

A methodology for supervised automatic document annotation

Kai Eckert

Computer Science Institute
University of Mannheim
Germany

**JCDL 2008, Doctoral Consortium
June 16th 2008
Pittsburgh, PA, USA**

Prototype

Please visit <http://www.kaiec.org>
for various Flash based demonstrations.

Background

- Traditional access to library contents: OPACs.
- Lot of users want something like this:



- 89% of college student information searches begin with a search engine (Rosa, 2006).
- “Grey Areas” between OPACs and Search Engines:
 - Google Books, Google Scholar
- Thesaurus Based Automatic Indexing as connector of OPACs and Search Engines?

Thesaurus Based Indexing

- Manual Indexing
 - Traditional bibliographic indexing
 - Using well-defined criteria, high quality
- Automatic Indexing
 - Crucial for large and fast growing document sets
 - News, Websites, Scientific Papers (Journal Papers and Conference Papers)
 - More index terms → Higher recall, especially for inexperienced users.
 - Precision?

State of the Art

- Thesaurus based automatic indexing is in use:
 - GoPubMed (Transinsight)
 - Collexis
 - CDS Invenio Document Server (CERN)
 - Medical Text Indexer (NLM Indexing Initiative)
 - CADIS
- Two motivations
 - Alternative approach for information retrieval
 - Supporting human indexers by generating suggestions

Problems of Automatic Indexing

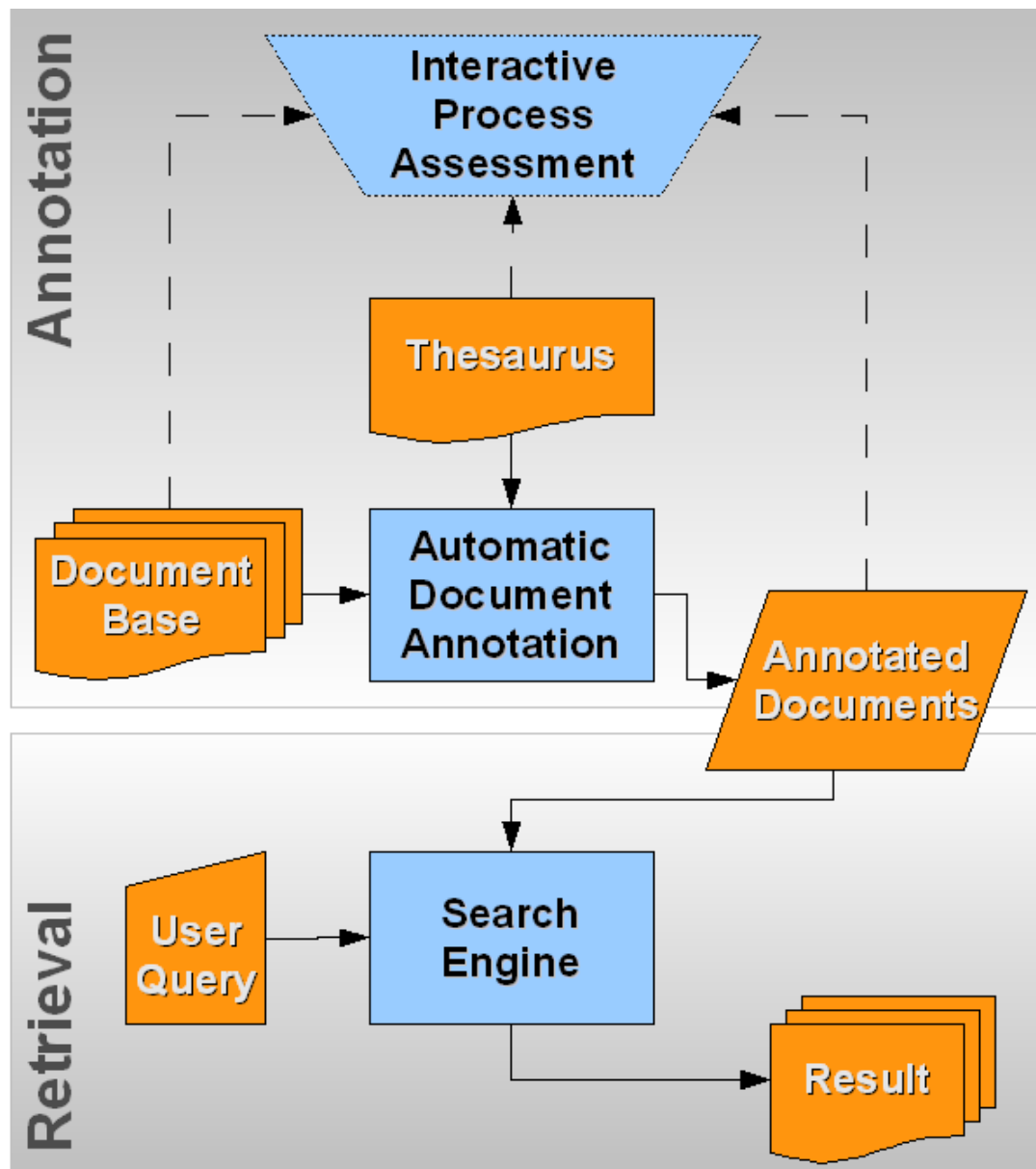
- Thesaurus:
 - Quality and suitability (regarding the domain)
- Indexing System:
 - Pitfalls in natural language processing
 - Quality of preprocessing (normalization, stemming)
 - Performance of disambiguation

Revision by a human expert needed.

Research Questions

- How can the quality of annotations be measured?
- Is it possible to visualize the annotation results globally?
- What kind of problems affect the annotation results?
- Is it possible to detect these problems?
Automatically?
- Which analysis methods and visualizations support the human detection of these problems?
- Is it possible to provide (proposed) solutions automatically?
- Is the quality of retrieval results improved significantly by the overall approach?

Semntinel Architecture





Performed Thesaurus Revisions

- Traditional
 - **Adaptation**, to reflect changes in the vocabulary.
 - **Deletion** and/or **Merging** of rarely used concepts.
 - **Splitting, extension** or **Restriction** of extensively used concepts.
 - **Review of the structure**, to avoid extensive subclassing.
- New
 - Identification of **problematic concepts** for automatic indexing.

Intuitive selection of problematic concepts

- Very **high** occurrence:
 - Too common – should be split into subconcepts
 - Not significant
 - *Indexing error*
- Very **low** occurrence:
 - Too specialized – could be merged with other concepts
 - Missing synonyms
 - Not significant
 - *Indexer failed to assign the concept*

Considering the hierarchy

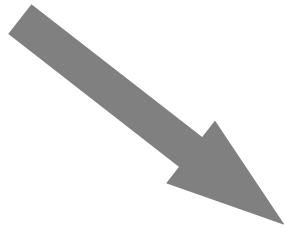
- **Higher** occurrence  **Higher** in the hierarchy
 - More common concepts 
- **Lower** occurrence **Lower** in the hierarchy
 - More specialized concepts

IC Diff Analysis

Information Content:

- Proposed by Resnik
- Depends on Frequency in Document Base

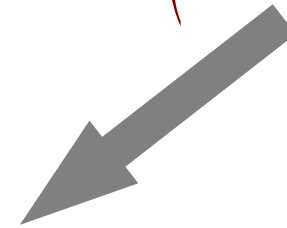
$$IC(c) = -\log P(c)$$



Intrinsic Information Content:

- Proposed by Seco, Veale und Hayes
- Based on the Number of Subconcepts

$$IIC(c) = -\log \left(\frac{hypo(c) + 1}{max} \right)$$



$$D_{IC}(c) = IC(c) - IIC(c)$$

Intuitive: A value between -1 and 1 that says, if a concept has a suspicious frequency regarding its position in the thesaurus.

Features and Limitations of IC Diff

- Supports all mentioned revisions with the exception of the adaption of the vocabulary.
- Judgement of the results is completely left to the user.
- Usability depends on the quality of the thesaurus or the quality of manually assigned index terms.
- The user has to search and browse the whole thesaurus to hunt down all (detected) potential problems.

Semtinel beyond IC Diff

- More sophisticated analysis, for example take the environment of concepts into account.
- Automated search for suspicious concepts
- Clustering techniques to find concepts with similar characteristics, revealing similar errors.
- Automatic identification of significant terms in the document base that do not exist in the thesaurus.

Next Steps

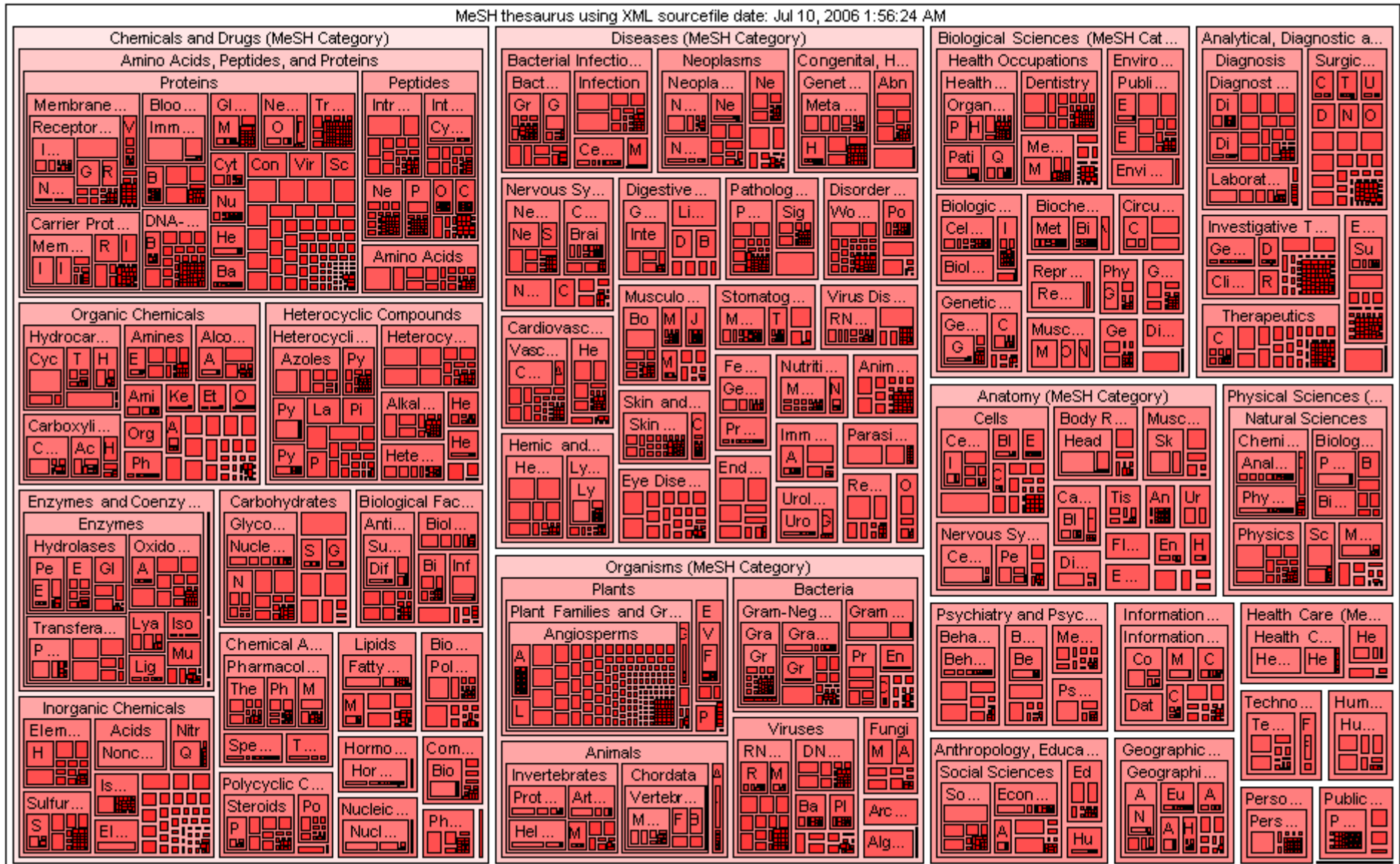
- Implementation of the sketched framework.
- Evaluation by library experts.
- Evaluation of actual retrieval results.
- Development of further analysis methods.

Thank you for your attention.

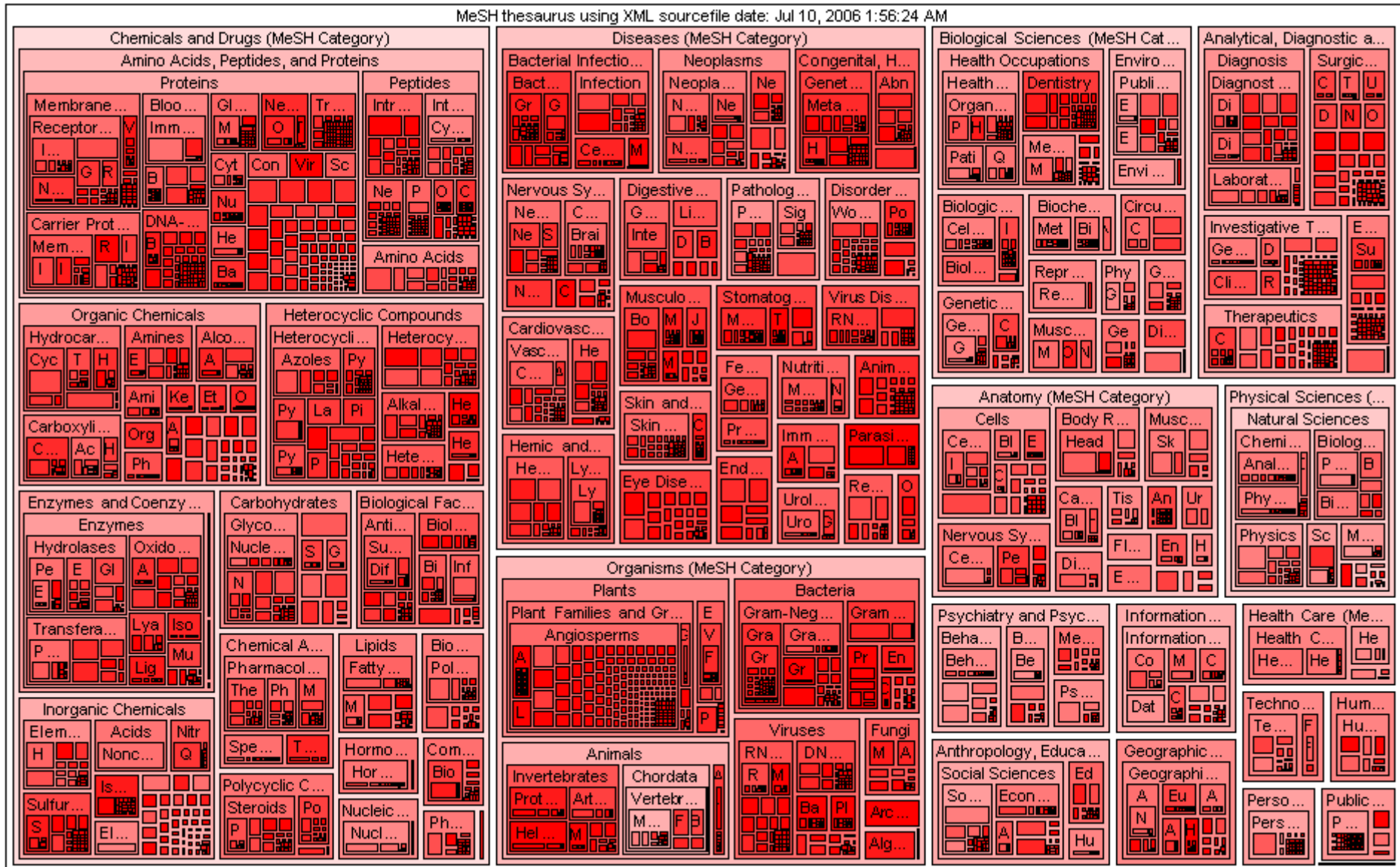
Questions or ideas?

kai@informatik.uni-mannheim.de

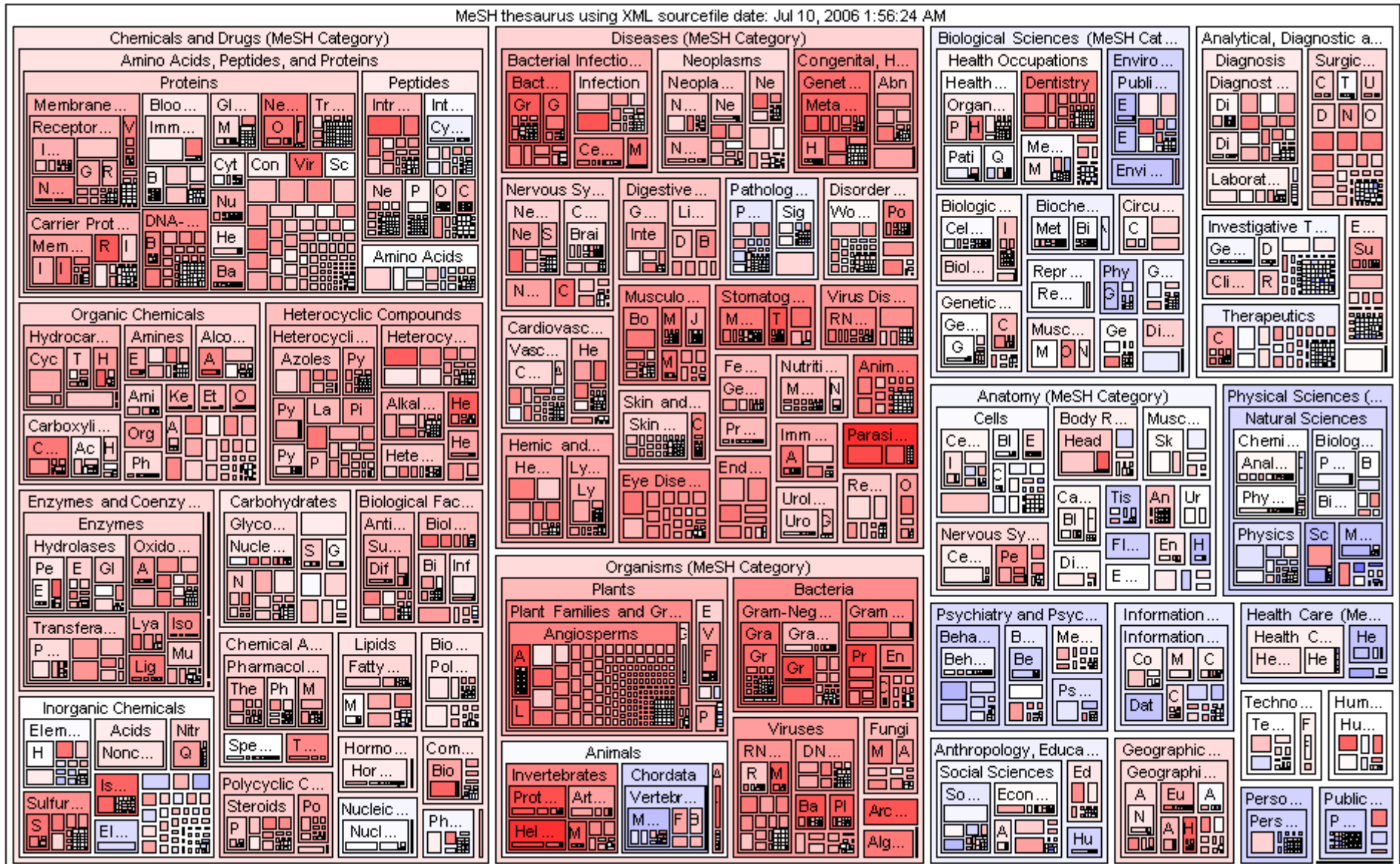
Intrinsic Information Content



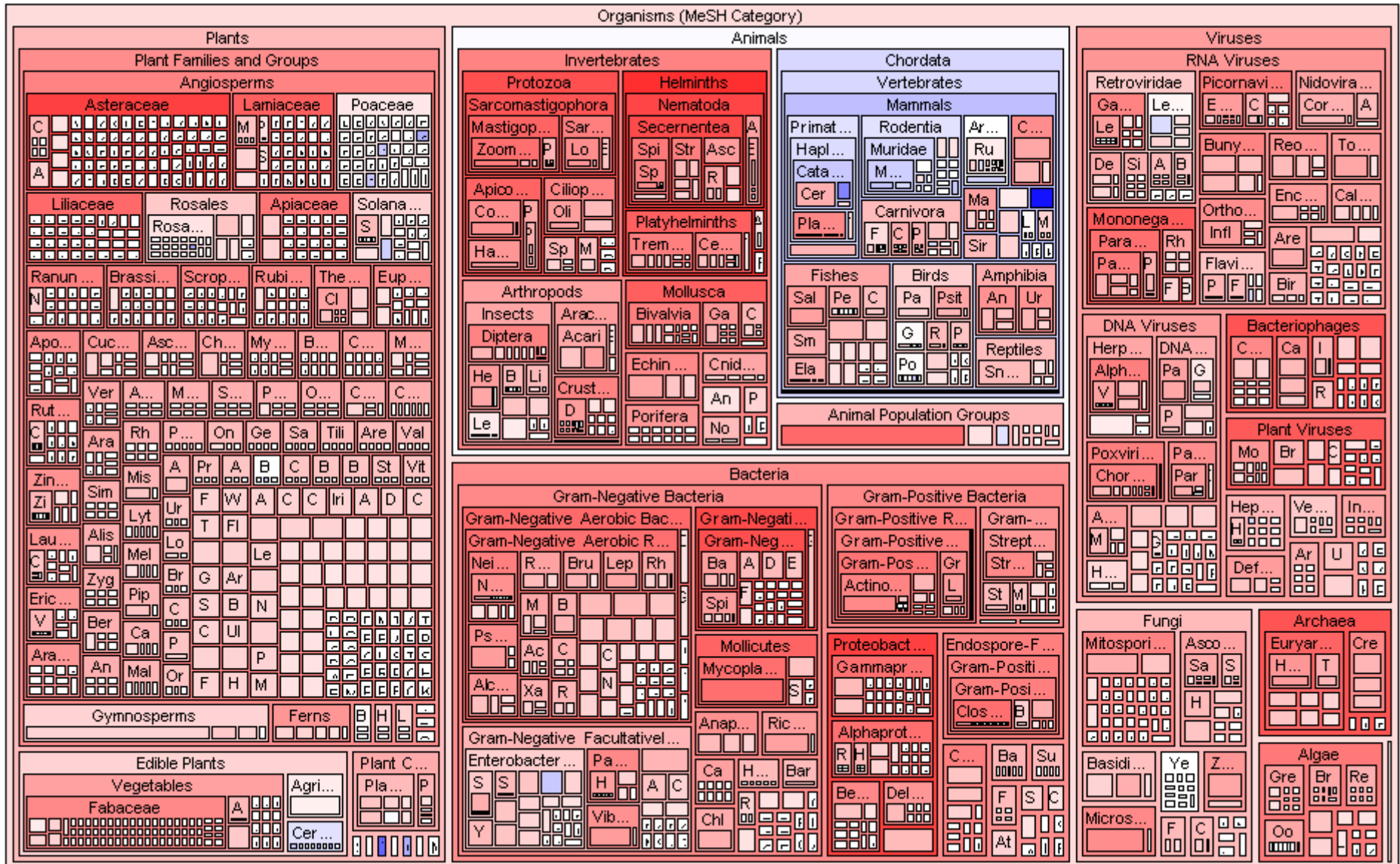
Information Content



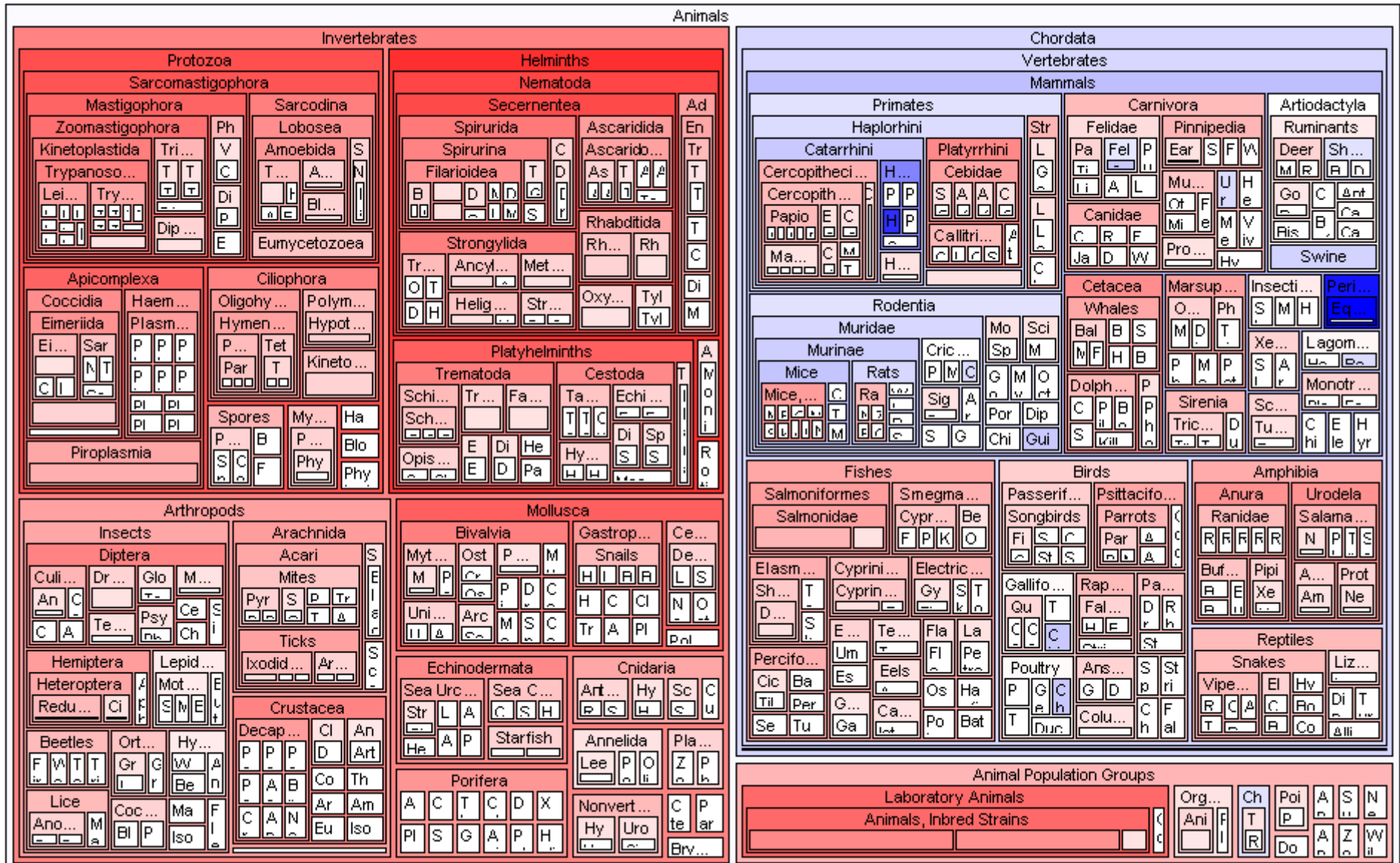
IC Diff



Organisms



Animals



Persons

Persons (MeSH Category)																																													
Occupational Groups										Persons																																			
Health Personnel										Age Groups				Patients																															
Allied Health Personnel					Nurses			Physicians		Infant		Adult		Child		Survivors		Patient																											
Dental Auxiliaries		Physician ...			Nurse Midwives	Male Nurses	Nurse Administrators	Women Physicians	Hospitalists	Infant, Low ...	Aged		Preschool Child	HIV Long-Term Survivors	Patient Dropouts	Hospitalized Adolescents																													
Dental ... Denturists	Dental Assistants	Pediatric Assistants								Infant, Very Low Birth Weight	Frail Elderly	Aged, 80 and over																																	
	Dental Hygienists	Ophthalmic Assistants			Nurse Practitioners	Nurse Anesthetists	Nurse Clinicians	Foreign Medical Graduates	Family Physicians	Infant, Small for Gestational	Middle Aged		Adolescent	Outpatients	Institutionalized Adolescents	Institutionalized Child																													
Medical S... Medical Receptionists		Pharmacists '		Home Health Aides	Medical Staff			Nursing ...		Dentists		Disabled Persons		Tissue Donors		Men		Exceptiona ...		Homeless ...																									
		Animal Technicians		Community Health Aides	Medical Staff, H... Hospitalists		Nursing Staff, Hospital		Women Dentists		Disabled Children		Mentally Disabled Persons		Living Donors		Blood Donors		Male Nurses		Gifted Child		Homeless Youth																						
Nurses &a...		Psychiatric Aides		Emergency Medical Technicians	Dental ... Dental Staff, Hospital		Infection Control Practitioners		Caregivers	Laboratory Personnel	Health Facility Administrators		Mentally Ill Persons		Hearing Impaired Persons		Volunt... Hospital Volunteers		Survivors		Child of Impaired Parents		Populatio n Groups		Medically Uninsured		Caregivers																		
													Amputees		Visually Impaired Persons																														
Hospital Personnel					Research Pers...			Librarians		Ethicists		Faculty		Clergy		Women		Multiple Birth Offspring		Homebound Persons		Siblings		Prisoners		Transient and Migrants		Research Personnel		Single Person															
Hospital Volu... Patient Escort Service		Hospital Auxiliaries		Nursing Staff, Hospital		Dental Staff, Hospital		Coroners and Medical Exami...		Health Educators		Medical Faculty		Physician Executives		Battered Women		Working Women		Women Dentists		Pregnant Women		Women Physicians		Triplets		Quadruplets		Twins		Quintuplets		Spouses		Friends		Parents		Students		Refugees		Sexual Partners	