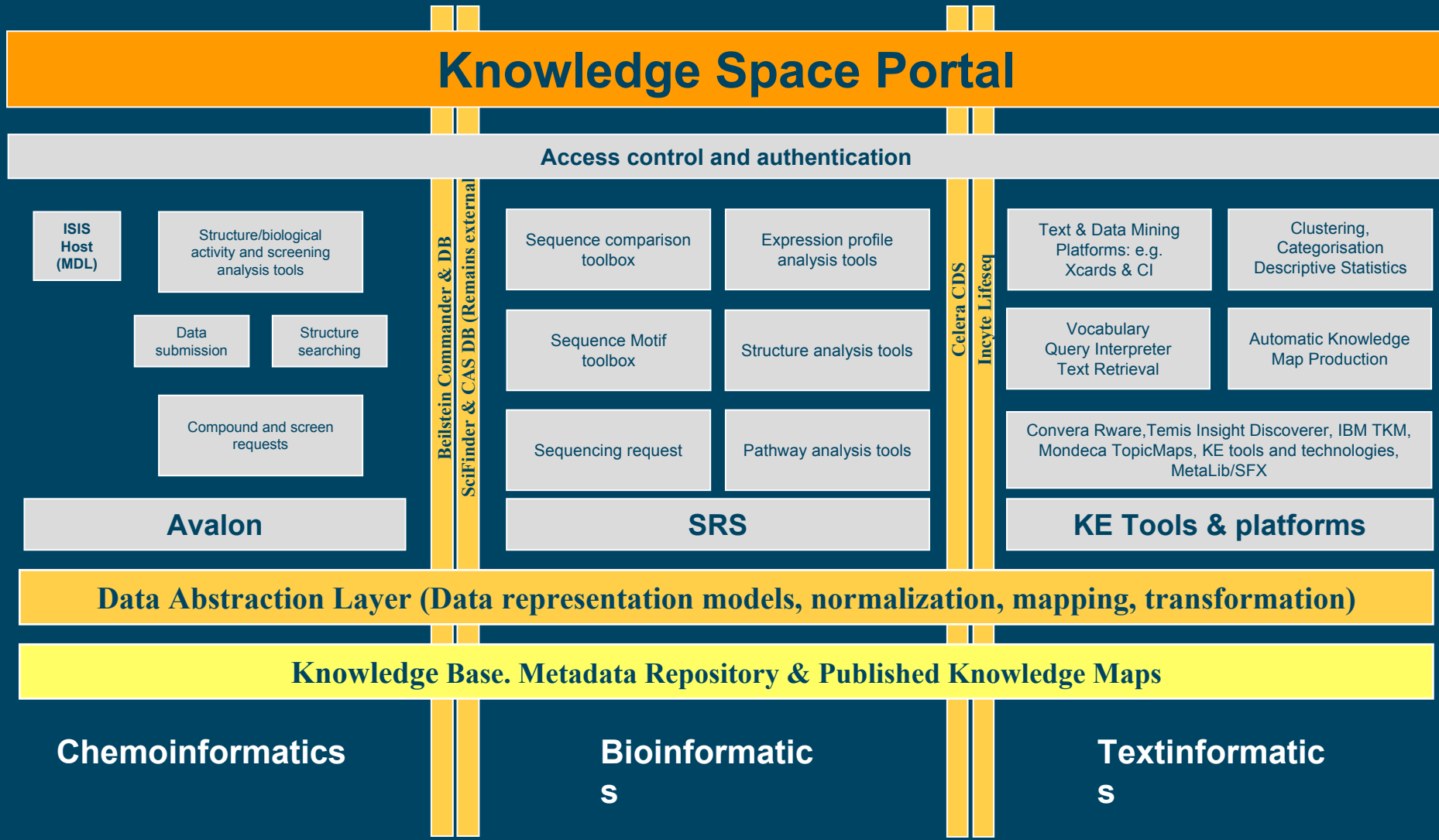
Text Mining in Life Science Informatics

Thérèse Vachon
Global Head of Novartis Information & Knowledge Engineering
IK@N, Informatics & Knowledge Management
Novartis Institutes for Biomedical Research

Basel Computational Biology Conference 2003

IK@N Knowledge Space Architecture



NOVARTIS

The implementation of the KS requires

- A knowledge representation model underlying the Knowledge Space
- A model for implementing the Novartis common terminology, for the validation and mapping of existing internal and external resources, and for the automatic production of consistent computational lexicons
- Advanced mining, information extraction and exploration techniques
- Advanced methods and tools for information searching and retrieval
- Advanced tools and components to be integrated on the Knowledge Space Portal

Knowledge Space Portal - Vision

The "Knowledge Space Portal" will, via a single customizable interface

- **Federate heterogeneous data resources and provide precise organization of the content**
- **Provide quick and intuitive access to information**
- **Provide data extraction, analysis and exploration tools**
- **Allow data integration, data exchange and interoperability of applications**
- **Provide mechanisms for data capture and annotation**
- **Provide knowledge sharing and collaborative tools**

Purpose of Text Mining

Text is by far the most important source of information

It remains largely untapped

- Unstructured
- Metaphoric
- Ambiguous
- Redundant
- Requires *a priori* knowledge of content
- Allows different viewpoints and different readings

Purpose of text mining

- *Ad-hoc* extraction of relevant information from structured or unstructured text
  - Relevant concepts, ideas, relationships between concepts
- Normalization of data representations
- Filtering
- Categorization

Tools and Technologies

Tools and Technologies for Text Mining

Development and integration of advanced text mining, information extraction and exploration techniques

- Lexical extraction, tagging & hyperlinking
- Natural language processing, information extraction
- Descriptive statistics and clustering, categorization

Business benefits

- Identification and extraction of meaningful objects and relationships between objects from text
- Consistent, business-relevant terminology across data sources
- Knowledge inference mechanism
- Discovery of unexpected data relationships
- Automatic tagging and hyperlinking across sources and disciplines (compound codes, citations, authors, accession codes, etc.)
- Detection of novel patterns rather than predefined patterns in specific classes
- Improved navigation across data sources and document sets

Knowledge Representations

- Develop flexible data representation models and tools for handling vocabularies, taxonomies, ontologies, etc.
- Design a robust and stable scheme for metadata and a common terminology (thesauri, ontologies etc) for describing objects in the KS
- Design and implement a dynamic conceptual network linking objects in the KS (Knowledge Map)

Business benefit

- Common representation scheme for describing data resources and associations between data elements
- Bridges between databases belonging to different disciplines
- Data analysis, categorization, navigation and exploration across data sources
- Smooth data integration and data exchange among applications
- Comprehensive, easy, and rapid access to all relevant data in the Knowledge Space
- Intuitive and dynamic navigation

Structured controlled vocabularies

Provide structured controlled vocabularies and vocabulary stores, used for validation, indexing, retrieval, navigation, data analysis, interactive data reduction and exploration tools

Business benefit

- Consistent search, retrieval, and analysis across databases
- Validation of metadata entries
- Increased data consistency
- Data exchange and interoperability

Text mining and exploratory statistics

Analysis and exploration of large document sets

- Unified view of heterogeneous sources
- Analysis of trends and patterns
- Analysis of complex relationships between data elements
- Detection of deviant or emerging information
- Knowledge inference, serendipity

Data reduction and exploration methods

- Common representation scheme across heterogeneous data sources
- Lexical extraction, information extraction
- Unbiaised analysis methods
- Intuitive data exploration and navigation tools
- Consistent graphical representations
- Link to underlying data

Data set construction

- Data acquisition
- Parsing
- Lexical extraction
- Information extraction
- Terminology mapping
- Normalization
- Variable typing
- Categorization

→ Homogeneous formal representation of heterogeneous data sources

NOVARTIS

Descriptive Statistics

Methods

- univariate (statistical properties of a single variable)
- bivariate (link between two variables)
- trivariate (effect of a third variable on that link)
- n-variate (effects of a third variable on multiple sets of two variables)
- multivariate (relationships between all variables (or modalities) in a data set)
  - relational analysis
  - K-means clustering
  - single and double hierarchical clustering
  - correspondence analysis
  - multidimensional scaling

On several types of native (contingency) or derived tables

NOVARTIS

Interactive graphical exploration

- Bar charts
- Bubble charts
- X-y plots
- Factorial maps
- Dendrograms
- Heat maps
- etc.

- Base lines
- Filtering
- Drill-down
- Expansion
- Localization

Links to data resources underlying the graphs

Problems with textual data sources

- **Analysis cannot be conducted on tables derived directly from ‘documentary data’, either full text or secondary sources**

  - drastic degradation content
  - lack of reactivity to new concepts
  - discipline-orientation
  - obsolescence of indexing schemes
  - heterogeneous representations
  - distribution of words / long tails / loss of information
  - overlaps of meaning / non-homogeneous variables
  - tables are not mathematically valid for most methods (void tables / ‘no response’)
  - results are trivial, unstable, or meaningless

Lexical *vs* Information Extraction

Lexical extraction

Extraction of meaningful concepts from text (or other data sources). Mainly based on the use of dictionaries

Information extraction

Extraction of objects and relationships between concepts (associations), in a goal-oriented manner. Mainly based on syntactic analysis (global / local) supplemented by dictionaries

NOVARTIS

Lexical extraction

Identification of objects in text:

- Morphological rules, separators, etc.
- Identification of idioms (meaningful noun phrases)
- Multiple ***(embedded or overlapping)*** identification
- Dictionary selection

Followed optionally by:

- Normalization
- Assignment of classes
- Keyword indexing

NOVARTIS

Usual problems

- Contextual identification (disease : Indication vs SE)
- Ambiguous acronyms
  - EGFR [1] = epidermal growth factor receptor
  - EGFR [2] = estimated glomerular filtration rate
- Homographs, Polysemy
  - Vistagen = drug (levobunolol)
  - VistaGen = company
- Objects not identified by names (*e.g.* anaphoric reference by pronouns)
- Extraction of concepts / not of associations between concepts
  - different from information extraction

NOVARTIS

Exploratory analysis

Robust analysis can only be carried out on tables prepared from generic variables (classes, categories)

- **Mathematically valid tables**
- **Retention of specific detailed information**
- **Drill-down and iterative analysis**
- **Links to underlying documents**

Extracted Objects

- *Terms*: lexical item which triggers a concept
- *Concepts*: what is actually extracted, attached to a hierarchical structure and synonym groups (terms)
- *Types*: simple hierarchical structure attached to concepts

Filtering based on Types can be combined with document structure filtering.

NOVARTIS

Applications

- Search & retrieval
- Extraction
- Categorization
- Information analysis
- Information exploration
- Navigation
- Data integration & data exchange

Applications currently being developed

- Ulix
- Knowledge Map
- Generic Text Analysis Platform
  - Applied to Competitive Intelligence
  - Applied to Genomics
  - Applied to NewsFlow
  - …
- Knowledge Space Portal

Ulix - Scope

- Consistent retrieval and analysis over 80 internal and external databases
- Lexical extraction
- Typed variables
- Hierarchical vocabulary
- Simple statistics and iterative K-Means clustering
- Filtering
- Links to underlying documents

ULIX Clustering

General graphical representation

Query: chronic pain
Libraries: /Medline & EMBASE (1988-Present)(376)
Clustering fields: Topics

Clusters	Stats	Chart	Documents
▶1. opiate ; m(; hyperalgesia	stats		48
▶2. drug thera dministration ; adverse drug reaction ; nonsteroid antiinflammato ; side effect ; magnesium ;	stats		44
▶3. fibromyalg eck pain ; whiplash injury ; r	stats		41
▶4. therapy ; cl ure ; review ; management ; psychophysiolc	stats		34
▶5. disability ; prevalence ; epidemiology ; patient ; distress syndrome ; world health organization	stats		32
▶6. anxiety ; complex regional pain syndrome ; spinal cord stimulation ; sympathetic dystrophy	stats		19
▶7. surgery ; abdominal pain ; knee ; laparoscopy ; ankle ; nuclear magnetic resonance imaging	stats		17
▶8. reliability ; questionnaire ; backache ; instrum on	stats		16
9. chronic pancreatitis ; tendinitis ; pharmaceuti	stats		6
10. vasectomy ; contraception ; prostate cancer	tats		3

Clusters 1-8 include sub-clusters [indicated by blue flag]

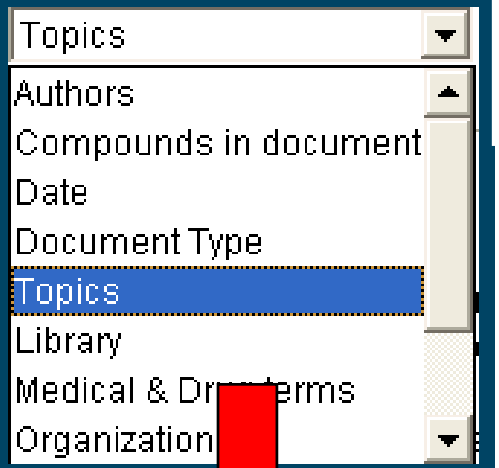
Graphical representation of sub-clusters

Access to a whole range of statistics

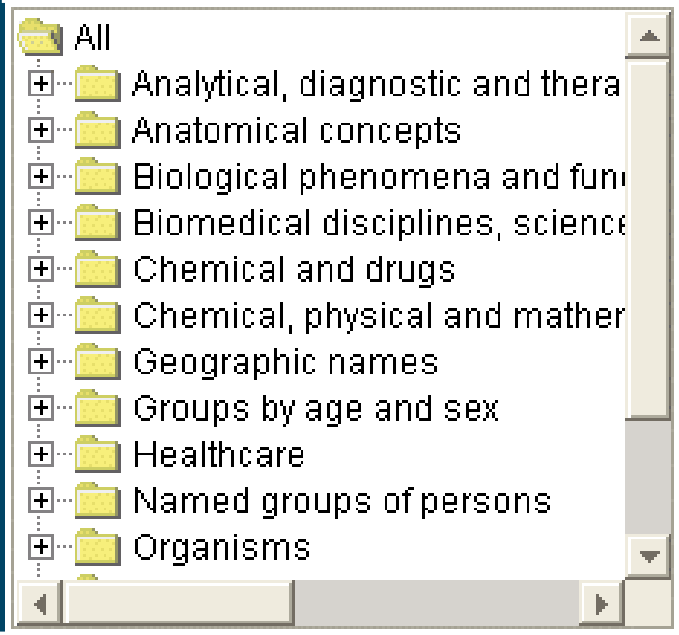
Access to documents

NOVARTIS

ULIX Clustering and Filtering

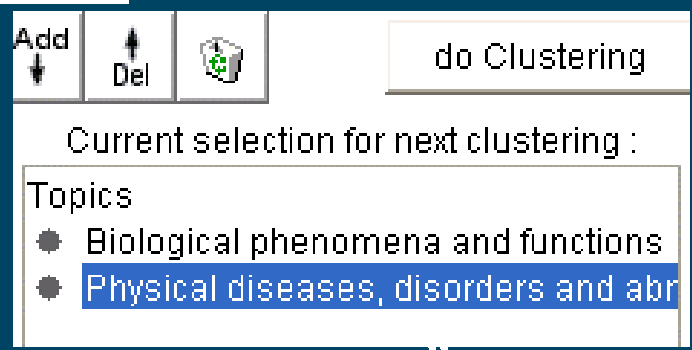


Select the main class for the clustering to be performed [here "topics"],...



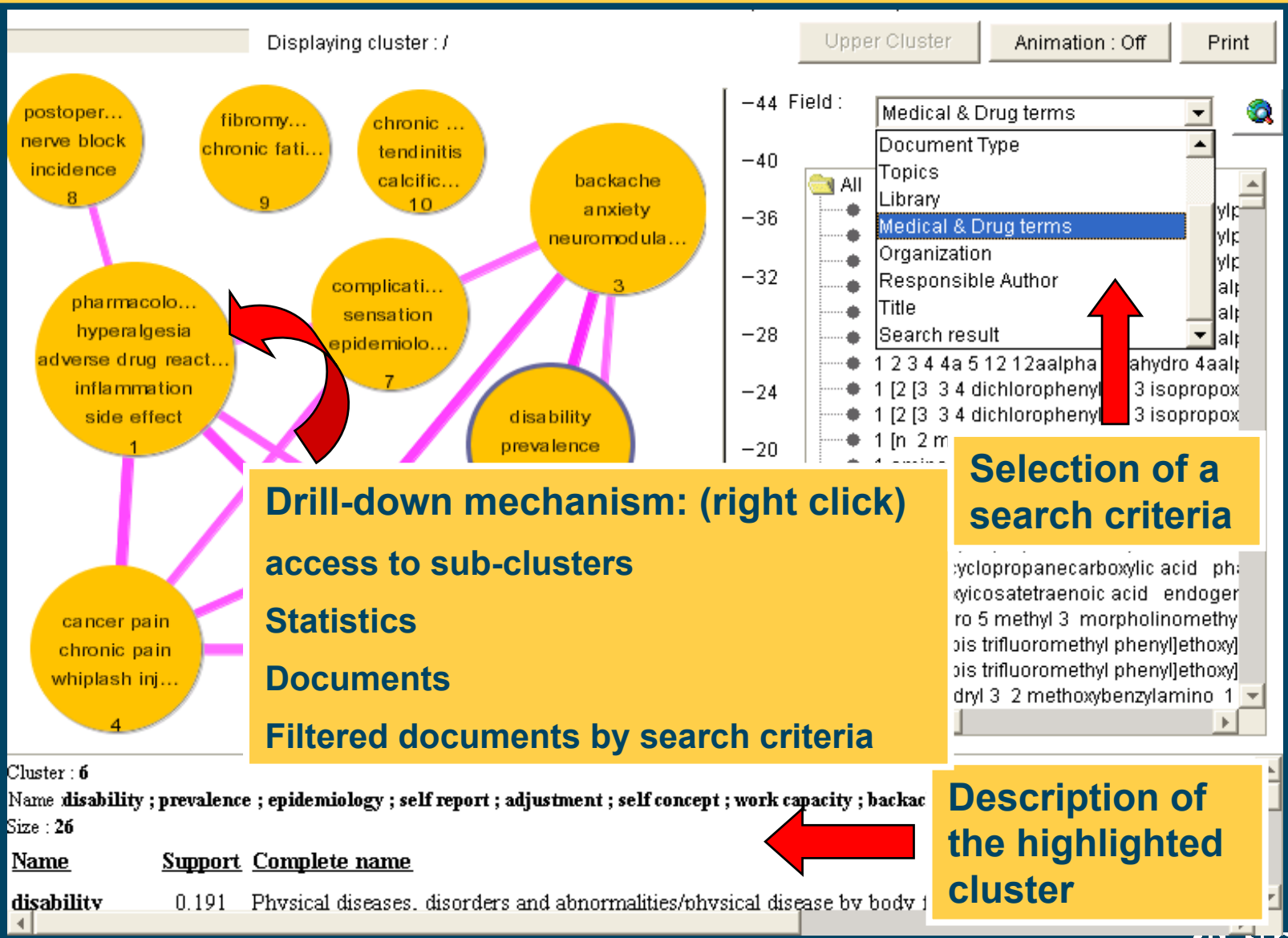
.....one or several sub-classes belonging or not to 2 different classes [here „Biological phenomena and functions", „Physical disorders and abnormalities" and Psychological and psychiatric phenomena]

Add your selection(s) to the current selection box and perform the clustering



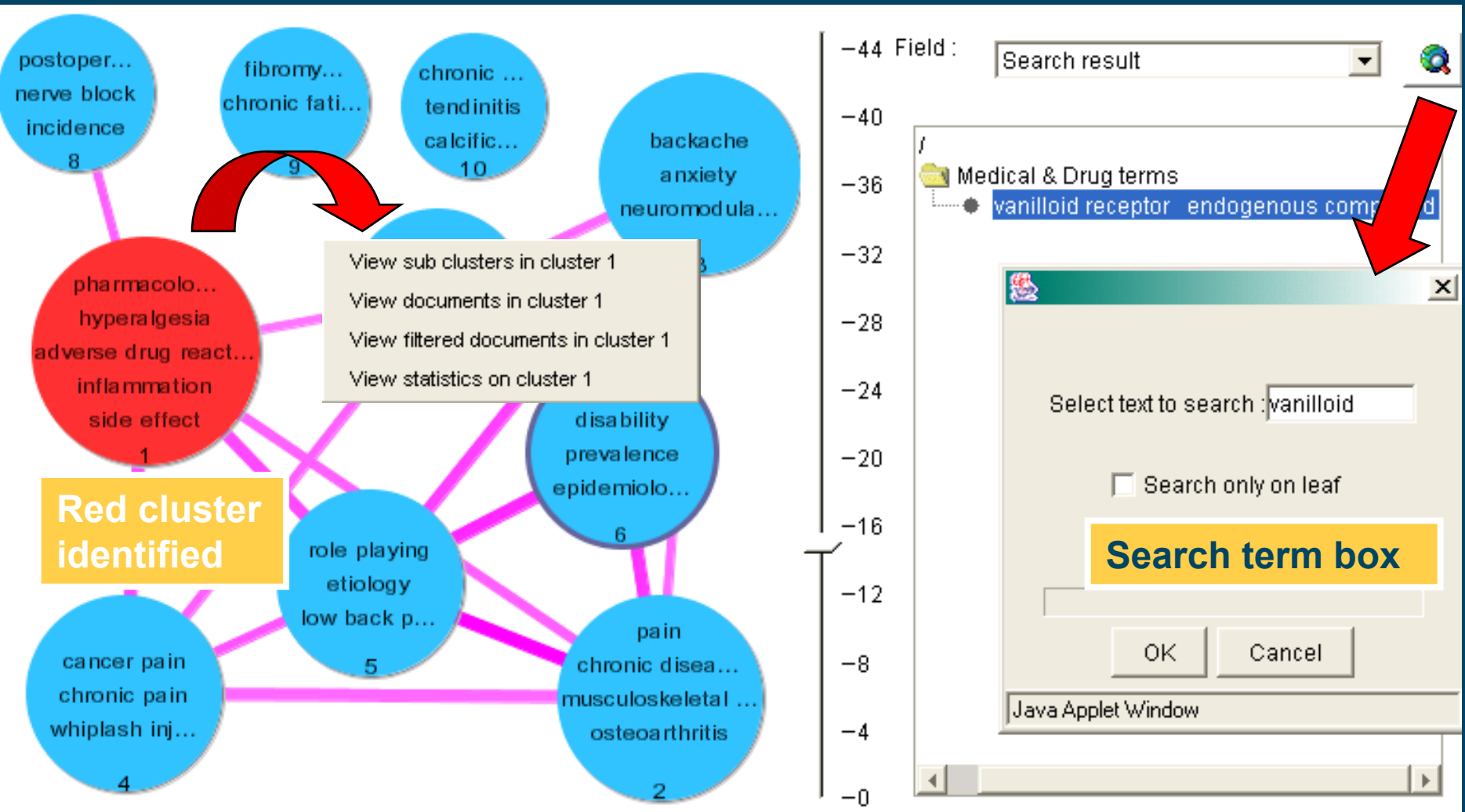
NOVARTIS

Ulix Clustering



NOVARTIS

Ulix Clustering and Filtering



NOVARTIS

Knowledge Map - Scope

- Tools for organizing retrieving, and navigating information resources
- Independent of the information resources themselves (knowledge layer)
- Node-link networks, where concept are nodes and associated relationships are links.
- Active, dynamic representations (hierarchies, networks, chains, etc.)

NOVARTIS

Metadata / Knowledge Map model

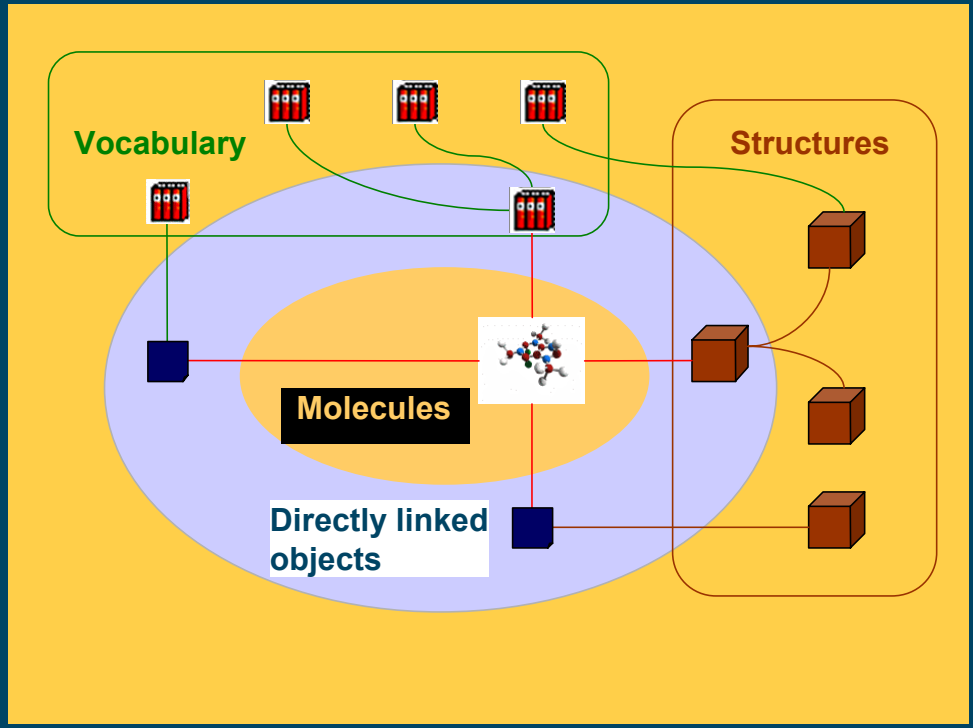
Molecule-centric model

- Organized and structured around the central concept of molecule and objects, attributes, parameters, properties, etc., attached directly or indirectly to those molecules
- Both types of objects are represented by topics, and the relationships between those objects by associations
- Together, they form the core Knowledge Base, further extended to two other classes of Topics
  - Vocabulary : terms from taxonomies, classifications, nomenclature, thesauri, etc.
  - Structures : real world individuals, structured objects and processes

Metadata/Knowledge Maps Model

Topic Classes

- Molecules
- Directly linked topics
- Structures
- Vocabulary

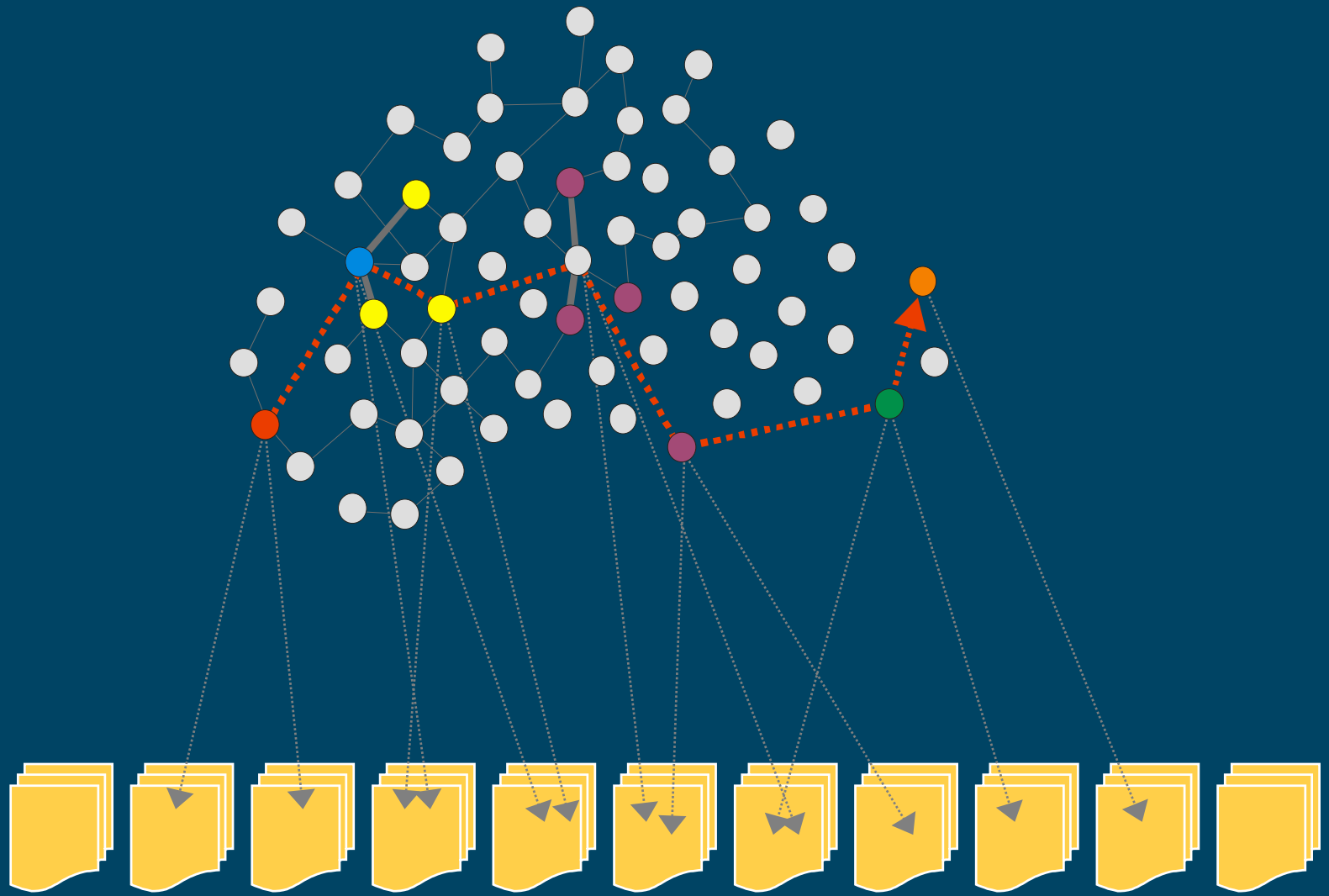


Association Types

- classified according to topics classes and subclassified as necessary by scopes
- Define the topic map "structural ontology".
- For each association type, the role types are defined

NOVARTIS

Navigation

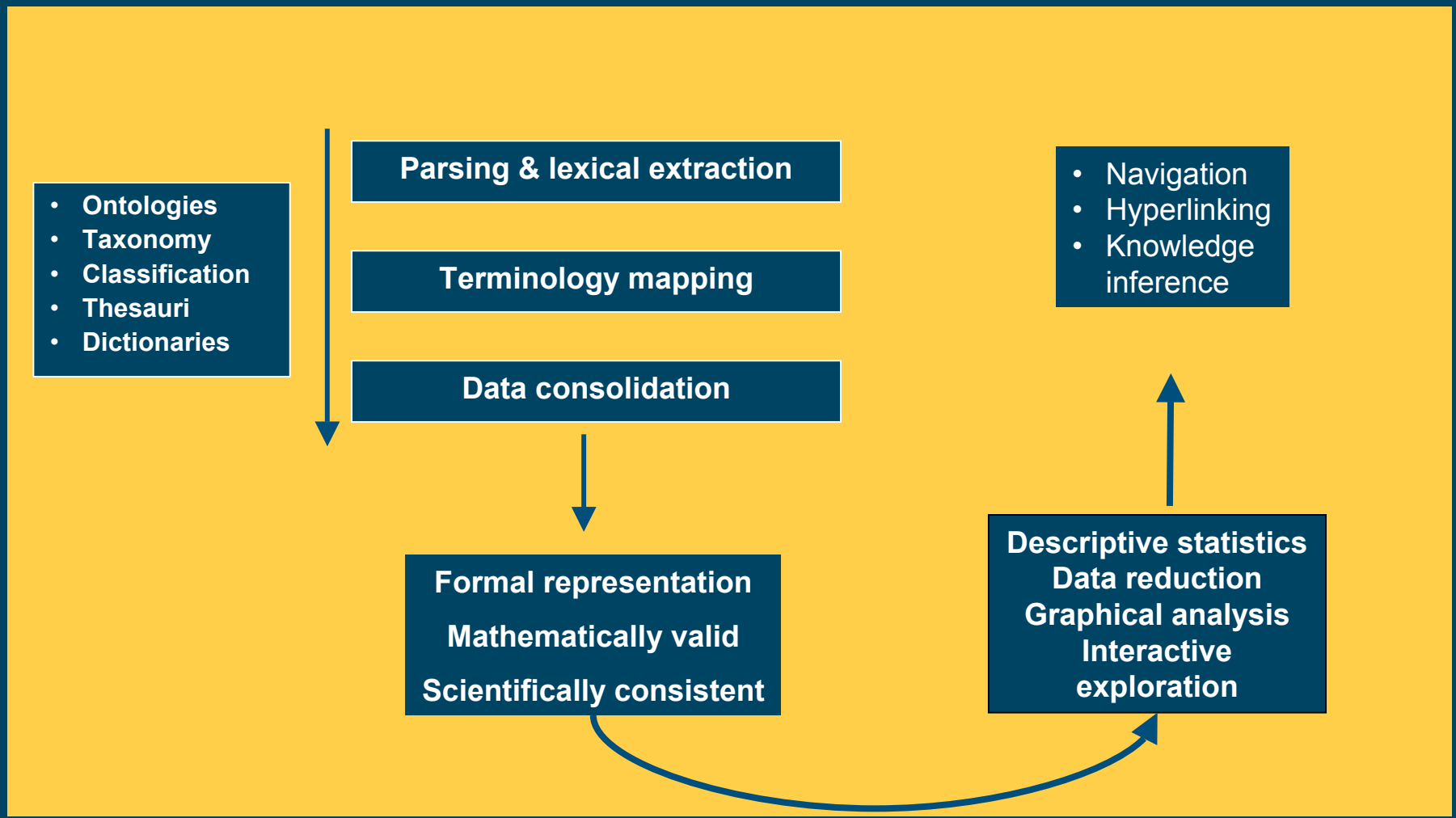


NOVARTIS

Topic Types

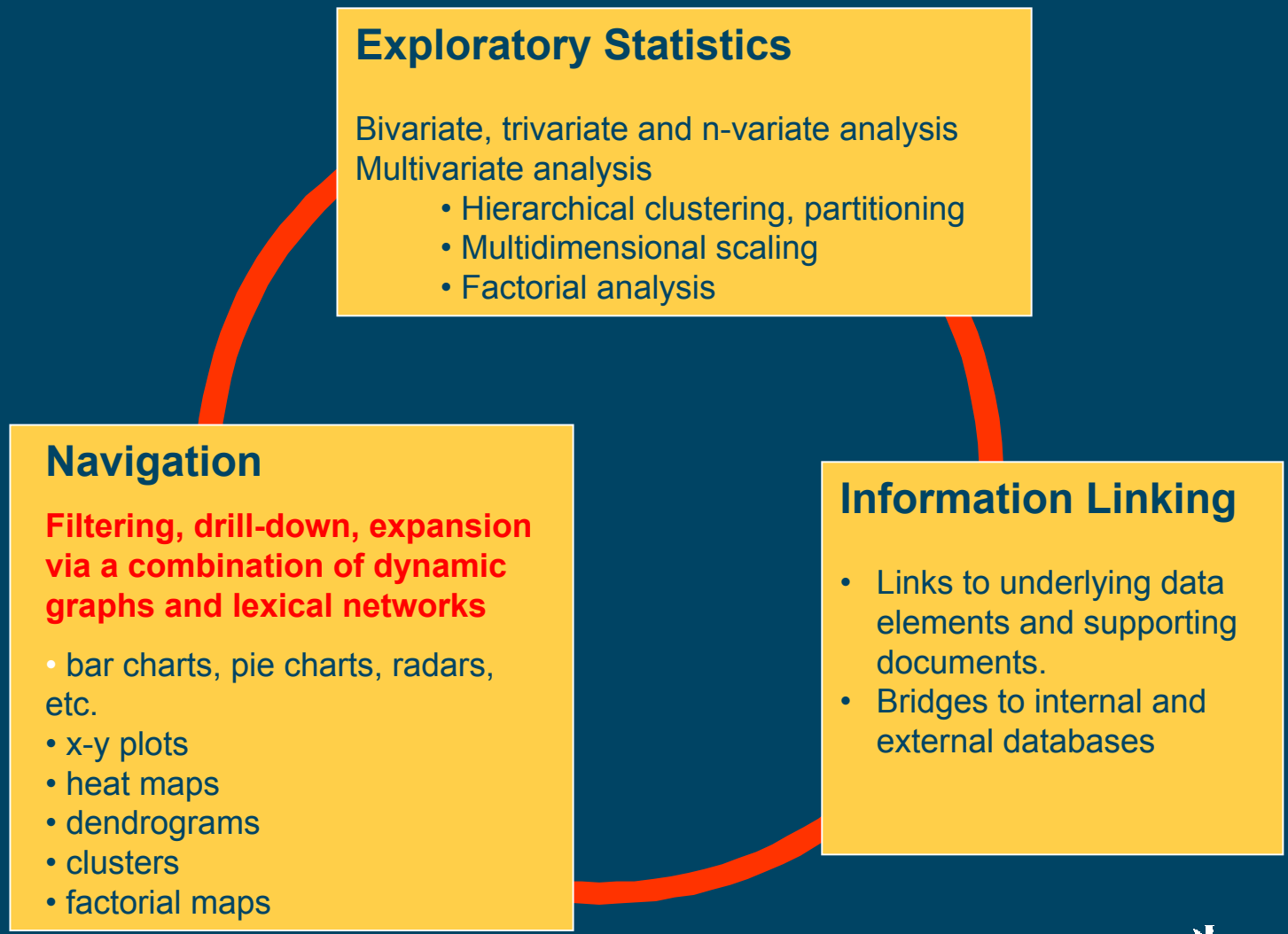
- Anatomy
- Assay
- Chemistry
- Date
- Development status
- External
- Galenics
- Diseases
- Molecular entities
- People
- Physiological Processes
- Organization
- Properties
- Targets

Generic Text Analysis Platform - Scope



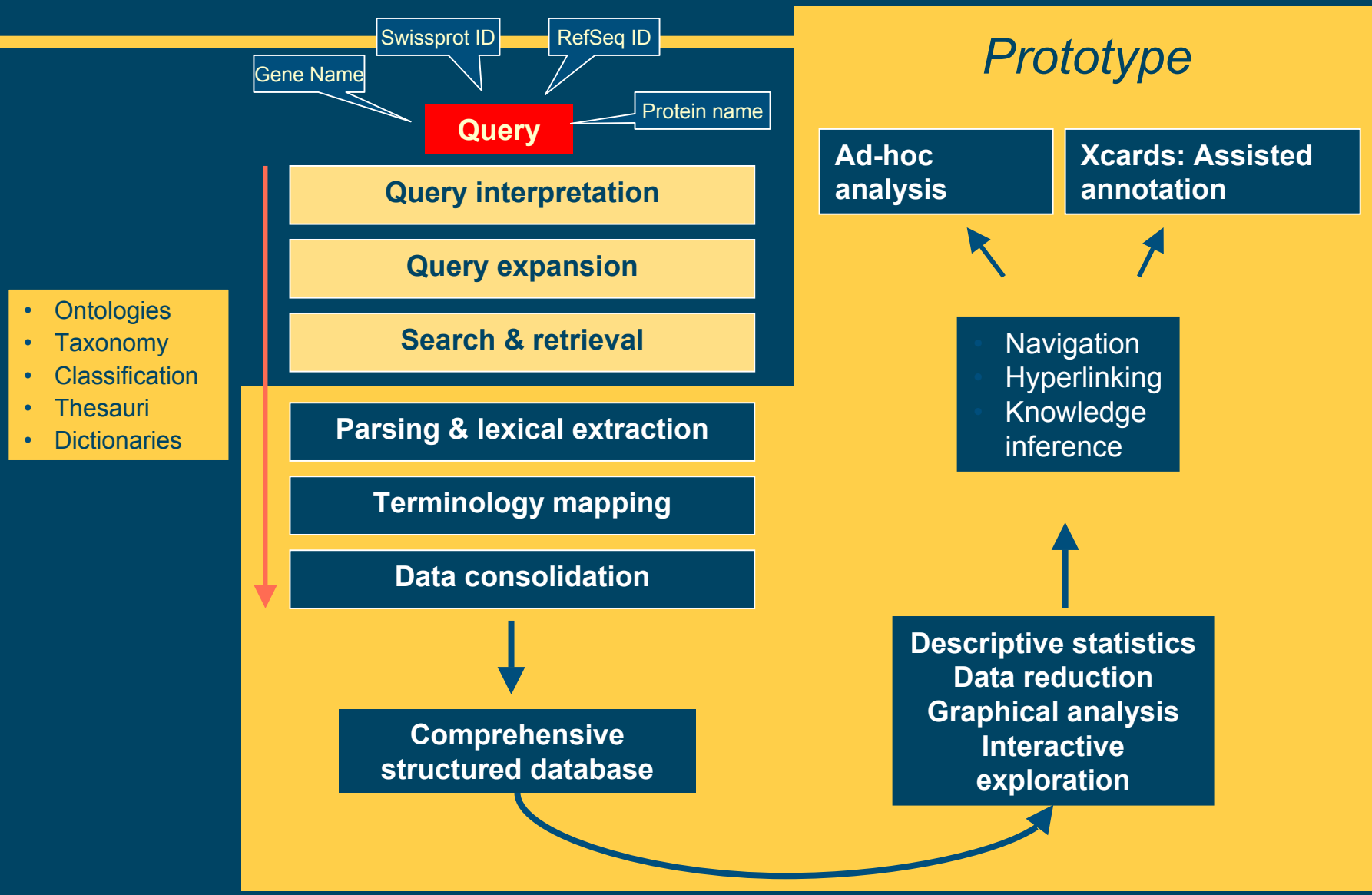
NOVARTIS

Interactive Exploration



NOVARTIS

Text Mining in Genomics - Prototype



NOVARTIS

Competitive Intelligence Analysis Platform

Consolidate all data essential for Competitive Intelligence (from multiple internal and external sources) into a single platform, together with interactive data analysis and exploration tools.

Consistent integration of data sources :

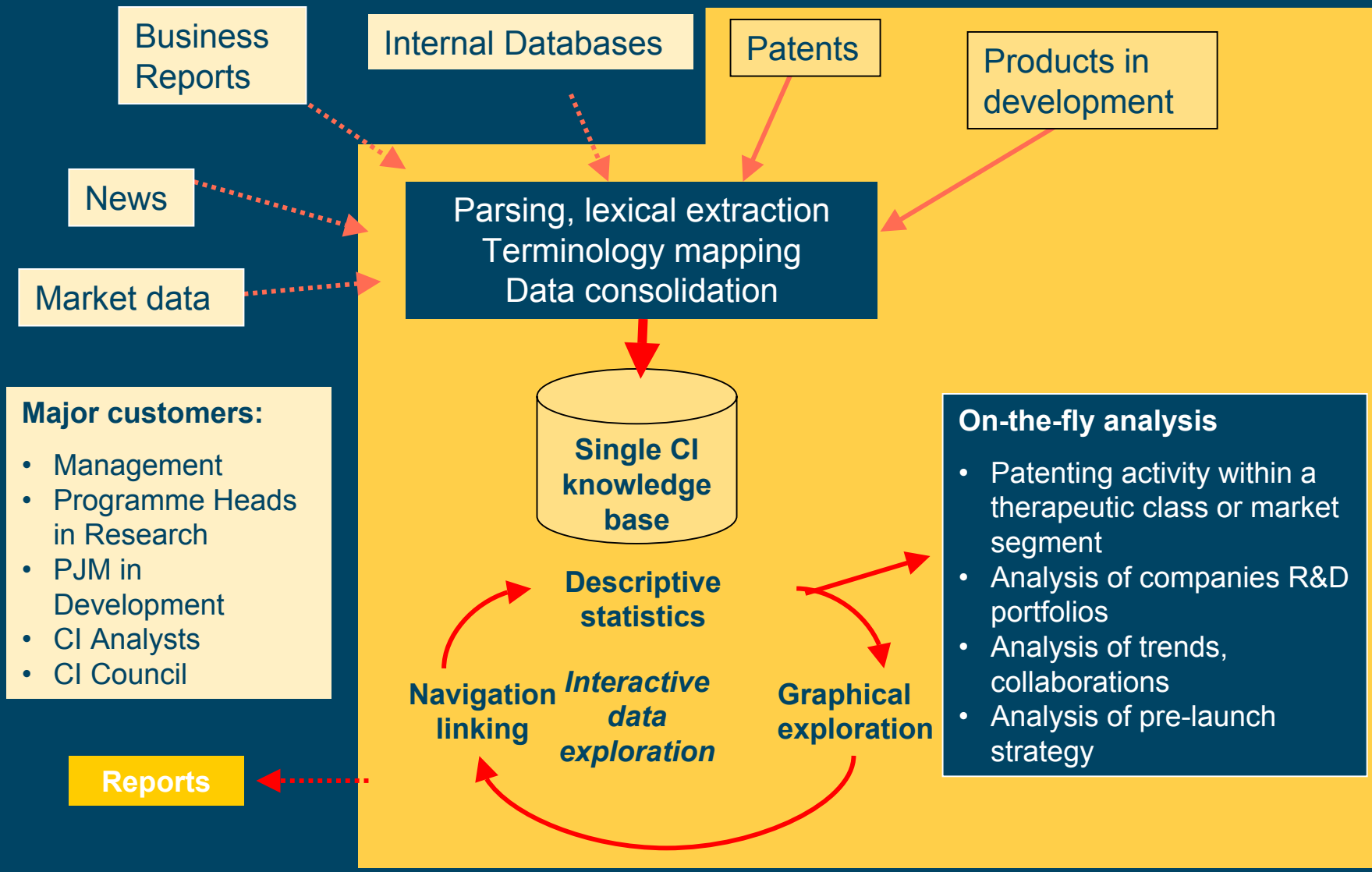
- Products in development
- Patents
- Internal CI sources
- Market data

(Mapped to a single representation scheme and taxonomy)

And

- Extensive data analysis, navigation, drill-down and reporting tools

Competitive Intelligence Analysis Platform



NOVARTIS

Competitive Intelligence Analysis Platform

- Comprehensive
  - Consolidating essential data from multiple internal and external sources into a single CI platform
- Consistent
  - Formats
  - Terminology
- Current
  - Daily updates
- Interactive analysis and data exploration tools

Examples of Analysis

- Patenting activity within a therapeutic class or market segment: type of protection, territorial coverage, build-up on original patents (process patents, formulations, *etc*.)
- Key inventors and teams
- Maturity / novelty of research projects
- Analysis of companies development portfolios: therapeutic classes, putative *vs* actual therapeutic indication, pharmacological classes, market segments, development phases, ranking, backups, speed of development, overlap of portfolios, pioneering research and me-too products
- Analysis of trends (over time and/or development phases)
- Collaborations (joint filings, product licenses, co-marketing)
- Analysis of pre-launch strategy

Data Sources

- Patent applications (primary and secondary sources)
- Products in development (commercial and internal sources)
- *Conference reports (internal and external)*
- *Published literature*
- *Market data*
- *Epidemiological data (prevalence, incidence)*
- *Business analysis reports*
- *Internal CI sources (internal analysis reports, annotations, validated 'human intelligence')*
- *Web crawling results, etc.*

Patents

- Research described in patents is approximately 2-yrs old
- Widely varying filing practices (broad *vs* specific applications, filing routes, territorial coverage, etc.)
- Lack of precision in some areas (e.g., potential therapeutic activities)
- Poor description of content

- The analysis of patent portfolios can give a reasonably accurate idea of the volume of activity in research, trends with time, *etc.*
- Not directly predictive of future clinical development activities.
- More sophisticated models must be applied to gain a clearer understanding of a company R&D strategy
- Also, a wealth of related information (collaborations, location of research, key inventors, *etc.*)

Patents remain the major source of information on R&D activity

NOVARTIS

News Flow Analysis Platform

Live news feed pulled every minute from News Edge

Lexical extraction to identify:

- Companies
- Products
- Diseases
- Company events (M&A, licences and agreements, product approvals)

Personalized categorization (*e.g.*, top 10, BUs, disease area, *etc.*)

Live display of customized news flow (filtered)

Links to reference data (company profiles, product profiles, *etc.*)

Link to the portfolio analysis platform

Automatic processing and mining of a NewsFlow

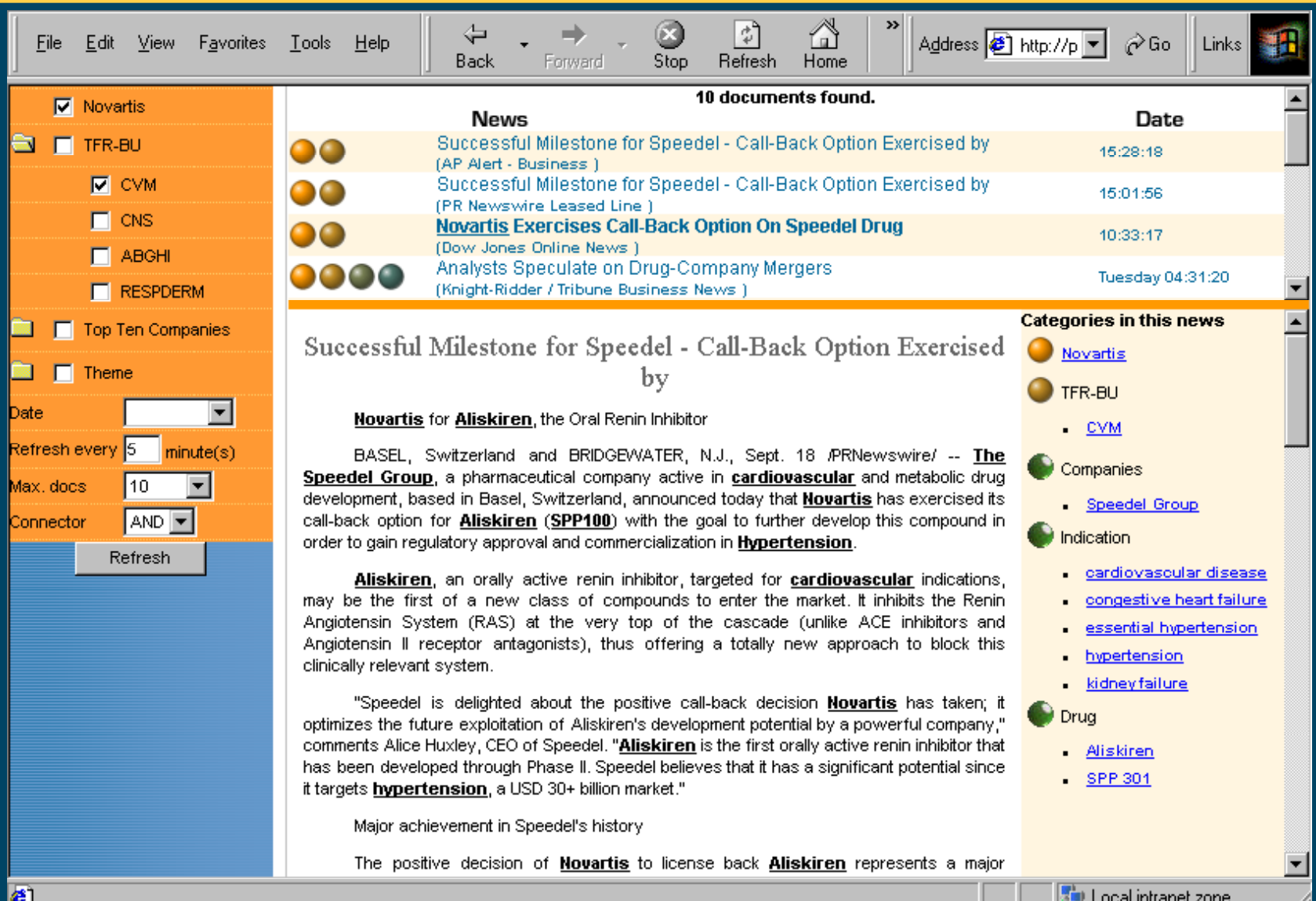
News are pulled out of NewsEdge's server every minute

Entities which are recognised and processed automatically by the lexical extractors currently include :

- Full list of drugs, launched or in development, with synonyms and brand names, normalized to the INNs
- Subset of ~2000 major indications, with synomyms and narrower terms, consolidated and mapped to the dictionary of indications used by the CI analysis platform
- List of companies with their affiliates in different countries, automatically extracted from CI sources (products & patents) and constantly updated.

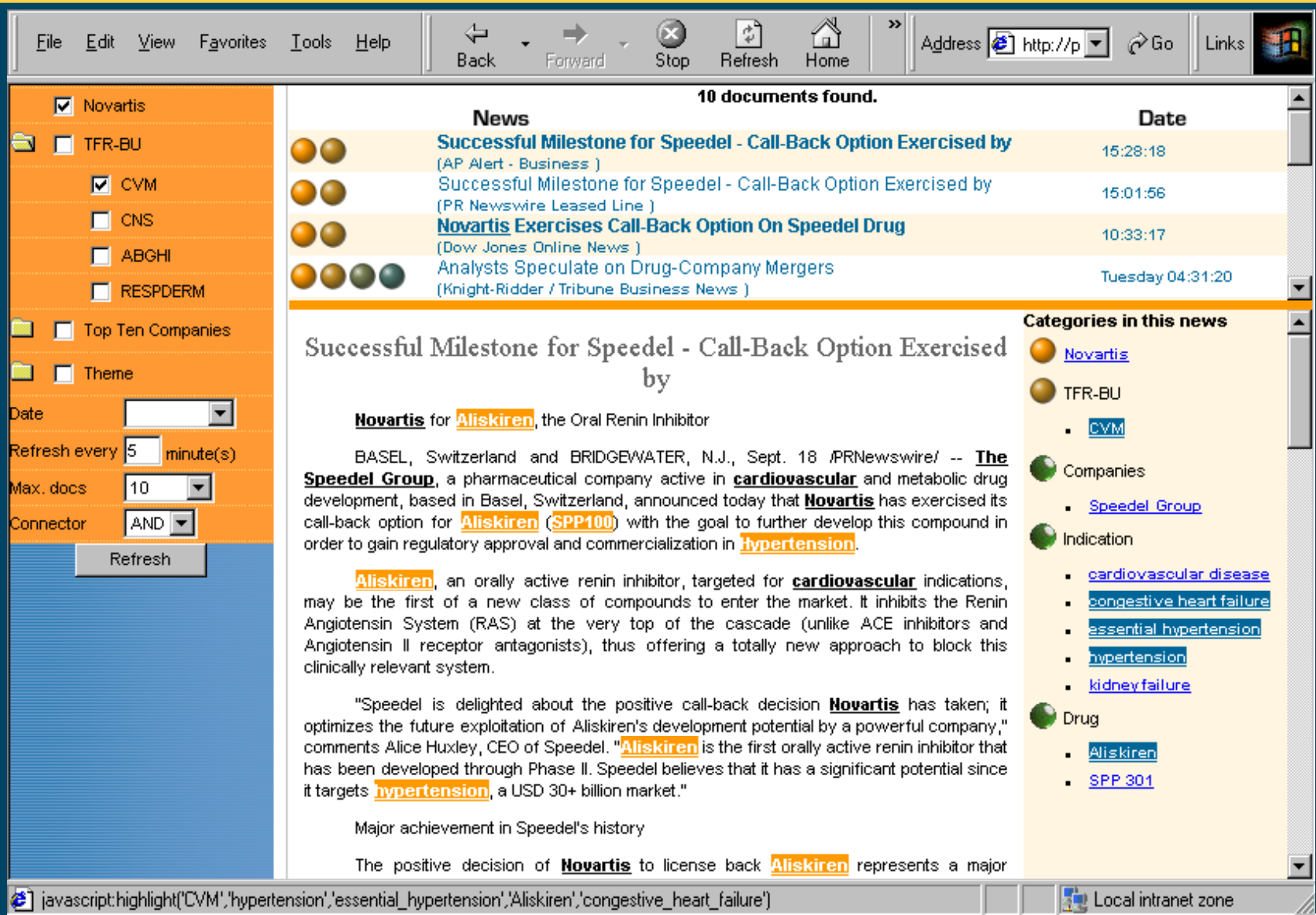
Information extraction prototype : Mergers and acquisitions, product approvals, licences are identified, marked and extracted

Annotation with lexical extraction and categorization



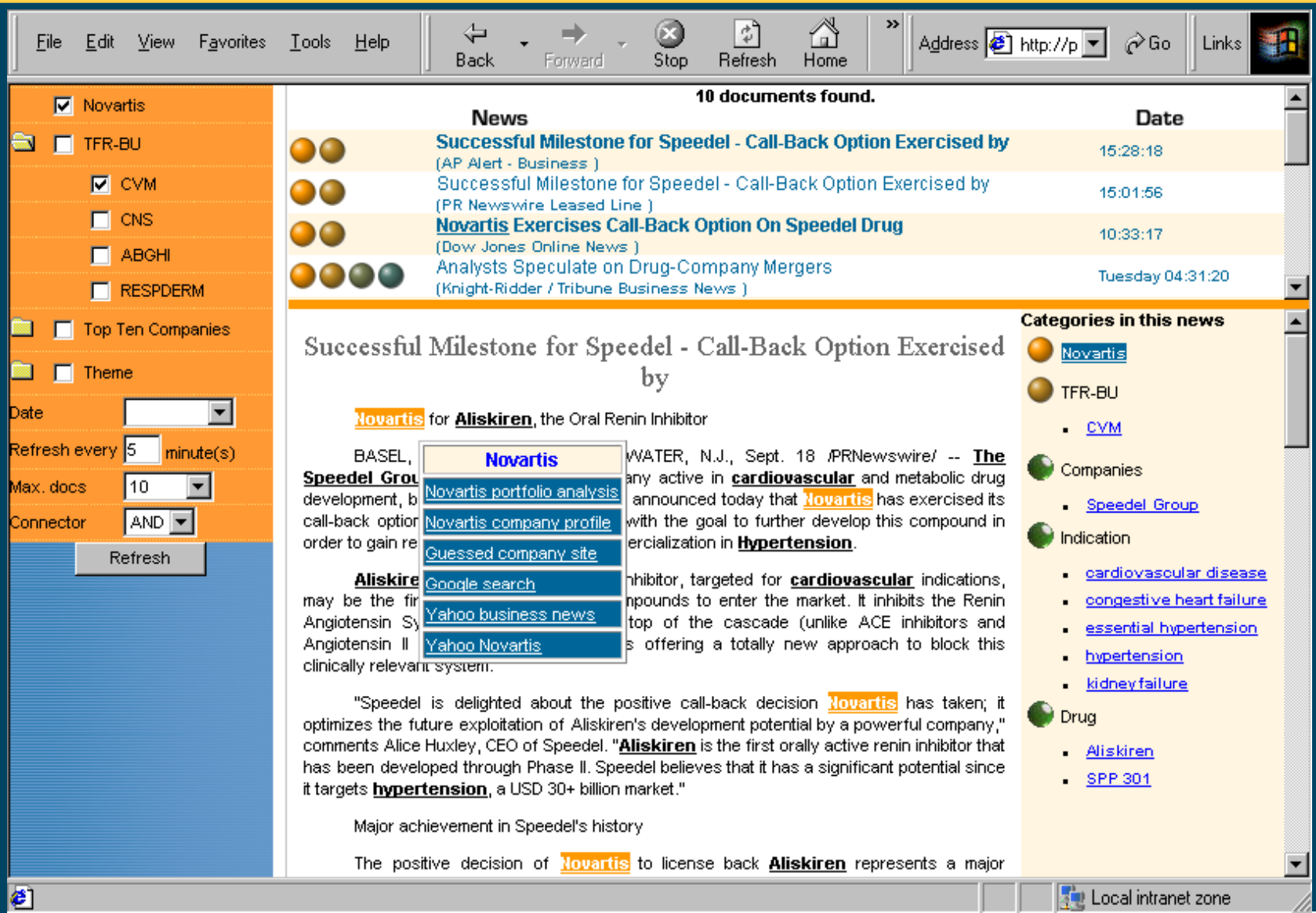
NOVARTIS

NewsFlow personalization



NOVARTIS

Ultralinks to pertinent and correctly accessed applications



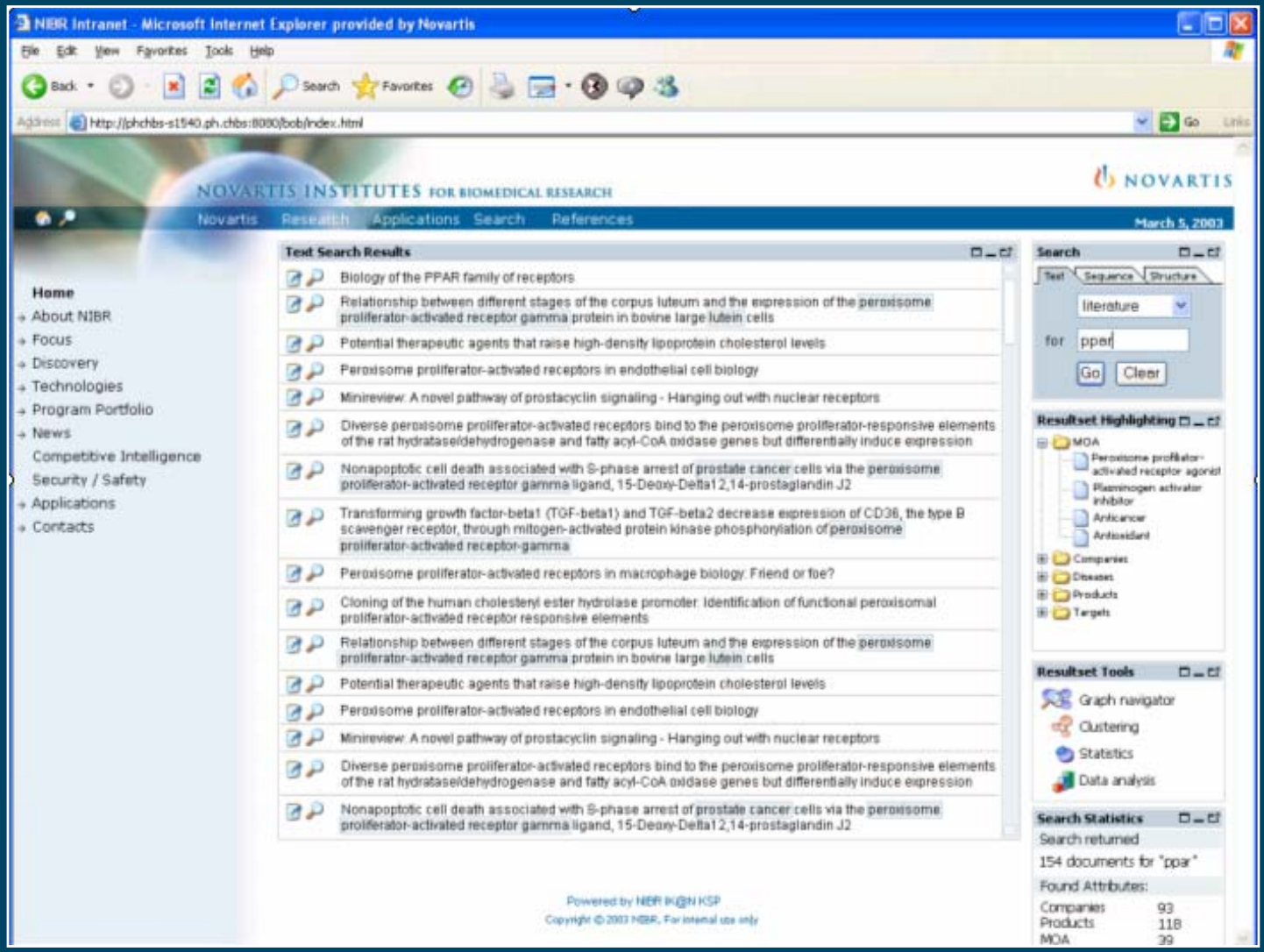
NOVARTIS

Knowledge Space Portal - Scope

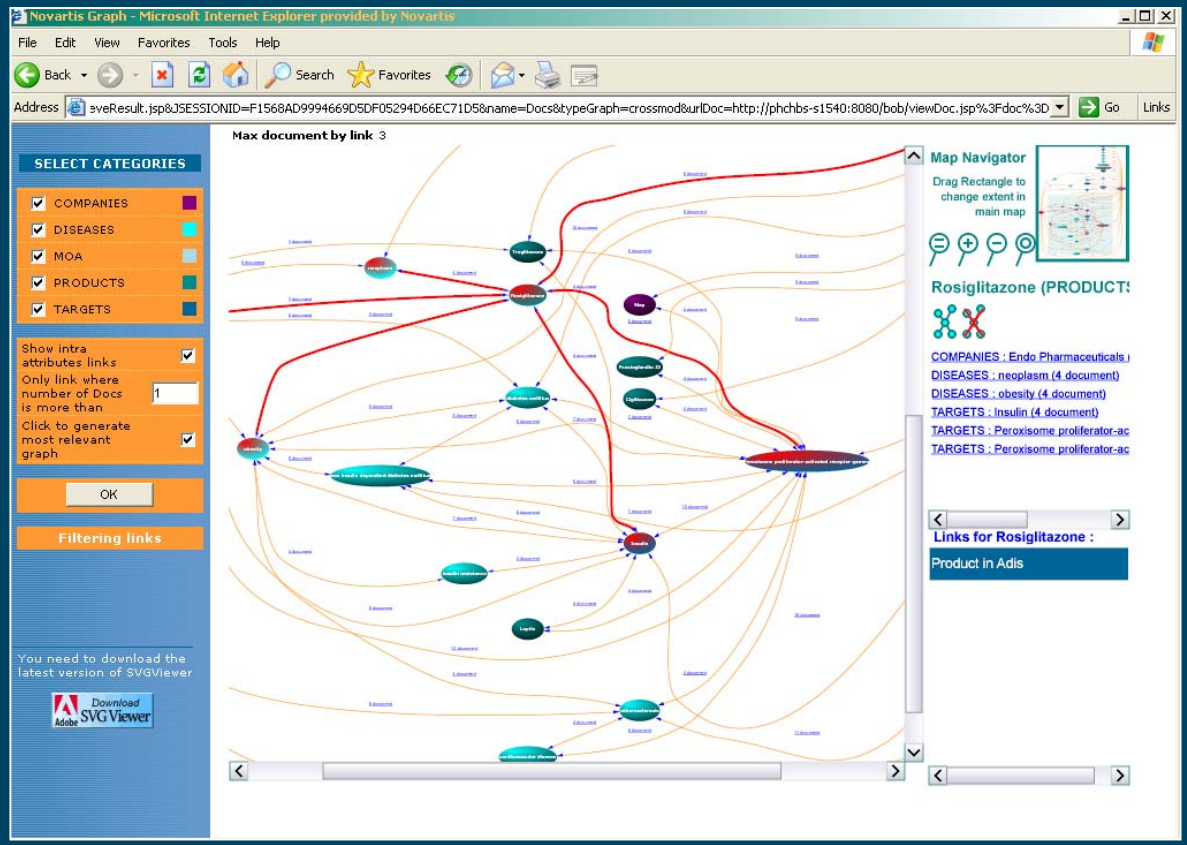
Provide key elements for efficiently accessing Novartis-internal and external information relevant to daily decision in the drug discovery and development process:

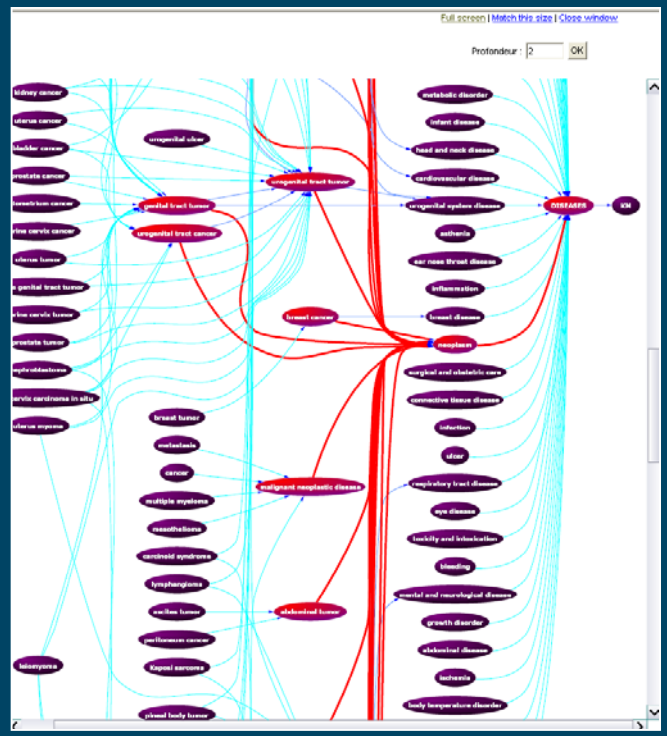
- Data integration across heterogeneous data sources and applications (internal and external)
- Consistent user interface for data retrieval, exploration and analysis across all data types
- Contextual (ultralink), tree-based (static or dynamic taxonomies) and semantic (knowledge map) navigation
- Data exploration and analysis methods
- Personalized views
- Collaborative, annotation and information sharing tools
- Alerting

Knowledge Space Portal Home Page



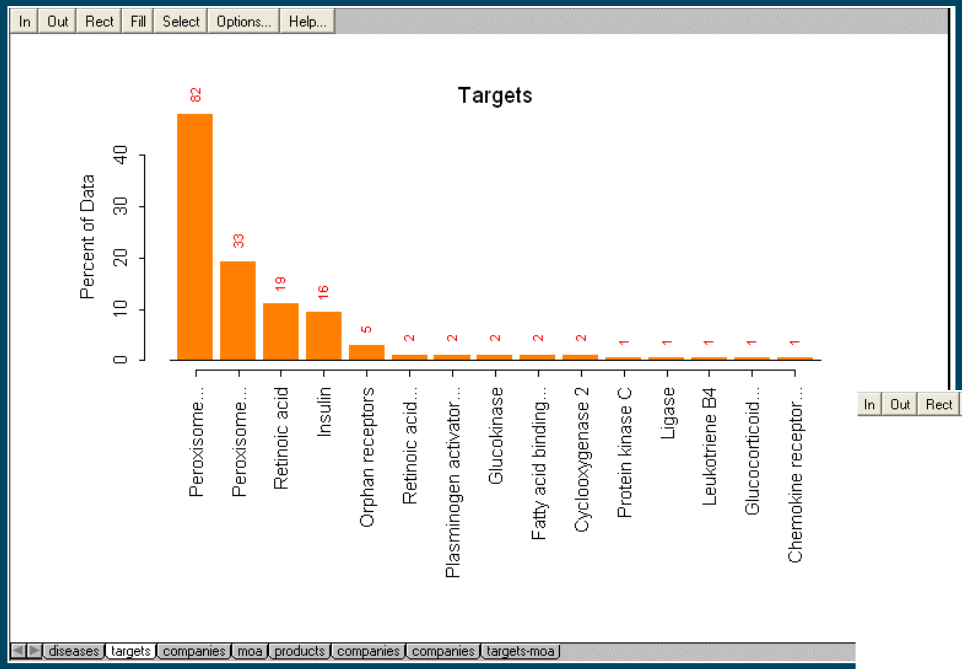
Navigation integrated on the Knowledge Space Portal

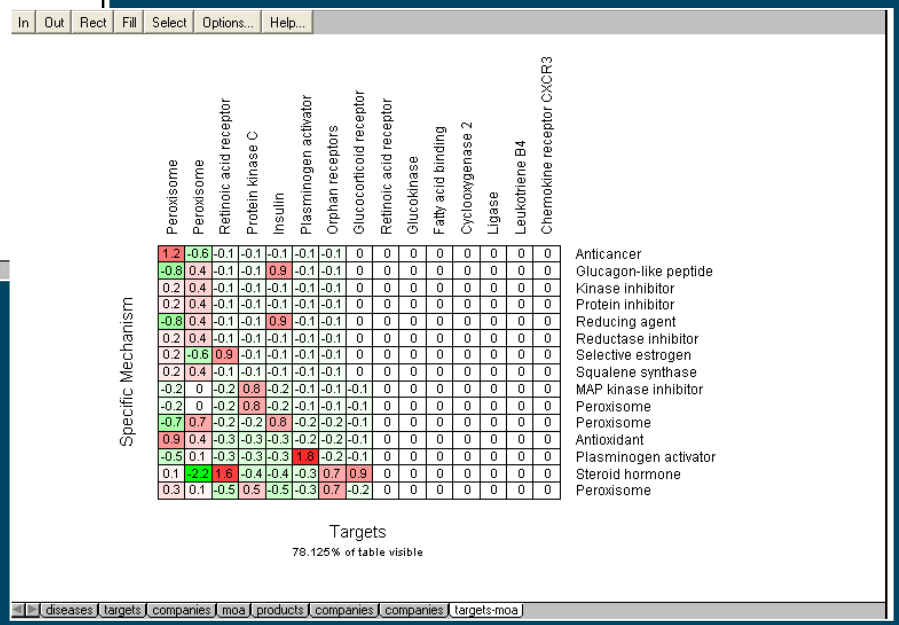




NOVARTIS

Data Analysis technologies integrated on the Knowledge Space Portal



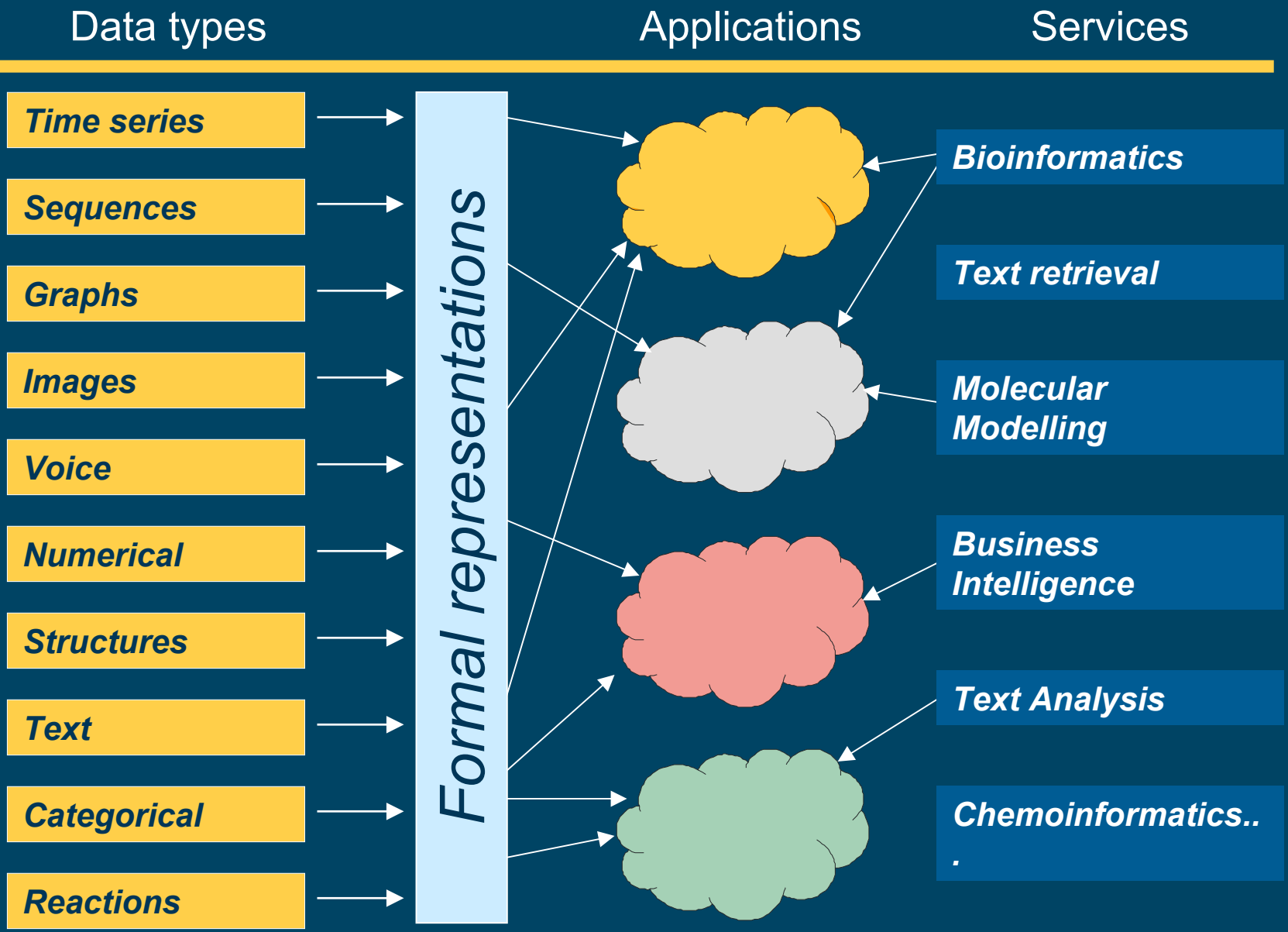


NOVARTIS

Future Steps

NOVARTIS

Data integration into a problem-solving environment



NOVARTIS