INterface FOr Rich MEtadata Resources

Thorsten Trippel, Sami Awad, Marc Bohnes, Patrick Dunkhorst, Carolin Kirchhof
Universität Bielefeld

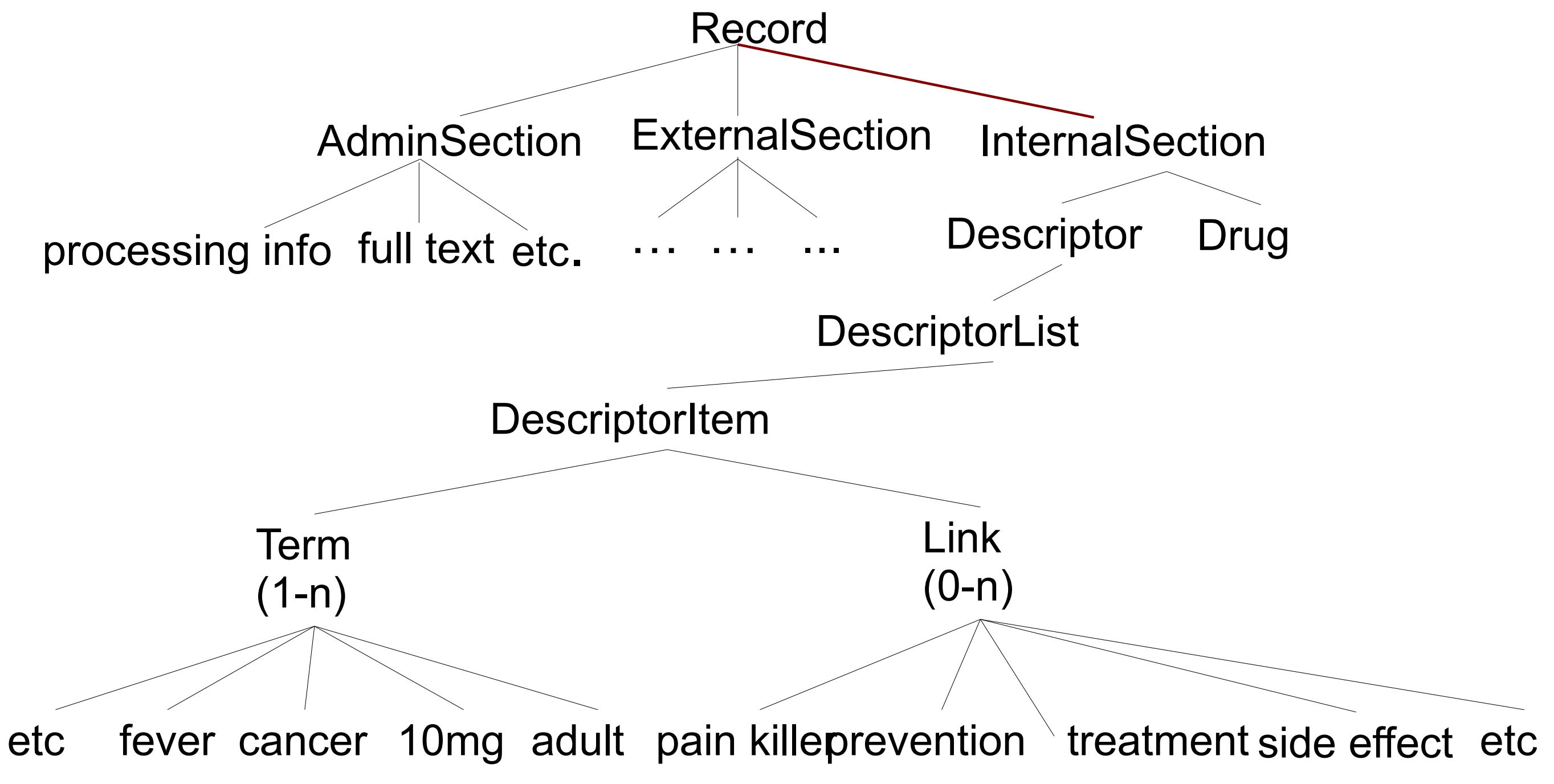
Motivation

- Language resources:
  - highly structured linguistic information
  - qualitative information: “small”
- Metadata for documentation
  - keywords
  - deep structure

- Challenge:
  - using the structure
  - providing access to data
  - harvesting metadata
- Technical threshold
  - without specialized querying skills
  - portable to other resources

Data Structure

- “Large” data set provided by project partner
  - NDA: data and concrete application
  - application in the medical and pharmaceutical domain
  - porting to other data of multimodal annotations
- data highly structured
- terminology database available for the appropriate domain



The keywords relevant for the search are present in the XML structure above (simplified) within DescriptorItem elements, either as term-link pairs, e.g. “treatment=pain killer” or as single terms such as “adult”.

Probabilistic Search Interface Element Ranking

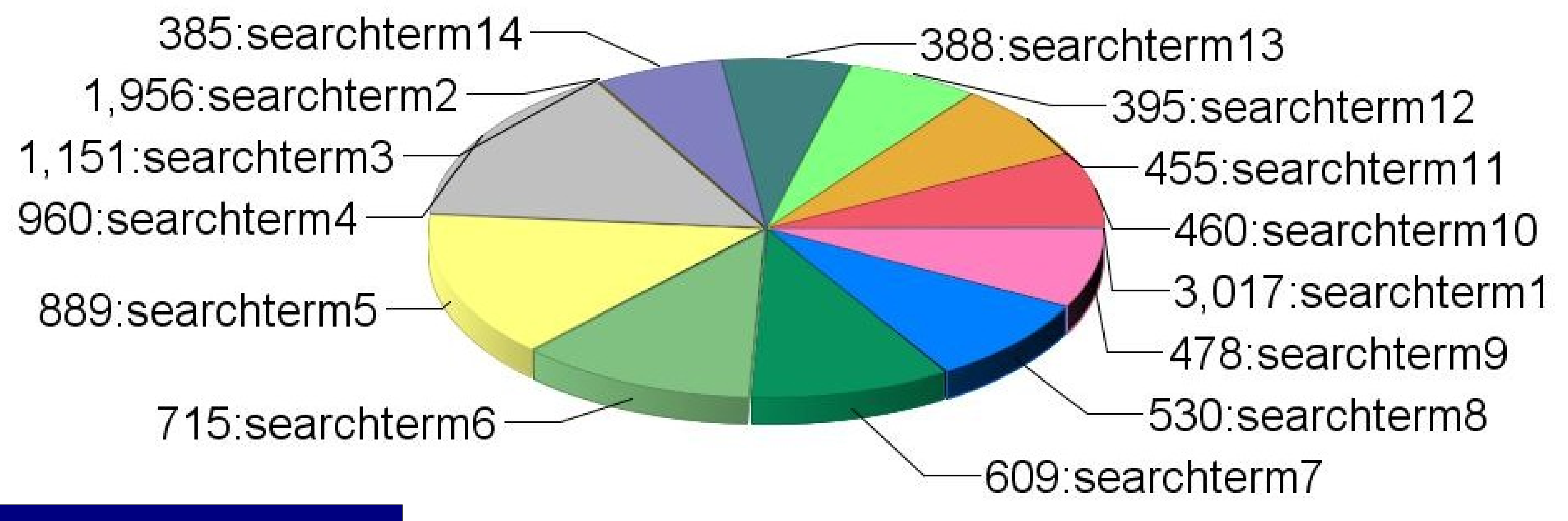
- regular grammar
- controlled vocabulary
- possibility of probability estimation for better interface integration

$$P(\omega_x \omega_y) = P(\omega_y \omega_x)$$

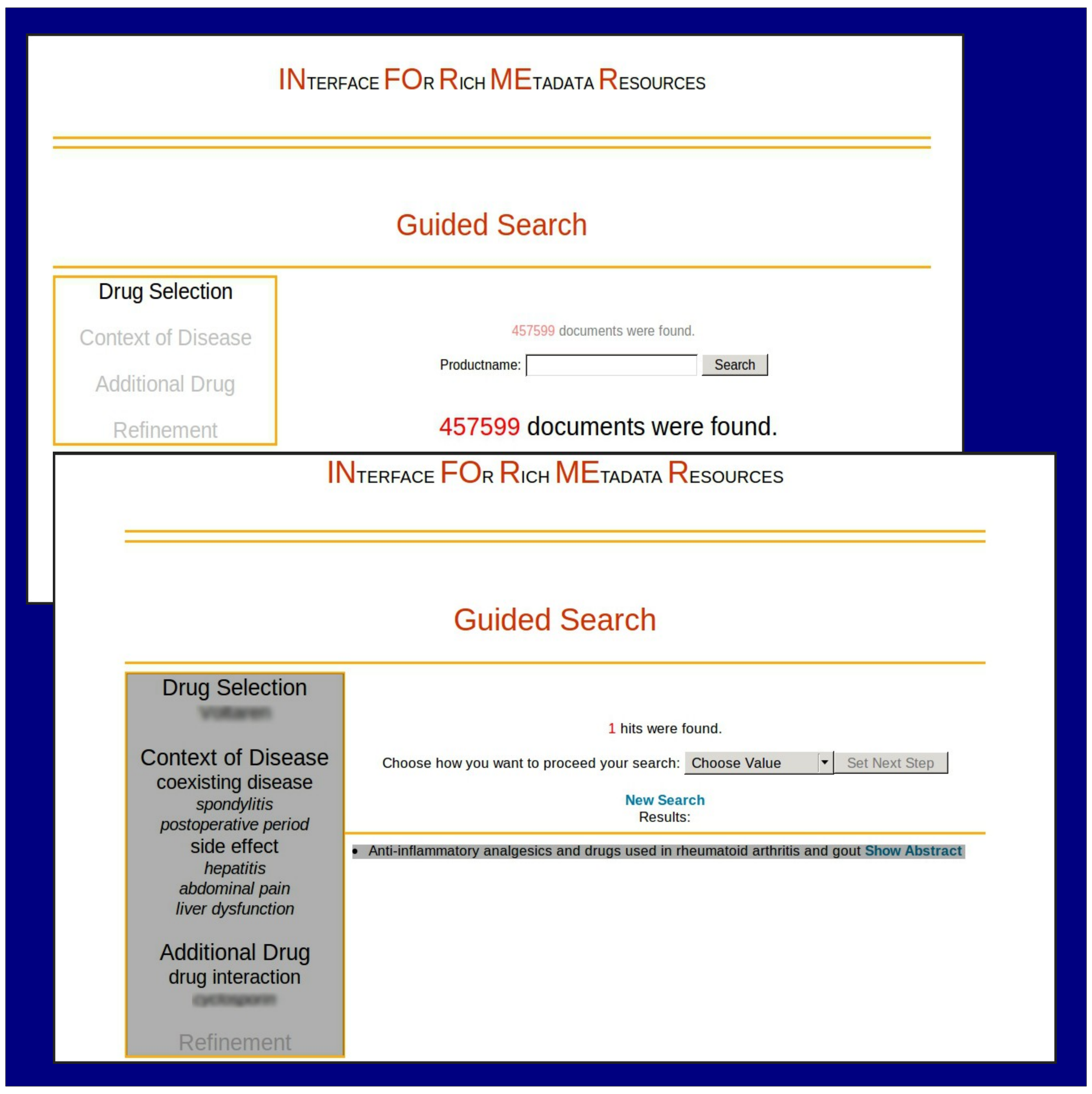
- Relative Frequency (RF) used for maximum likelihood estimation (MLE)
- general equation for word sequence of length n:

$$P(\omega_n) = \frac{C(\omega_n)}{\Sigma(C(\omega_{1+n}))} = 0.00..1.00$$

- assumption: recurrent patterns of search queries
- INFORMER can be optimized by statistically modeling user behavior
- RF provides statistical joint distribution of search queries in use



INFORMER Interface



457599 documents were found.

Productname:

457599 documents were found.

Drug Selection
[redacted]
Context of Disease
coexisting disease
spondylitis
postoperative period
side effect
hepatitis
abdominal pain
liver dysfunction
Additional Drug
drug interaction
[redacted]
Refinement

1 hits were found.

Choose how you want to proceed your search: Choose Value

Results:

- Anti-inflammatory analgesics and drugs used in rheumatoid arthritis and gout Show Abstract

Search Grammar

G = < Φ, T, R, Search>

Search ∈ Φ

Non-terminal symbols:
Φ = {Search, Drug, prevention, treatment, side effect, Links, etc.}

Terminal Symbols:
T = {pain killer, adult, 10mg, cancer, fever, Terms, etc. }

Rules:

Search → Drug (Context of Disease) (Additional Drug) (Refinement)
Context of Disease -> (Drug Therapy)* (Prevention)* (Diagnosis)* (Coexisting Disease)* (Side Effect)*

Additional Drug → (Combination) (Comparison) (Interaction)

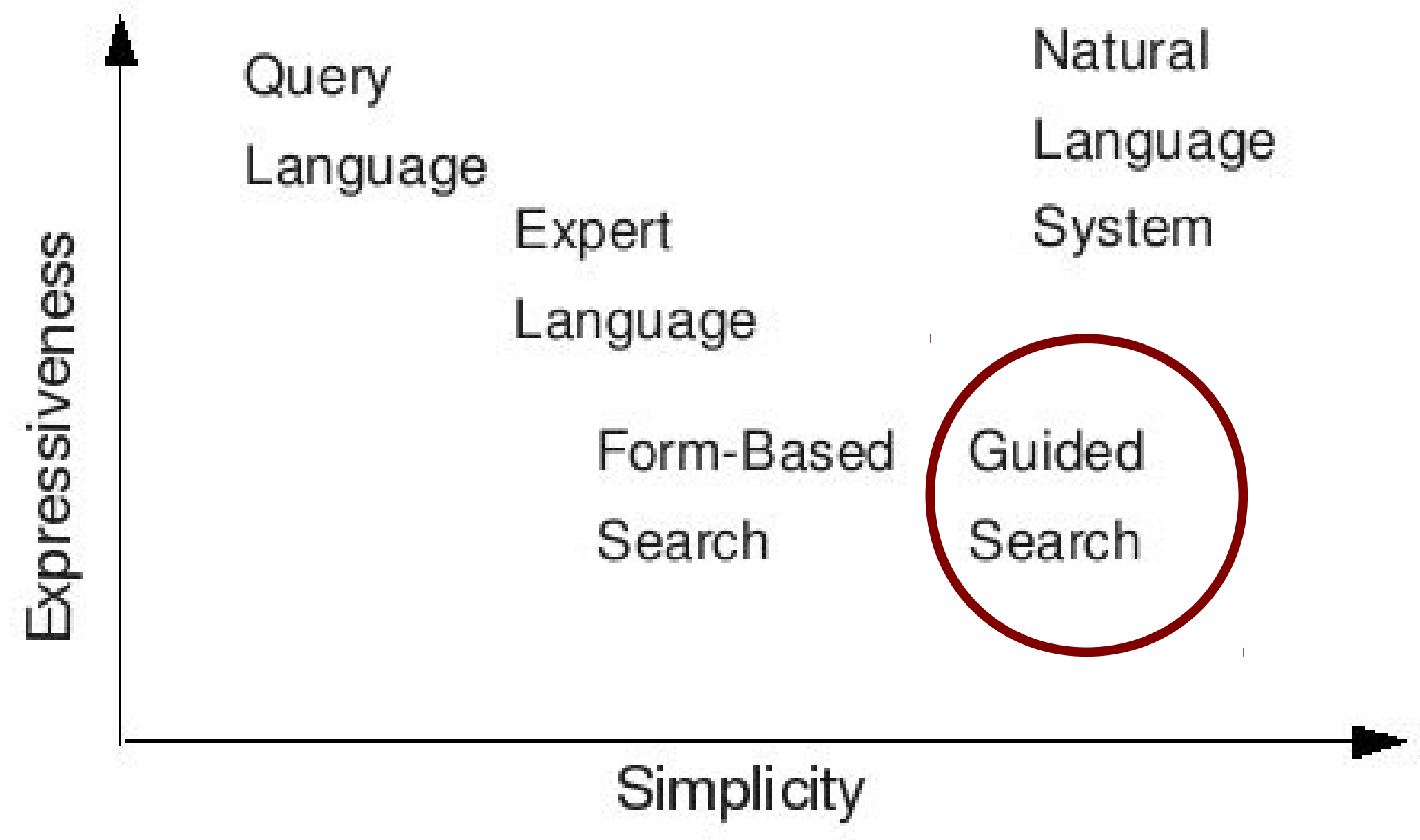
Refinement → (Type of study) (Age group) (Sex) (Route of Administration) (Dosage) (Free Search) (Duration of Treatment)

Drug → {...}

Complexity

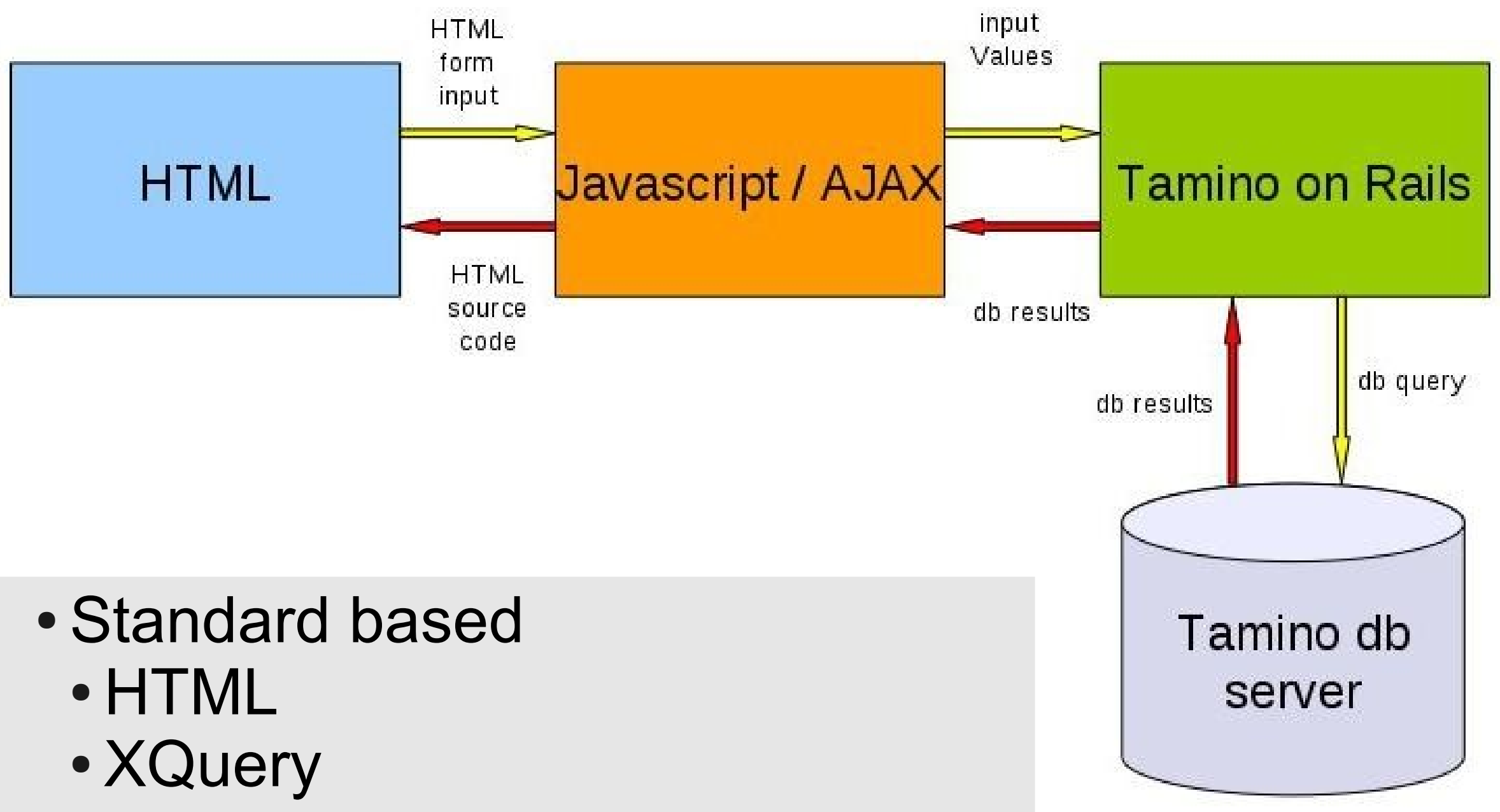
Complexity and its reduction:

- front-end: usability
- back-end: computational complexity, processing time



Idealized schema of interfaces to search engines. INFORMER ∈ Guided Search.

Implementation



- Standard based
  - HTML
  - XQuery
  - TBX
  - AJAX
- High performance
  - XML-Database engine
  - Rails framework

Synonym Search

Termbank use: TBX-Termbase
- concept based lexical resource
- search by synonym, hyponyms, related concept
- language restrictions: same language, all languages, specific language

```
<termEntry id="d1e64">
    <langSet xml:lang="en">
        <ntig>
            <termGrp>
                <term>painkillerXYZ</term>
                <termNote type="termType">GenericName</termNote>
            </termGrp>
        </ntig>
        <ntig>
            <termGrp>
                <term>12345</term>
                <termNote type="termType">EAN</termNote>
            </termGrp>
        </ntig>
    </langSet>
    <langSet xml:lang="de">
        <ntig>
            <termGrp>
                <term>SchmerzmittelXYZ</term>
                <termNote type="termType">GenericName</termNote>
            </termGrp>
        </ntig>
    </langSet>
</termEntry>
```

Conclusion

- usability of resources improved
- metadata used
- processing complexity reduced to linear complexity
- Untrained users:
  - high precision
  - high recall
- selection of sub-corpora for linguistic phenomena

Future work

- advancement to more linguistic resources
- more generic approach for tailoring the interface
- visualization of results
- reporting for technical analysis and optimization

Contact:{thorsten.trippel}@uni-bielefeld.de