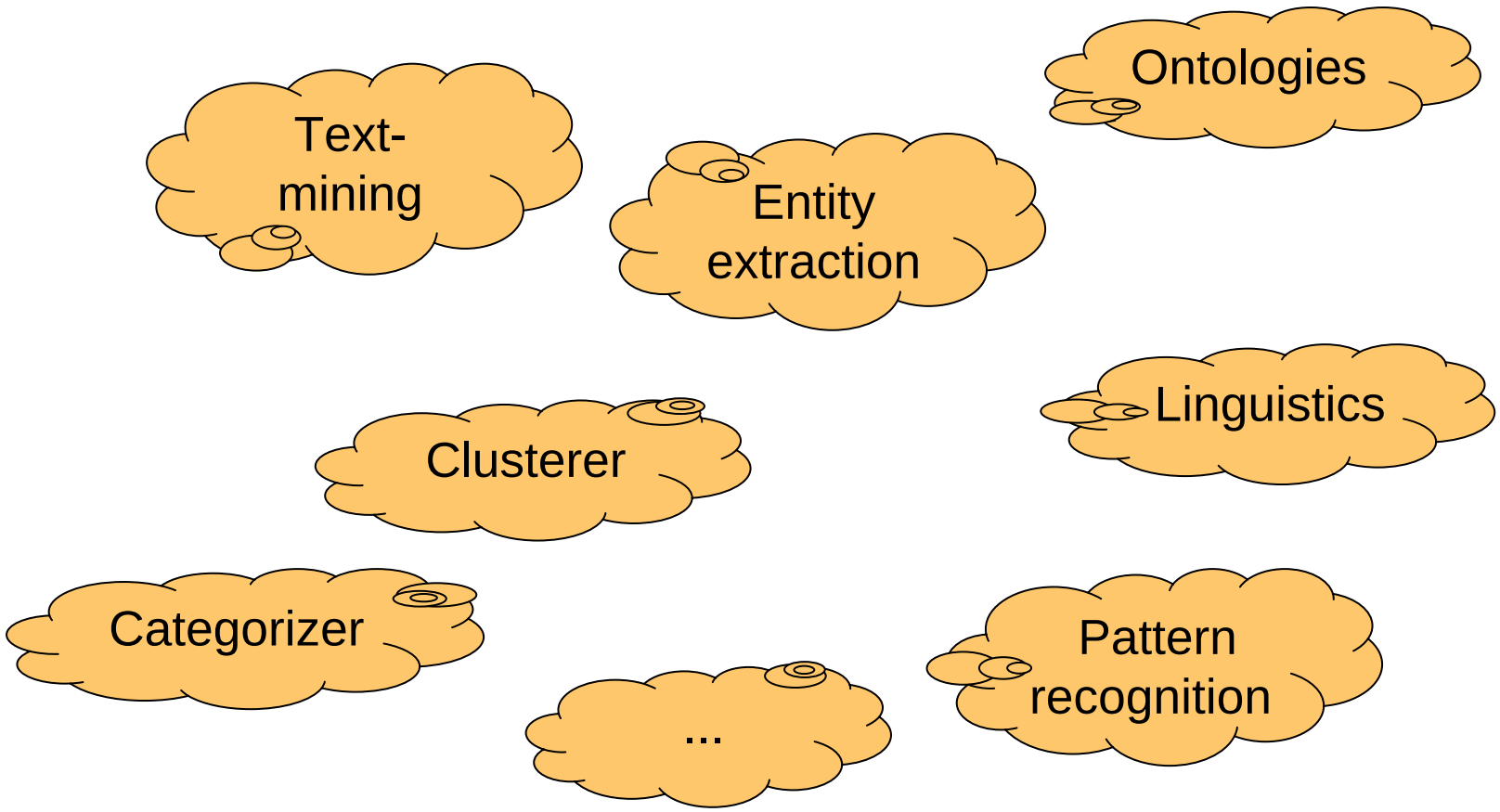
Semantic Technologies in a Chemical Context

Quick wins and the long-term game

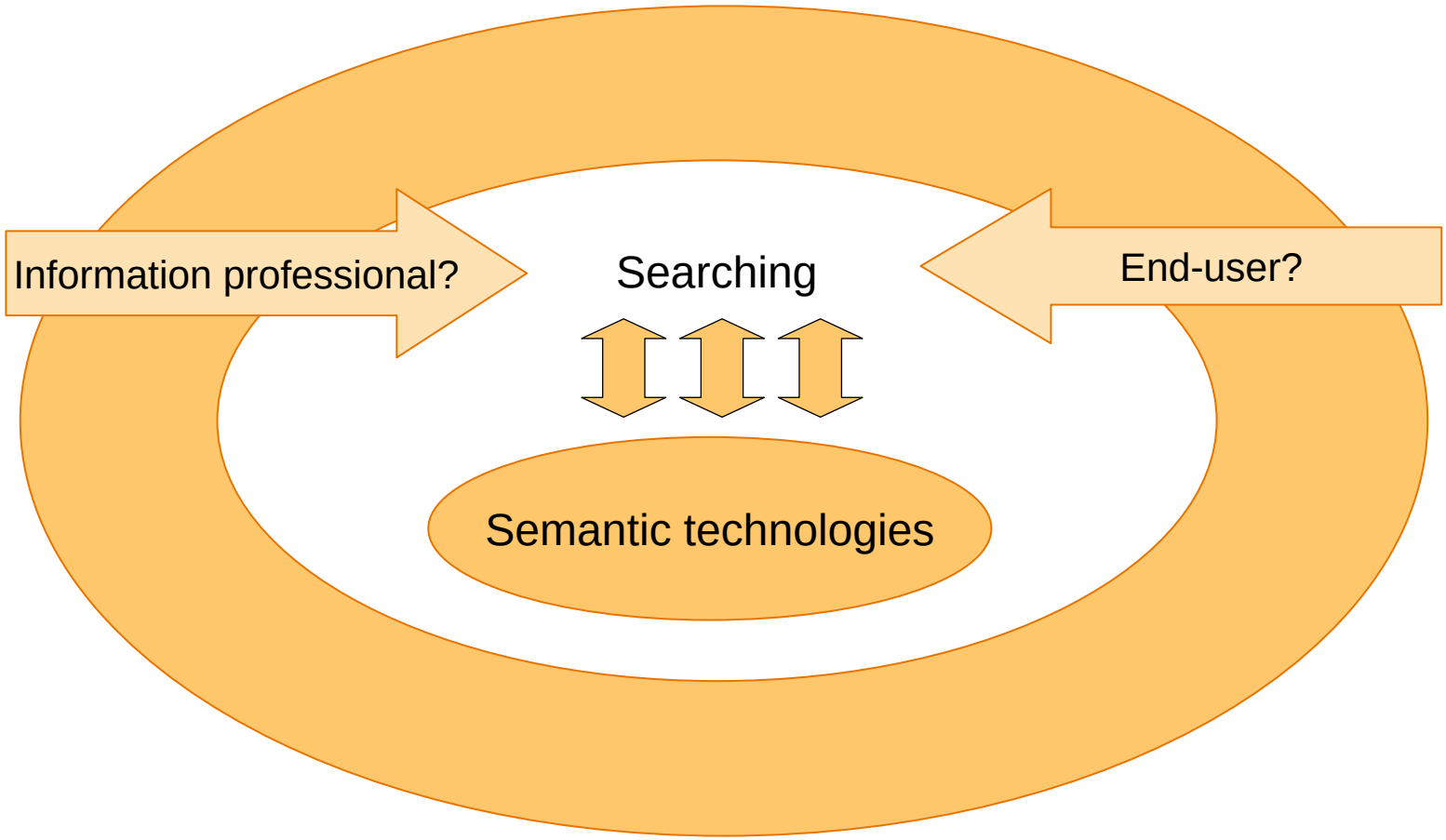
Dr. Heinz-Gerd Kneip
BASF SE

The International Conference on Trends for Scientific Information Professionals
ICIC 2010, 24.-27.10.2010, Vienna

Semantic technologies =



What the talk is about

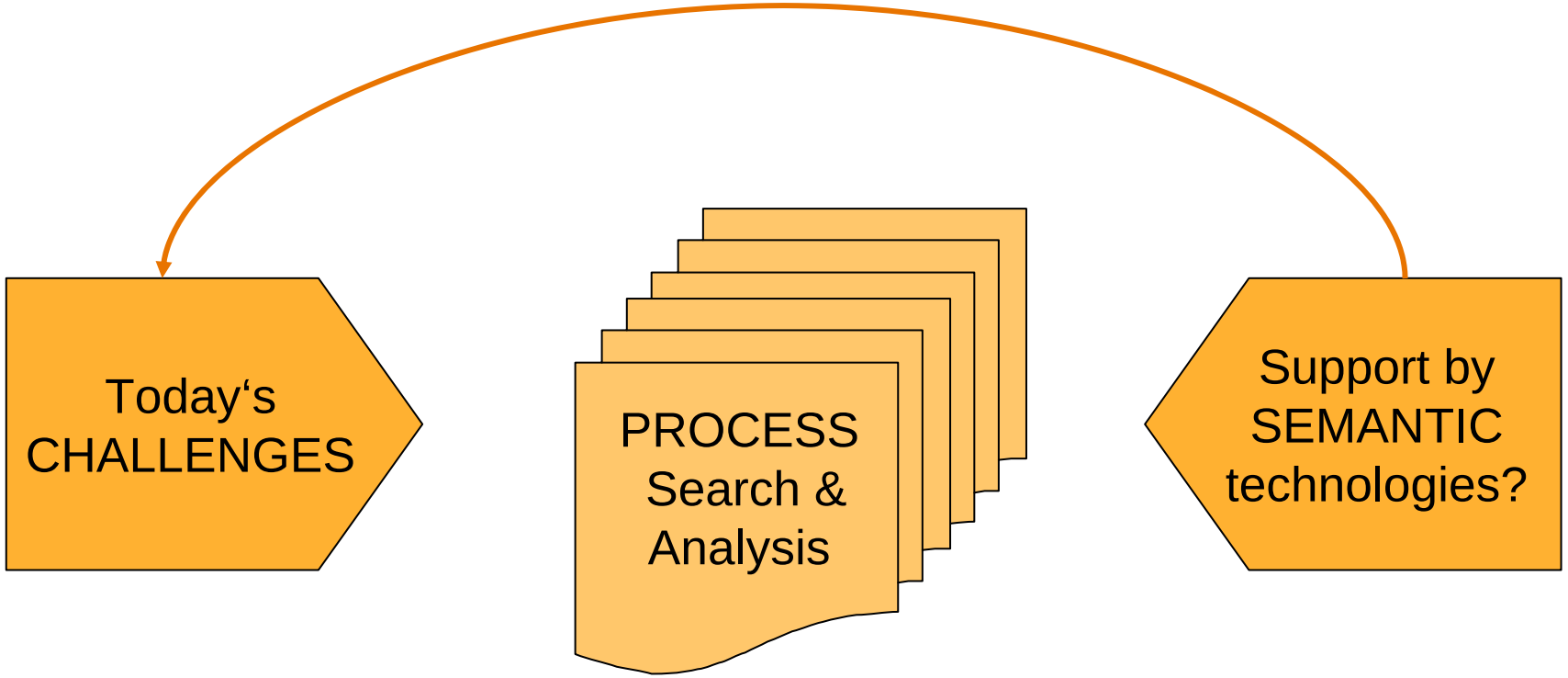


End-users and information professionals: Searching: different behaviours and expectations

	End-users	Information professionals
Information sources	„Which sources?“	Carefull choice
Querying	Expect the system to be intelligent	Mistrust black boxes, rely on own competency and do everything to crack the nut
Search result	Some good answers	As comprehensive as required

Totally different needs during search (and analysis)
→ Today focus on information professionals

Professional search & analysis
Challenges → semantic solutions?



Main topic

Examples

Professional search & analysis
(Idealized) process

- Demands of customers/partners
- Selection of sources
- Building query strategies
- Searching + result download
- Relevance-checking
- Consolidation for post-processing
- Data-cleanup
- Data enrichment
- Analysis, visualization, reporting
- Result delivery

- *Intellectually: hit (yes / no)?*
- *Many formats to one format*
- *Assignee names; patent families*
- *Indexing, annotation, categorizing*
- *Tables, charts, ...*
- *Supporting tools*

Professional search & analysis
Challenges (i)

- The complexity of search requests for information professionals is ever growing 1
- Database indexing should increase recall and precision: where it is not the case 2
- Increasingly (patent) full text databases have to be searched in addition to indexed abstract databases 3
- Due to large result sets sometimes documents can no longer be relevance checked before analysis 4

Professional search & analysis
Challenges (ii)

- When you want to mine information to add value, you perhaps are not allowed to mine the information 5
- More often customers want answers to their questions instead of documents only as search result 6
- You have to analyze 10.000 documents or more, but the software asks you to limit your search result to < 1.000 7
- We have to deliver search results in a way that customer‘s processes are supported 8

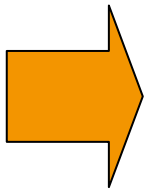
Professional search & analysis
Challenges (iii)

BASF
The Chemical Company

Disclaimer

At no time the presenter had the intention to be complete:

There are much more challenges!



8 Short stories for 8 challenges:

- with examples for semantic solutions
- asking if there might be a semantic solution
- even without relation to semantic technologies

Challenges:
Complexity of requests (i)

1

The complexity of search requests for information professionals is ever growing

- The strategy of BASF‘s Info Center is to provide relevant information sources for end-users through an information portal for the less critical searches to utilize the resources of the well educated information professionals for the business critical searches and analyses

- Search requests for information professionals changed with the increasing interdisciplinarity and complexity of research and globalizing markets

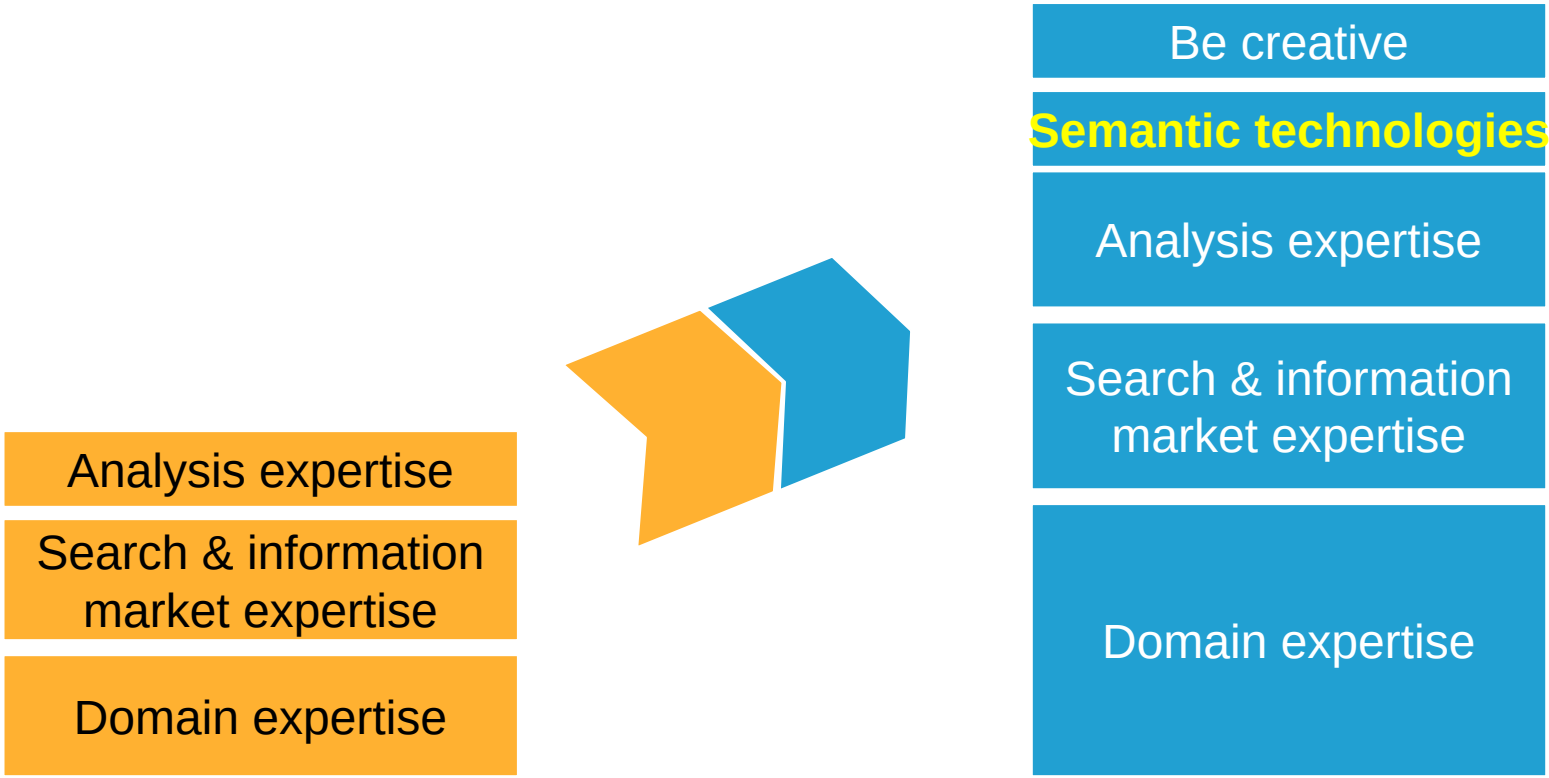
Challenges:
Complexity of requests (ii)

- Search requests changed from
  - structures → systems
  - anything published about xyz → what is not patented in area xyz
  - simple problem solutions → technology and competitor monitoring

- Typical requests:
  - which formulation techniques are used by which companies?
  - emerging white biotech companies in China?
  - which intermediates can be used in tomorrows drugs?

Challenges:
Complexity of requests (iii)

- Due to increasing complexity of requests the profile for information professionals has changed and is changing:



Challenges:
Inadequate database indexing (i)

2

BASF
The Chemical Company

Database indexing should increase recall and precision:
where it is not the case

- Use case: Search ternary compositions of A + B + C

- Ternary compositions of A + B + C are rarely indexed as described
- As a consequence we have to search documents describing optionally A and/or optionally B and/or optionally C
- This increases the number of documents to be relevance-checked dramatically

Challenges:
Inadequate database indexing (ii)

- Since many years patenting compositions, formulations, mixtures has increased rapidly. This is true for most of the domains in chemistry: polymers, performance chemicals, agrochemicals, catalysts, ...

- Searching for compositions, formulations, mixtures is a daily challenge

???

How can semantic technologies
assist to enrich abstracts or
full text to more precisely identify
compositions, formulations, mixtures?

???

Challenges:
Abstract plus full text (i)

BASF
The Chemical Company

Increasingly (patent) full text databases have to be searched in addition to indexed abstract databases

- Esp. newer technologies (nanotechnology, green biotechnology) do not have precise coding for searching
- Specific terms of a domain or common names often can only be searched and found in full text
- Search hits from codes in abstract databases often can only be reviewed when having the appropiate paragraph from full text

- Once again the number of documents (with full text: also the number of pages) to be relevance-checked increases heavily

Challenges:
Abstract plus full text (ii)

- Wanted: software which assists to relevance-check (and later to deliver) documents from different abstract und full text databases (in a patent family context).
  Ideally relevant parts can be merged in a result set to be delivered (record from database A + indexing from database B + claims from database C)

- Where are we going?
  Is there an unavoidable need in the (near?) future to semantically process full-text to
  - bypass deficiencies of indexing in abstract databases
  - substitute intellectual database indexing as it becomes too costly due to the sheer number of published documents?

Challenges:
Intellectual relevance-check impossible (i)

4

BASF
The Chemical Company

Due to large result sets sometimes documents can no longer be relevance checked before analysis

- Often search results with 10.000 to 100.000 documents
- Relevance-checking is elementary to remove false positives being related to generic search terms – otherwise during analysis biased results will be produced
- Why does generic searching produce false positives?
  search for: anti-wrinkling composition
  also search for: skin cosmetics
  anti-acne ointment *as false positive*

Challenges:
Intellectual relevance-check impossible (ii)

- - Broad search for green biotechnology
  - How to remove breeding, pharma, nutrition, ...?

- Solution with semantic technologies
  - Develop categorizers (with TEMIS LUXID®)
  - Categorize complete search result
  - Clever validation strategies needed to control correctness of black box approach

- But:
  - Parameters for domain A not transferable for domain B
  - One never knows before, if and when categorizing will be successfull or not
  - Process support and perfomance for categorizing can be improved

Challenges:
You are not allowed to mine the data (i)

5

BASF
The Chemical Company

When you want to mine information to add value, you perhaps are not allowed to mine the information

- Sometimes the terms & conditions for data usage read like: „you can search and read the retrieved data – afterwards please destroy the data without utilizing them with additional (not allowed) techniques to generate value“
  (note from the presenter: this phrasing is only close to – but not identical with - reality)

- Providers want to protect their investment e.g. in indexing, but for additional money we can get rights to mine data and add specific value

Challenges:
Answers to questions (i)

BASF
The Chemical Company

More often customers want answers to their questions instead of documents only as search result

- Use case: which companies (esp. in emerging countries) are using which formulation techniques?
  - Define emerging countries
  - Search for formulation with max. recall → precision: --
  - No. of documents too high to be relevance-checked
  - Extract entities, check for technologies and consolidate
  - Relevance-check those docs with extracted technologies
  - Clean-up and group company information
  - Present results in charts, tables or reports (with links to docs)

Challenges:
Answers to questions (ii)

- Extract technologies and consolidate: solution approach

  - untyped entity extraction (TEMIS relevant term cartridge or inhouse solution for concept extraction)
    → list of entities plus doc frequency + link to doc
  - check lists for hit concepts with domain experts or search for patterns in the list
  - consolidate concepts (synonyms, clusters)

Use results to built up a cartridge for latter requests

Challenges:
Limits in analysis tools (i)

BASF
The Chemical Company

You have to analyze 10.000 documents or more, but often analysis software asks you to limit your search result to < x00 or x.000

- Inhouse tool Maurice: works with acceptable performance up to 10.000 documents
- TEMIS Luxid®:
  - performance for entity display has been optimized heavily
  - matrix limited to a few 100 documents
  - semantic net limited a few 100 documents

→ Analysis tools far away from real life challenges

Challenges:
Result delivery (i)

BASF
The Chemical Company

We have to deliver search results in a way that customer‘s processes are supported

- Result delivery: ideally a combination of
  - the relevant documents
  - extracted and cleaned-up concepts for navigation and analysis
  - easy functionalities to process the information in a business context

- Collaboration: shared commenting and annotating, distributing
- Analysis and visualization templates
- Understand the content quickly: alternatives to reading text

Challenges:
Result delivery (ii)

- Example: how to use untyped entities to understand content quickly

TRANSCRIPTION FACTOR E2FA PROTEIN USEFUL IN INCREASING SALT TOLERANCE IN PLANTS AND OBTAINING HIGHER GROWTH RATE AT ELEVATED SALT CONCENTRATIONS

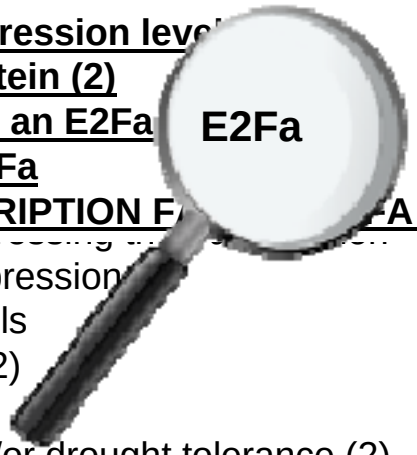
Use of an E2Fa protein to obtain salt and/or drought tolerance in eukaryotic cells. An INDEPENDENT CLAIM is also included for a method to obtain a salt tolerant plant, comprising a mutation resulting in the increase of the E2Fa expression level. The eukaryotic cells in obtaining salt and/or drought tolerance using an E2Fa protein are plant cells, or constitute a whole plant. The use is overexpression of the protein that is realized by operably linking the gene, encoding an E2Fa protein to a strong heterologous promoter, preferably a constitutive promoter that is the 35S promoter (all claimed). The methods and compositions of the present invention are particularly useful in increasing salt tolerance in plants by overexpressing the transcription factor E2Fa to obtain higher growth [illegible] elevated salt concentrations.

Automatic extraction

35S promoter
constitutive promoter
E2Fa expression level
E2Fa protein (2)
elevated salt concentrations (2)
encoding an E2Fa protein
eukaryotic cells (2)
factor E2Fa
higher growth rate (2)
increasing salt tolerance (2)
Mutation
operably linking the gene
overexpressing the transcription
Overexpression
plant cells
Plants (2)
Protein
salt and/or drought tolerance (2)
salt tolerant plant
strong heterologous promoter
TRANSCRIPTION FACTOR E2FA PROTEIN
whole plant

Challenges:
Result delivery (iii)

35S promoter
constitutive promoter
E2Fa expression level
E2Fa protein (2)
elevated salt concentrations (2)
encoding an E2Fa protein
eukaryotic cells (2)
E2Fa expression leve
E2Fa protein (2)
encoding an E2Fa
factor E2Fa
TRANSCRIPTION F **FA PROTEIN**
Overexpression
plant cells
Plants (2)
Protein
salt and/or drought tolerance (2)
salt tolerant plant
strong heterologous promoter
TRANSCRIPTION FACTOR E2FAPROTEIN
whole plant

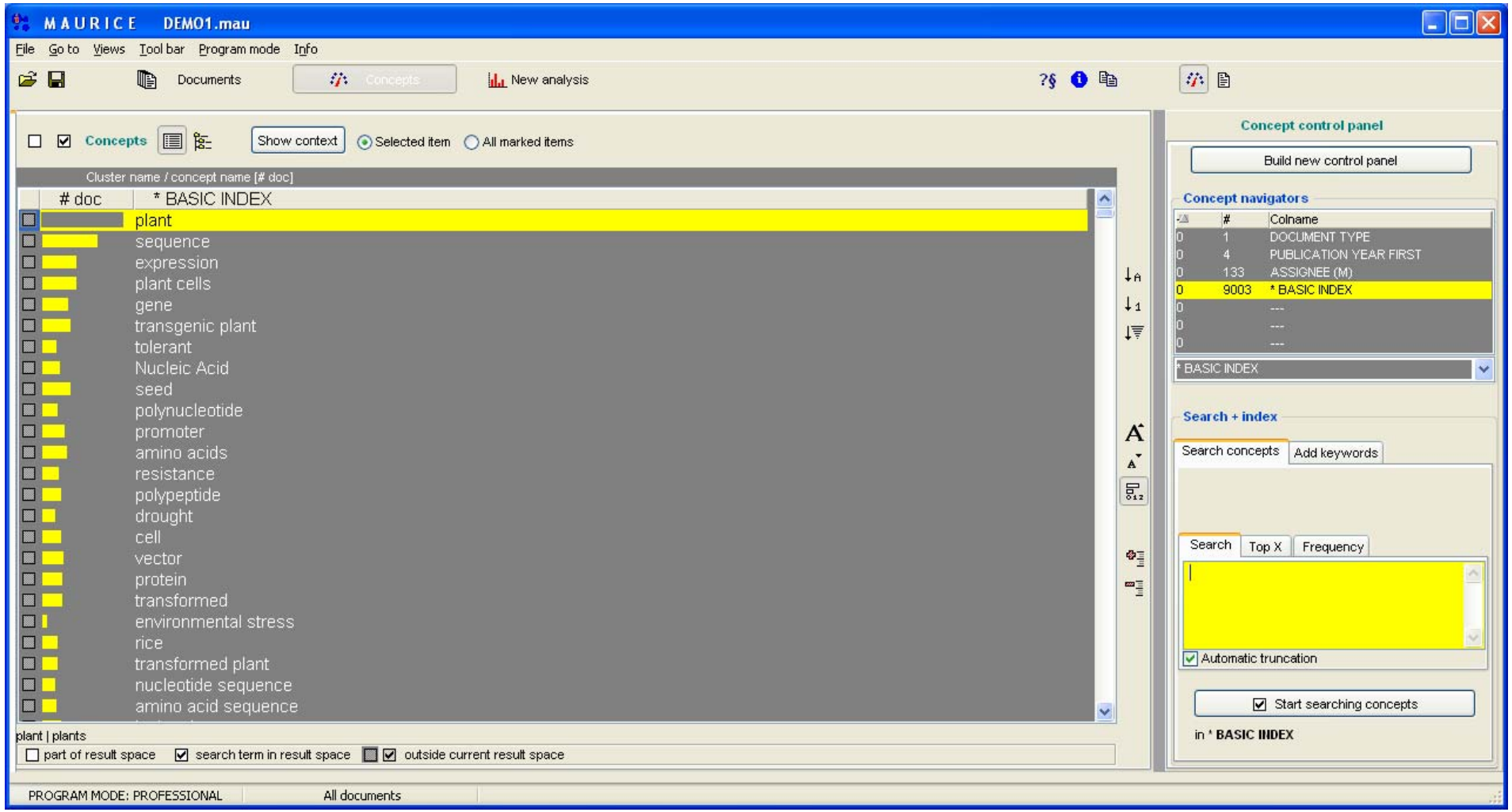


Context

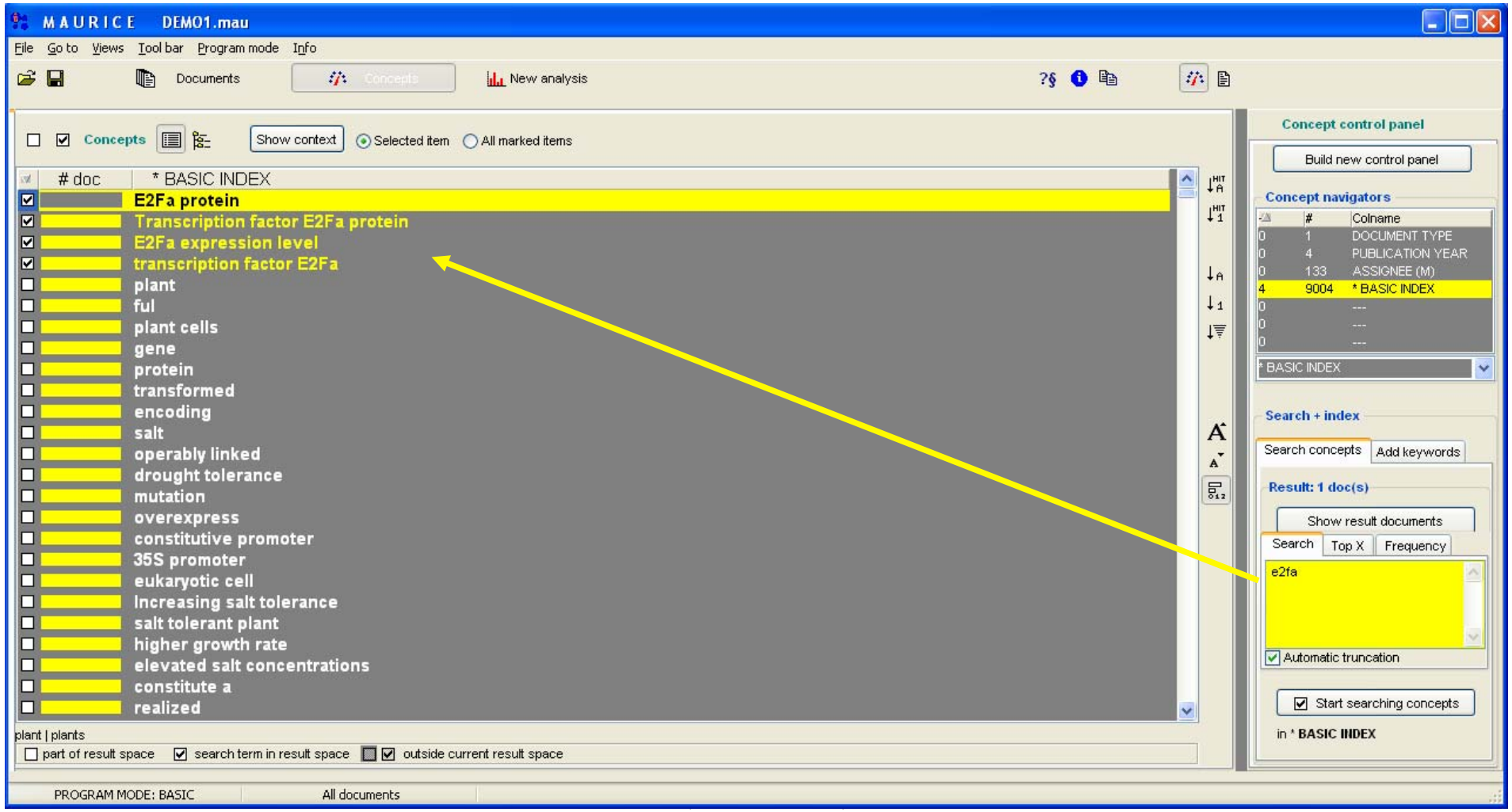
elevated salt concentrations (2)
eukaryotic cells (2)
higher growth rate (2)
increasing salt tolerance (2)
Plants (2)
salt and/or drought tolerance (2)

35S promoter
constitutive promoter
Mutation
operably linking the gene
overexpressing the transcription
Overexpression
plant cells
Protein
salt tolerant plant
strong heterologous promoter
whole plant

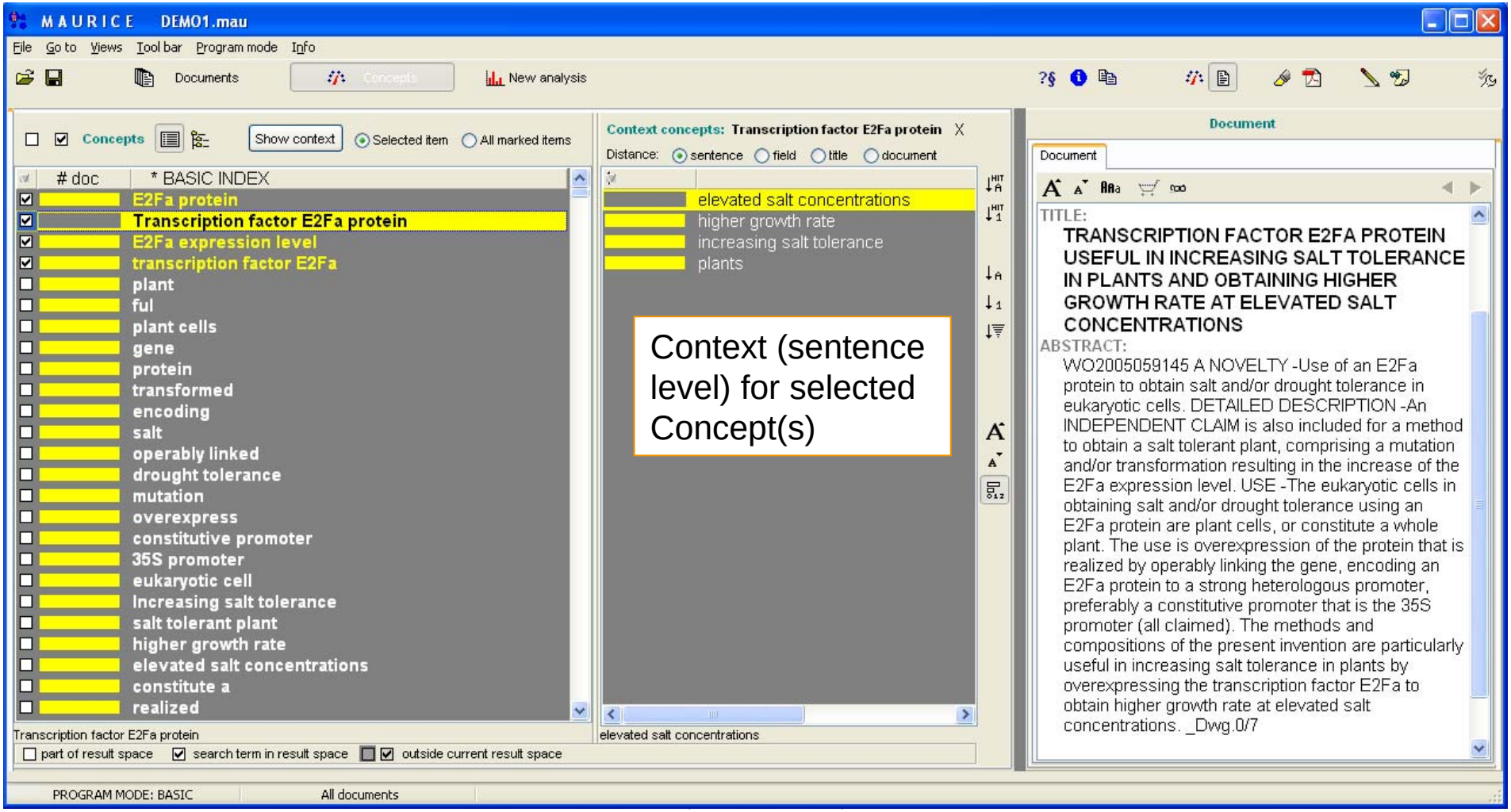
Result delivery:
Navigate through / search in concepts



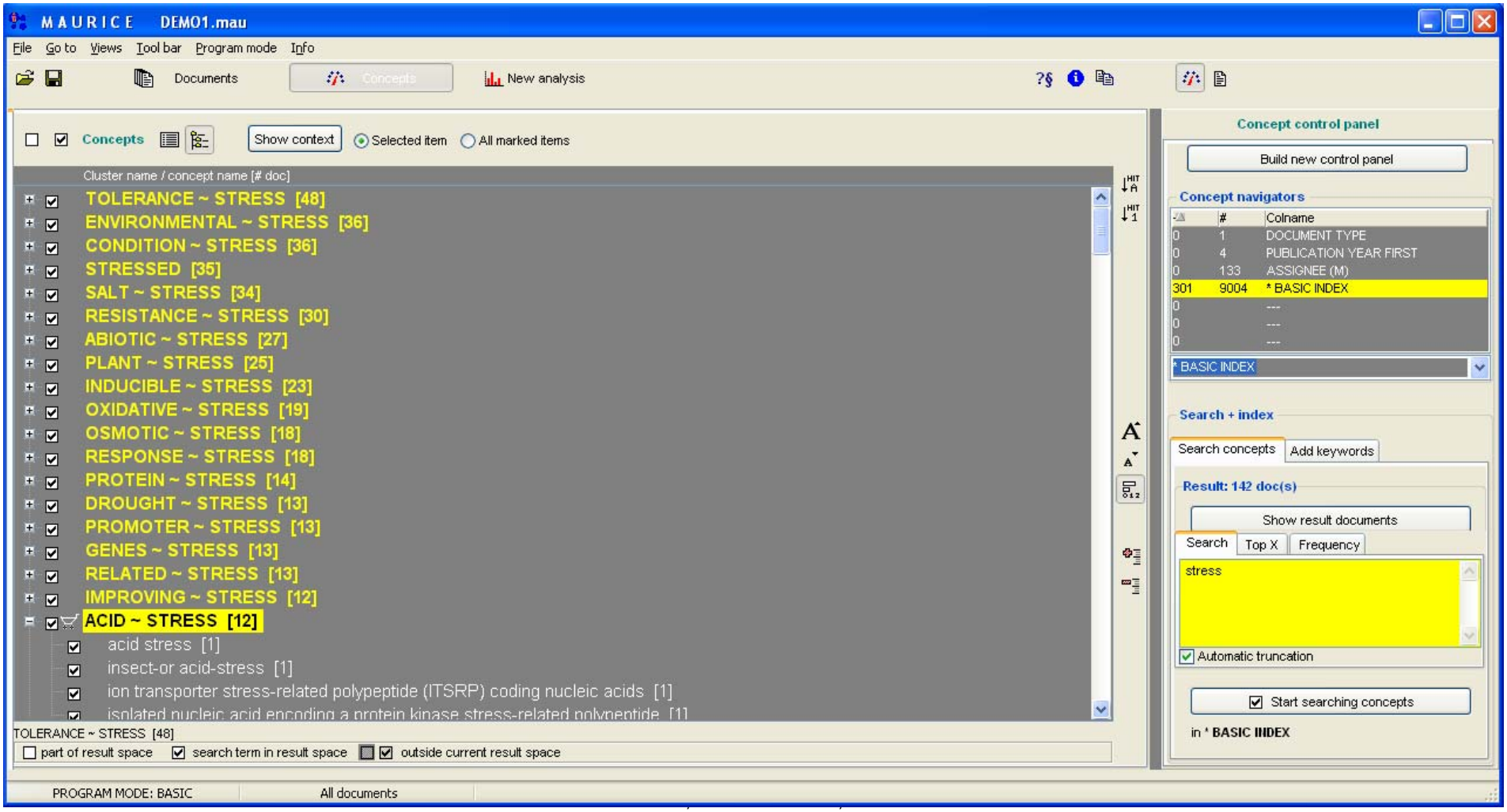
Result delivery:
Specific concepts found by searching



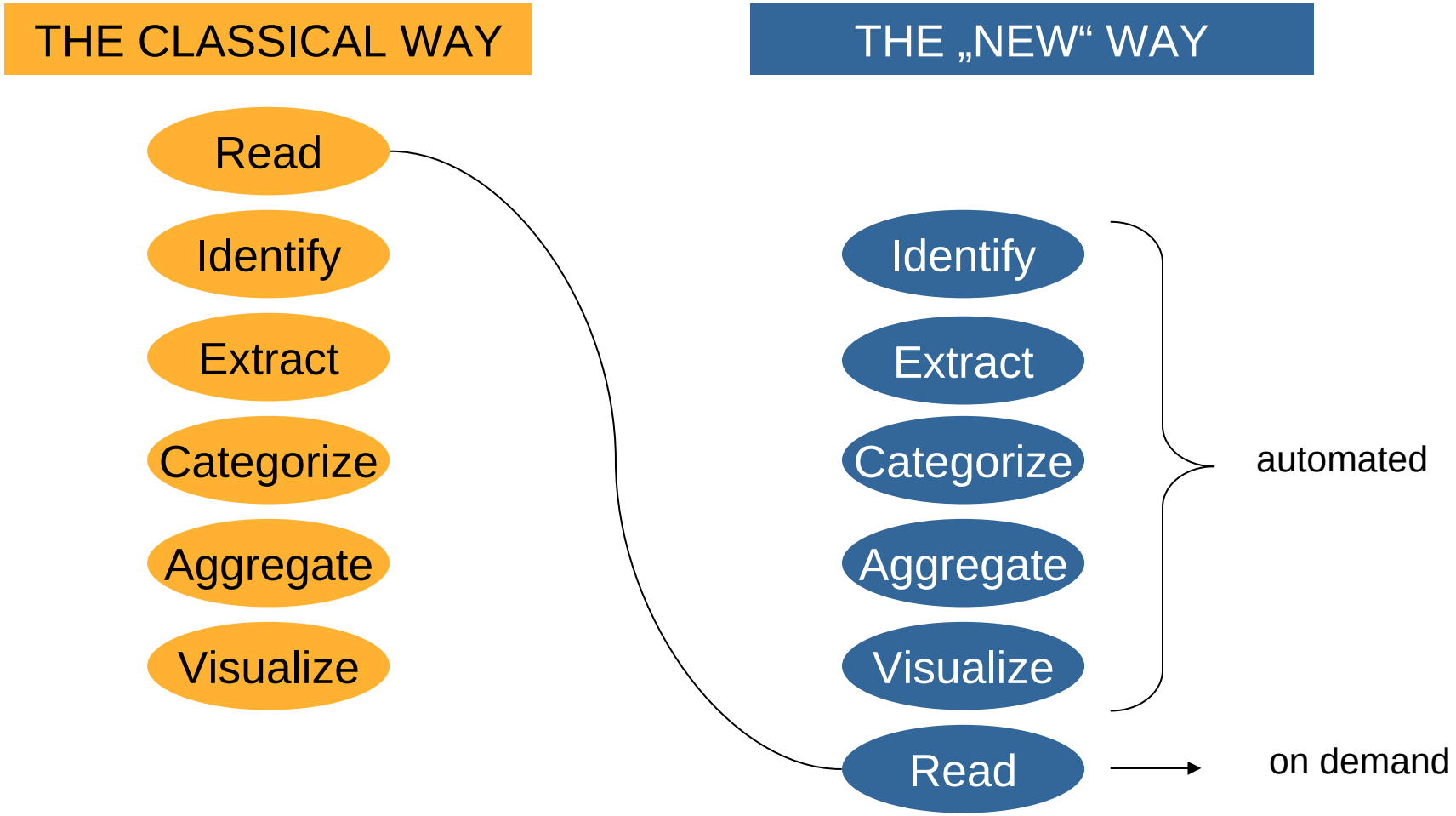
Result delivery:
Inspect context, drill down to documents



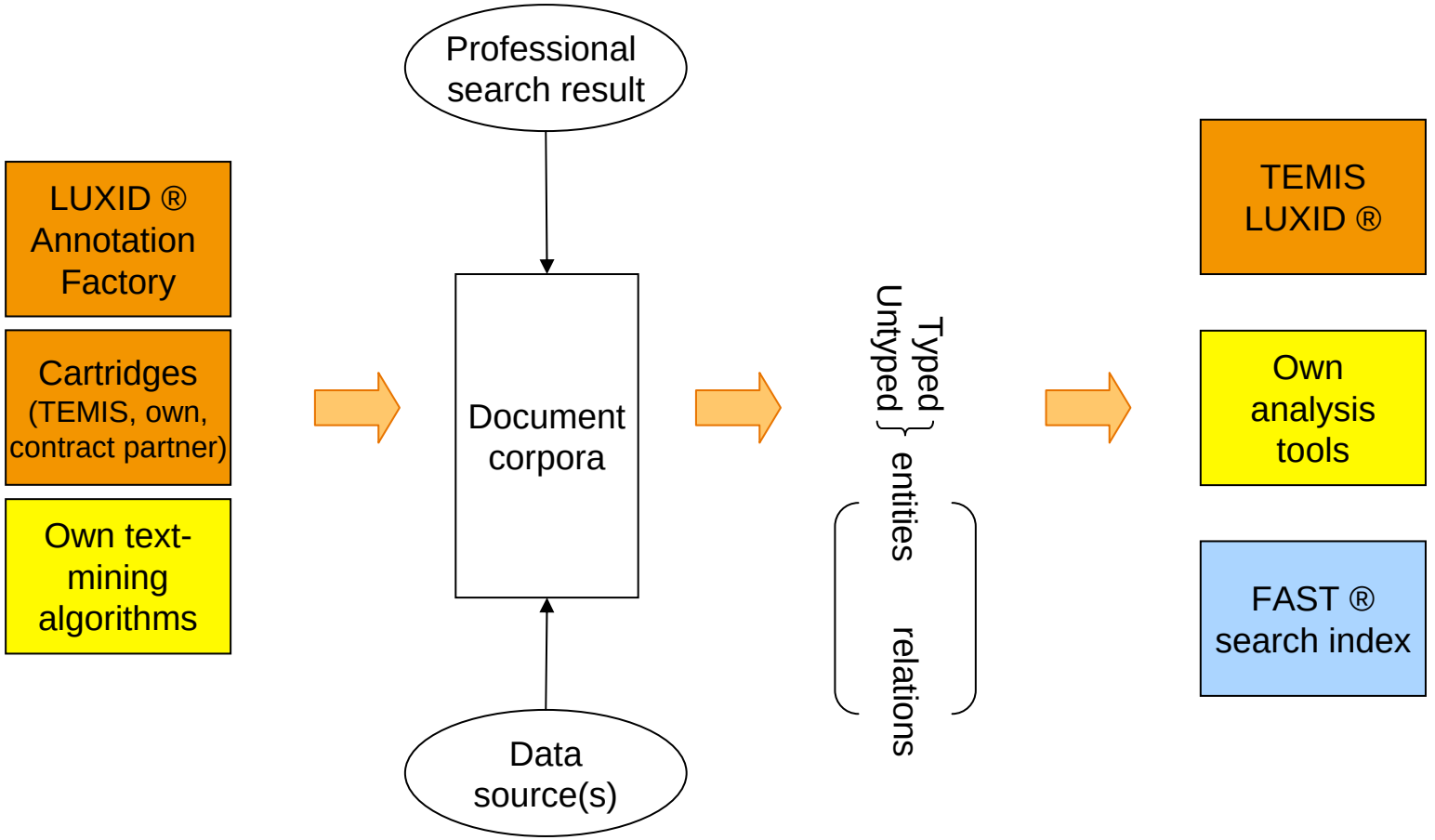
Result delivery:
Cluster concepts - if too many



Paradigm shift:
from documents to concepts



High level view on semantic architecture at BASF Info Center



Semantic technologies
Learnings and expectations

- Semantic technologies complement search & analysis
- Semantic technologies can help to
  - automate existing processes
  - set up new processes where classical methods do not work
- LuxidBar is an innovative idea to demonstrate the value of ontologies
- To build up the expertise to understand and utilize semantic technologies is a significant investment
- The development of own ontologies (as they are rarely available in chemistry) is manpower-intensive
- Process support to build ontologies or to cluster/categorize can be improved
- Analysis capabilities are far away from real life challenges
- Semantic technologies still need to be adapted to specific domains

Thank you very much for your attention.

Any questions?