# Taxonomy Based Indexing

- HEP taxonomy
- automatic indexing

HEP taxonomy

# HEP taxonomy

- hierarchical structure
- extension of DESY thesaurus
- elements:
  - preferred label
  - alternative labels (synonyms, abbreviations)
  - hidden labels (different spellings etc, regexp)
  - broader terms
  - narrower terms
  - related terms
  - history (e.g. name changes)
  - scope (range of application)
  - definition
- stored as rdf

HEP taxonomy

# Taxonomy entry

- main keyword:

  dynamical symmetry breaking

  a: DSB
  h: /dynamical break\w+/
  h: dynamically broken
  b: symmetry breaking
  r: spontaneous symmetry breaking

- combined keyword (pair of 2 main keywords):

  Higgs particle: doublet

  a: Higgs doublet
  a: 2HDM
  b: Higgs particle: multiplet

# Automatic indexing program

- Python script
- counts the occurrence of preferred, alternative and hidden labels of a keyword in a pdf or text file
- goes through the list of found main keywords, checks whether 2 of them appear as immediate neighbors and adds a count if this pair constitutes a legal combined keyword

HEP taxonomy

# Example

gr-qc/0607071

Title: A study of gravitational collapse with decaying of the vacuum energy

Authors: M. de Campos

Abstract: We study the gravitational collapse of a spherically symmetric star made of a dark matter dust in a background with $\Lambda $ and consider that $\Lambda$ decay into dark particles. The approach adopted assumes a modified matter expansion rate and we have formation of a black hole since that we have the formation of an apparent horizon. A brief comparison of the process of the matter condensation using the gravitational collapse approach and the linear scalar perturbation theory is considered.

HEP taxonomy

# Sample output

Composites:

40 gravitation: collapse [24, 28]
5 energy: density [23, 9]
3 dark energy: density [17, 9]
2 matter: condensation [26, 2]
2 tensor: energy-momentum [3, 2]
2 radiation: background [2, 5]
2 star: collapse [18, 28]
2 perturbation: linear [3, 2]
2 perturbation: scalar [3, 5]
2 particle: production [11, 6]
2 field theory: scalar [0, 0]
1 black hole: mass [11, 5]
1 galaxy: rotation [4, 1]
1 energy: vacuum [23, 12]
1 matter: density [26, 9]
1 energy: decay [23, 8]
1 dark matter: density [18, 9]

Main Keywords:

11 horizon
9 formation
9 cosmological constant
7 cloud
7 surface
4 pressure
4 supernova
4 general relativity
3 redshift
3 fluid
2 field equations
3 conservation law
2 perturbation theory
2 anthropic principle
2 equation of state
2 anisotropy
2 acceleration

manually:

dark matter: star
symmetry: rotation
gravitation: collapse
cosmological constant: decay
black hole: formation
matter: condensation
dark energy: density
energy: vacuum
perturbation: scalar
perturbation theory: linear

HEP taxonomy

# Applications

- Current:
  - runs regularly on arXiv pdf's
  - proposes keywords for human indexers
- Planned:
  - run on journal abstracts
  - create temporary keywords immediately with bibliographic data

HEP taxonomy

# Future developments

- taxonomy
  - short term:
    - Hierarchical relations
    - Definition of classes (arxiv classifications, detectors, accelerators…)
    - Filling of alt and hidden labels
    - Selection and adaptation of an rdf editor
  - long term:
    - User-friendly presentation of the ontology
    - Links to external resources
    - Search tool
    - Basis for author keywords

HEP taxonomy

# Future developments

- keyword extraction program:
  - Weighting
  - Further noise reduction
  - Preprocessing of text file
  - Exploit syntax
  - Speed up performance
  - Convert into layered library with plug-in support