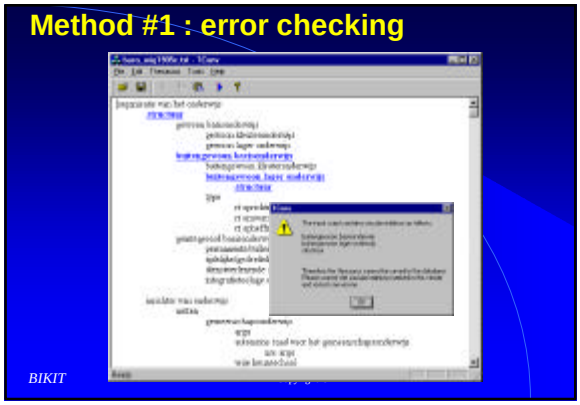

What is indexing about ?
D.Vervenne
Weighting the concept distances that are used to represent the implicit knowledge from a document (pat.pend.)
BIKIT
BIKIT Copyright '97


Searching documents ?
The thesaurus supports the Document Searcher
Hierarchic browsing
Search results
All thesaurus-terms
Full text search
Keyword search
Concept search
Update search


The document viewer
supports the XML-based document-structure recognition
Structure of document


Thesaurus construction
And
Maintenance models

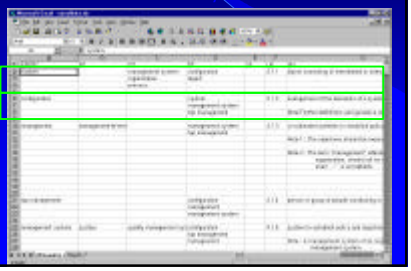

Creation and updates of thesauri
Options available in IKEM™-TDocs
1) Import of hierarchic relations {BT,NT,U,UF,RT} using a tabbed ascii text format
2) Import of all relations {BT,NT,U,UF,RT} and Scope Note using a table format (like Excel)
3) (*) Import of hierarchic relations {BT,NT} from a tagged document collection (Autor's tagging)
* this method is currently under development
BIKIT
BIKIT Copyright '97


Method #1 : error checking
BIKIT

Method #2 : filling in a relations table

- A table is used with the following column headings:
  | ENTRY | BT | NT | RT | UF | LB | SN |
- Relations are entered in table rows, in this case in Excel

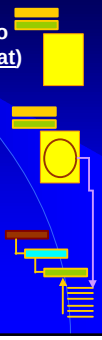
relations of "system"
relations of "configuration"
...



BIKIT

Method #3 : using tagged documents

- **Step 1**
  Author's classification terms are attached to documents as tags (preferably in XML format)
- **Step 2**
  Relevant expressions are extracted from document contents and saved in the update table
- **Step 3**
  The update table is processed and relevant terms are attached to the authors classification terms in the thesaurus

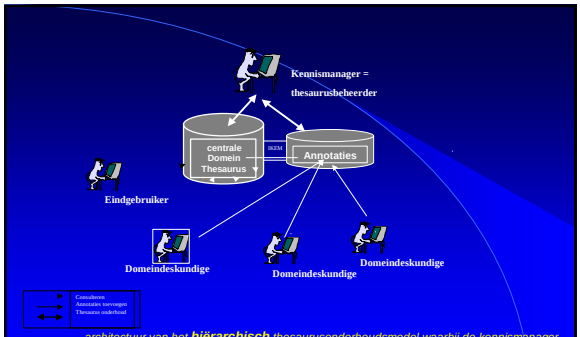


BIKIT BIKIT Copyright '97

How to maintain a thesaurus ?

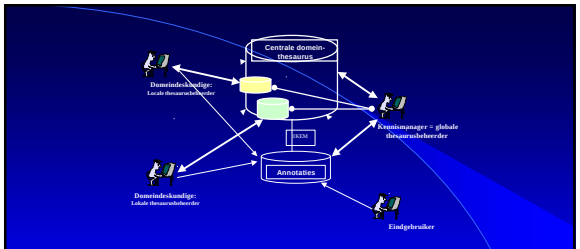
3 workflow models:

- *hierarchic workflow*
- *consensus workflow*
- *heterarchic workflow*



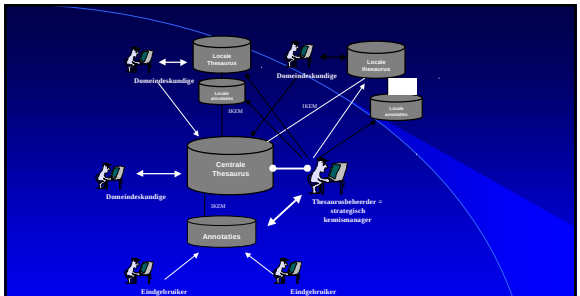
*architectuur van het **hiërarchisch** thesaurusonderhoudsmodel waarbij de kennismanager exclusieve beheerverantwoordelijkheden heeft. In dit model is sprake van één enkele coöperatiegroep waarbij de verantwoordelijkheid geconcentreerd is bij de kennismanager.*

BIKIT BIKIT Copyright '97



*architectuur van het **consensus** thesaurusonderhoudsmodel: de domeindeskundigen kunnen op lokale plaatsen in de thesaurus termen toevoegen gezien ze verantwoordelijk zijn voor een thesaurusonderdeel. In deze benadering is er sprake van één coöperatiegroep met gedeelde verantwoordelijkheid voor het thesaurusonderhoud.*

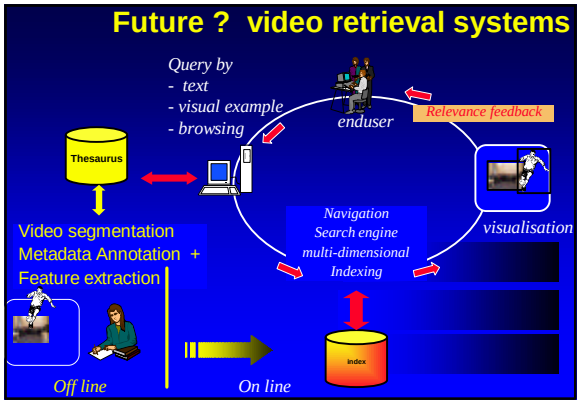
BIKIT BIKIT Copyright '97



*architectuur van het **heterarchisch** thesaurusonderhoudsmodel: elke domeindeskundige heeft een eigen thesaurus waarvoor hij verantwoordelijk is. Dit model is technisch zeer complex doch stimuleert hybridische ontologieën*

BIKIT BIKIT Copyright '97


Conclusion ?
Indexing is a strategic process
with important implications for many document
management processes:
new trend is
to start the indexing process
during the creation process
and to integrate with content managem.


Future ? video retrieval systems
Query by
- text
- visual example
- browsing
enduser
Relevance feedback
Thesaurus
Video segmentation
Metadata Annotation +
Feature extraction
Navigation
Search engine
multi-dimensional
Indexing
visualisation
index
Off line
On line