(Semi)-automated subject indexing of Swedish resources:

Evaluating (a combination of) cataloguers’, end users’ and automated index terms in retrieval

Koraljka Golub

International Conference on Dublin Core and Metadata Applications 2016 (DC-2016)

Linnæus University

LIBRIS

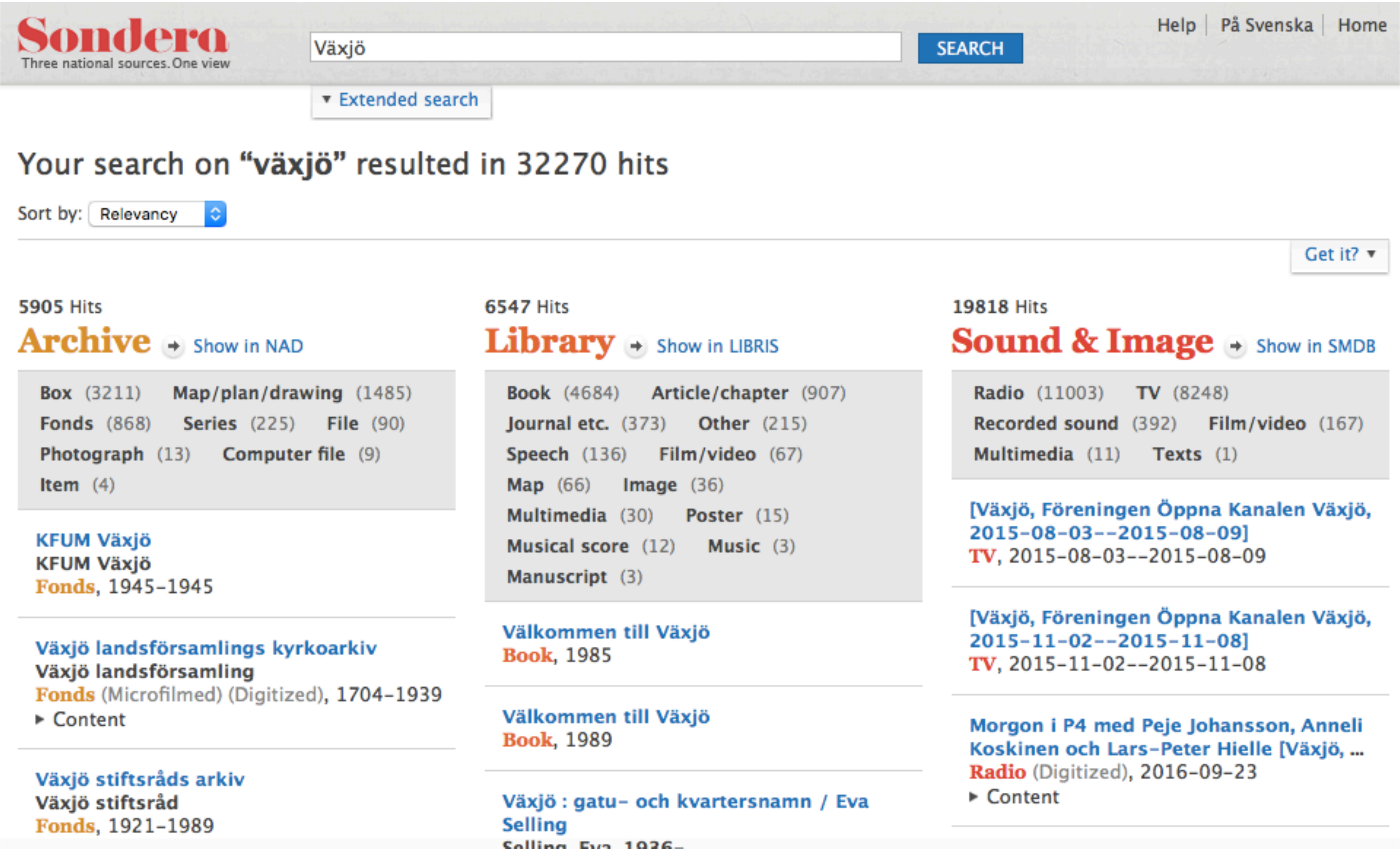
Sondera

Three national sources. One view

Purpose

SwePub

- Improve searching and browsing for information in the Swedish language, esp. in interoperable information systems like LIBRIS, SwePub, Sondera
- Subject index terms from controlled vocabularies like SAO and DDC offer:
  - Uniformity of term format
  - Provide context
  - Browsing
- But, expensive while increase of digital documents → 2 possible solutions:
  - 1) (semi)-automated solutions
  - 2) author/end-user tagging



Your search on "växjö" resulted in 32270 hits

Sort by: Relevancy

5905 Hits

Archive → Show in NAD

Box (3211) **Map/plan/drawing** (1485)
Fonds (868) **Series** (225) **File** (90)
Photograph (13) **Computer file** (9)
Item (4)

KFUM Växjö
KFUM Växjö
Fonds, 1945-1945

Växjö landsförsamlings kyrkoarkiv
Växjö landsförsamling
Fonds (Microfilmed) (Digitized), 1704-1939
▸ Content

Växjö stiftsråds arkiv
Växjö stiftsråd
Fonds, 1921-1989

6547 Hits

Library → Show in LIBRIS

Book (4684) **Article/chapter** (907)
Journal etc. (373) **Other** (215)
Speech (136) **Film/video** (67)
Map (66) **Image** (36)
Multimedia (30) **Poster** (15)
Musical score (12) **Music** (3)
Manuscript (3)

Välkommen till Växjö
Book, 1985

Välkommen till Växjö
Book, 1989

Växjö : gatu- och kvartersnamn / Eva Selling
Selling, Eva, 1936-

19818 Hits

Sound & Image → Show in SMDB

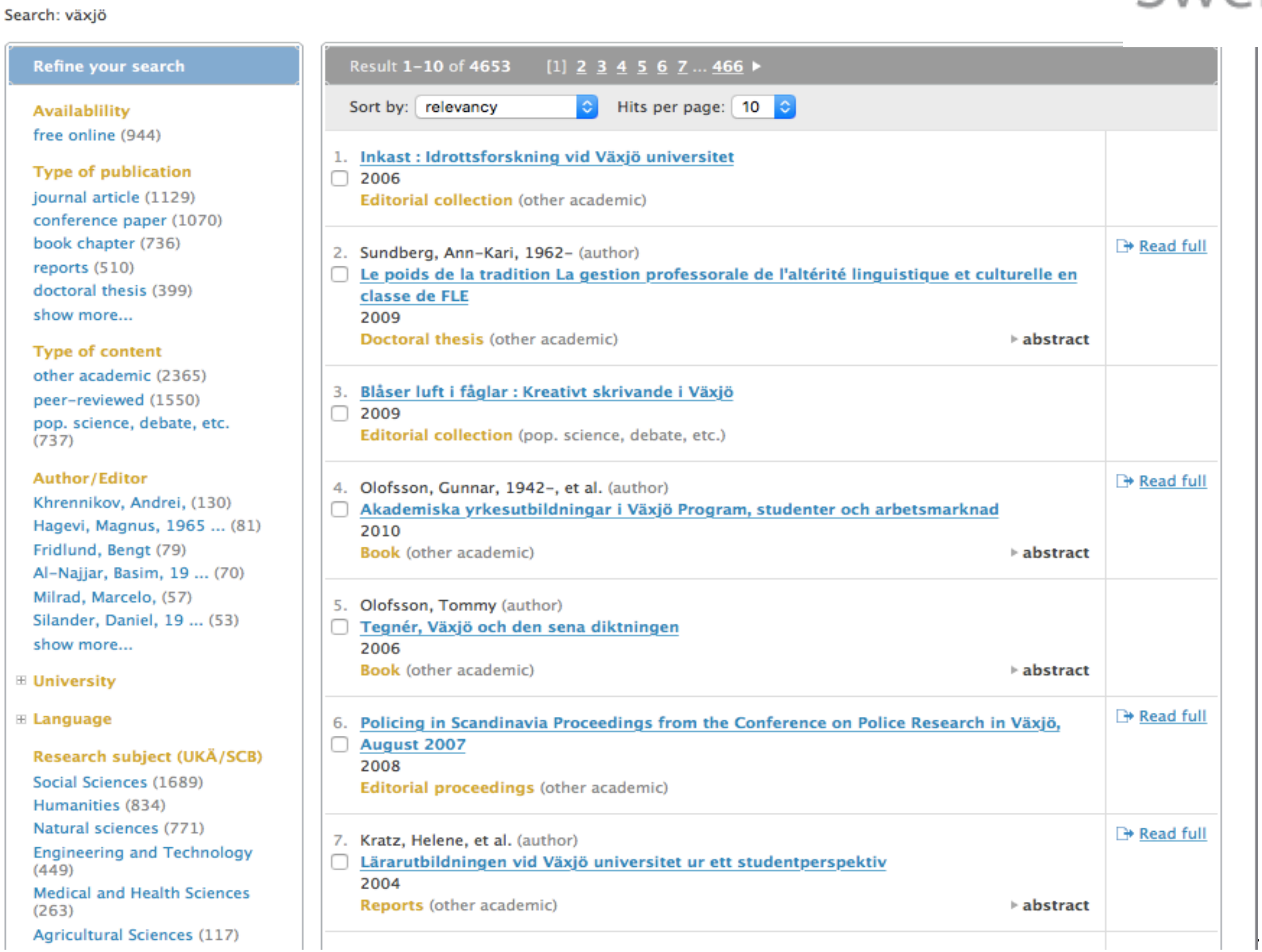
Radio (11003) **TV** (8248)
Recorded sound (392) **Film/video** (167)
Multimedia (11) **Texts** (1)

[Växjö, Föreningen Öppna Kanalen Växjö, 2015-08-03--2015-08-09]
TV, 2015-08-03--2015-08-09

[Växjö, Föreningen Öppna Kanalen Växjö, 2015-11-02--2015-11-08]
TV, 2015-11-02--2015-11-08

Morgon i P4 med Peje Johansson, Anneli Koskinen och Lars-Peter Hielle [Växjö, ...
Radio (Digitized), 2016-09-23
▸ Content

SwePub



Search: växjö

Refine your search

Availablility
free online (944)

Type of publication
journal article (1129)
conference paper (1070)
book chapter (736)
reports (510)
doctoral thesis (399)
show more...

Type of content
other academic (2365)
peer-reviewed (1550)
pop. science, debate, etc. (737)

Author/Editor
Khrennikov, Andrei, (130)
Hagevi, Magnus, 1965 ... (81)
Fridlund, Bengt (79)
Al-Najjar, Basim, 19 ... (70)
Milrad, Marcelo, (57)
Silander, Daniel, 19 ... (53)
show more...

⊞ **University**

⊞ **Language**

Research subject (UKÄ/SCB)
Social Sciences (1689)
Humanities (834)
Natural sciences (771)
Engineering and Technology (449)
Medical and Health Sciences (263)
Agricultural Sciences (117)

Result 1-10 of 4653 [1] 2 3 4 5 6 7 ... 466 ▸

Sort by: relevancy Hits per page: 10

1. □ Inkast : Idrottsforskning vid Växjö universitet
2006
Editorial collection (other academic)

2. □ Sundberg, Ann-Kari, 1962- (author)
Le poids de la tradition La gestion professorale de l'altérité linguistique et culturelle en classe de FLE
2009
Doctoral thesis (other academic) ▸ abstract
Read full

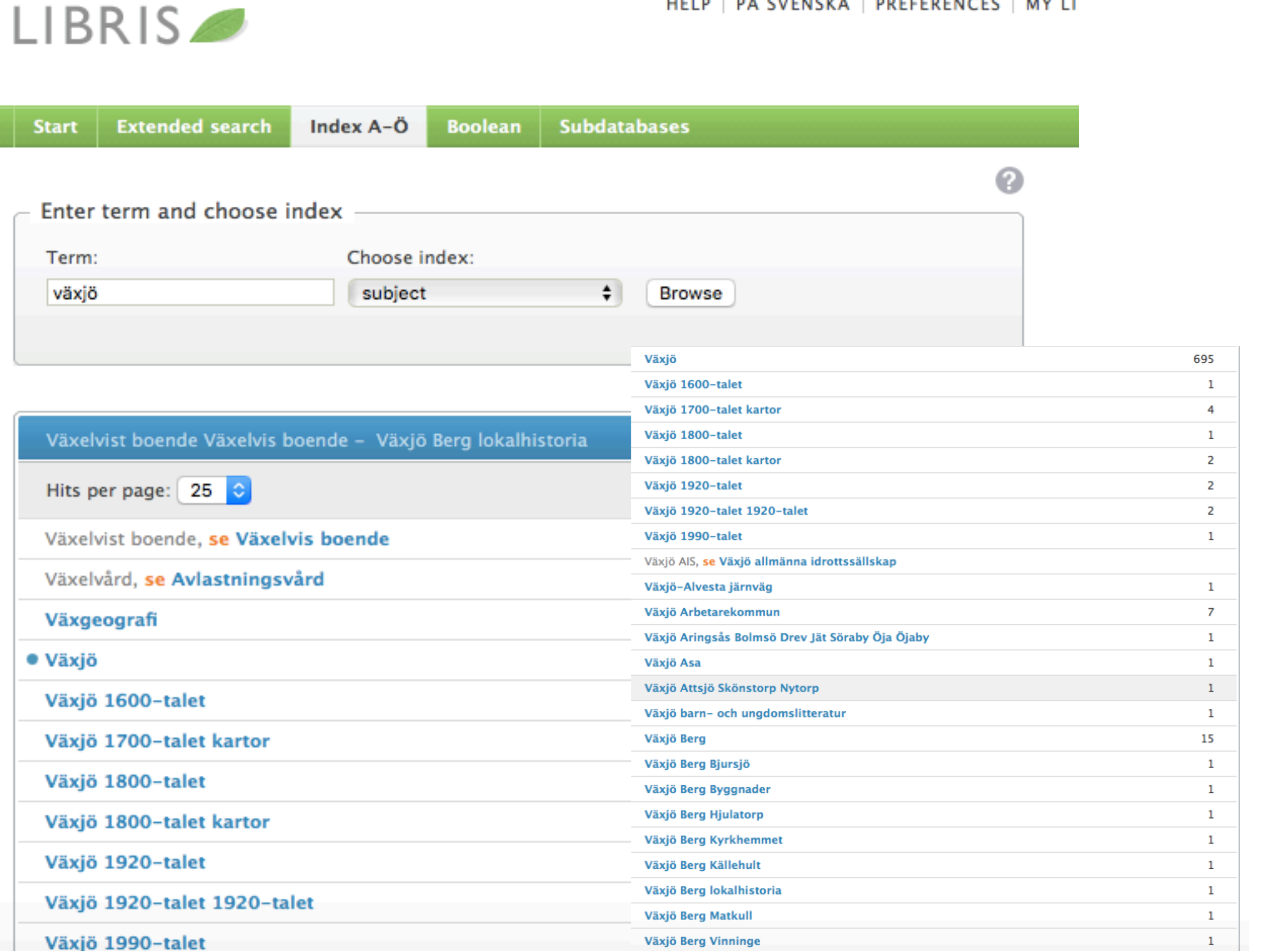
3. □ Blåser luft i fåglar : Kreativt skrivande i Växjö
2009
Editorial collection (pop. science, debate, etc.)

4. □ Olofsson, Gunnar, 1942-, et al. (author)
Akademiska yrkesutbildningar i Växjö Program, studenter och arbetsmarknad
2010
Book (other academic) ▸ abstract
Read full

5. □ Olofsson, Tommy (author)
Tegnér, Växjö och den sena diktningen
2006
Book (other academic) ▸ abstract

6. □ Policing in Scandinavia Proceedings from the Conference on Police Research in Växjö, August 2007
2008
Editorial proceedings (other academic)
Read full

7. □ Kratz, Helene, et al. (author)
Lärarutbildningen vid Växjö universitet ur ett studentperspektiv
2004
Reports (other academic) ▸ abstract
Read full



Enter term and choose index

Term: växjö

Choose index: subject

Växelvist boende Växelvis boende – Växjö Berg lokalhistoria

Hits per page: 25

Växelvist boende, se Växelvis boende

Växelvård, se Avlastningsvård

Växgeografi

● Växjö

Växjö 1600-talet

Växjö 1700-talet kartor

Växjö 1800-talet

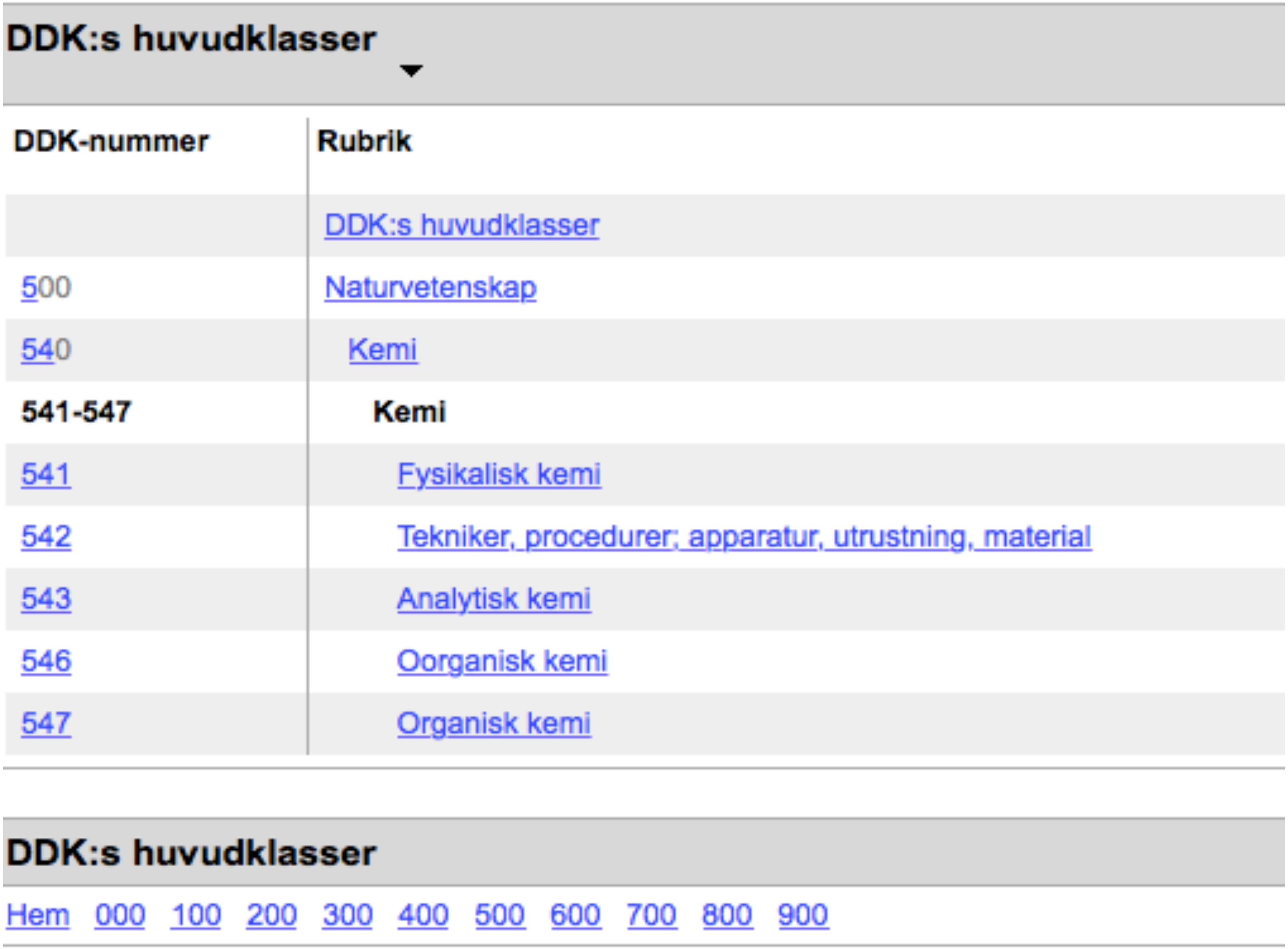
Växjö 1800-talet kartor

Växjö 1920-talet

Växjö 1920-talet 1920-talet

Växjö 1990-talet

Växjö	695
Växjö 1600-talet	1
Växjö 1700-talet kartor	4
Växjö 1800-talet	1
Växjö 1800-talet kartor	2
Växjö 1920-talet	2
Växjö 1920-talet 1920-talet	2
Växjö 1990-talet	1
Växjö AIS, se Växjö allmänna idrottssällskap	
Växjö-Alvesta järnväg	1
Växjö Arbetarekommun	7
Växjö Aringsås Bolmsö Drev Jät Söraby Öja Öjaby	1
Växjö Asa	1
Växjö Attsjö Skönstorp Nytorp	1
Växjö barn- och ungdomslitteratur	1
Växjö Berg	15
Växjö Berg Bjursjö	1
Växjö Berg Byggnader	1
Växjö Berg Hjulatorp	1
Växjö Berg Kyrkhemmet	1
Växjö Berg Källehult	1
Växjö Berg lokalhistoria	1
Växjö Berg Matkull	1
Växjö Berg Vinninge	1



DDK:s huvudklasser

DDK-nummer	Rubrik
	DDK:s huvudklasser
500	Naturvetenskap
540	Kemi
541-547	**Kemi**
541	Fysikalisk kemi
542	Tekniker, procedurer; apparatur, utrustning, material
543	Analytisk kemi
546	Oorganisk kemi
547	Organisk kemi

DDK:s huvudklasser

Hem 000 100 200 300 400 500 600 700 800 900

Aims

1. Find out to what degree it is possible to apply automated subject indexing based on:

- Controlled indexing languages like the Dewey Decimal Classification (DDC) and Swedish Subject Headings (SAO)
- Derived indexing of keywords from the resource itself

2. Determine the value of automatically assigned index terms, in combination and comparison with end-user and cataloguers index terms in the process of information retrieval by end users

BACKGROUND

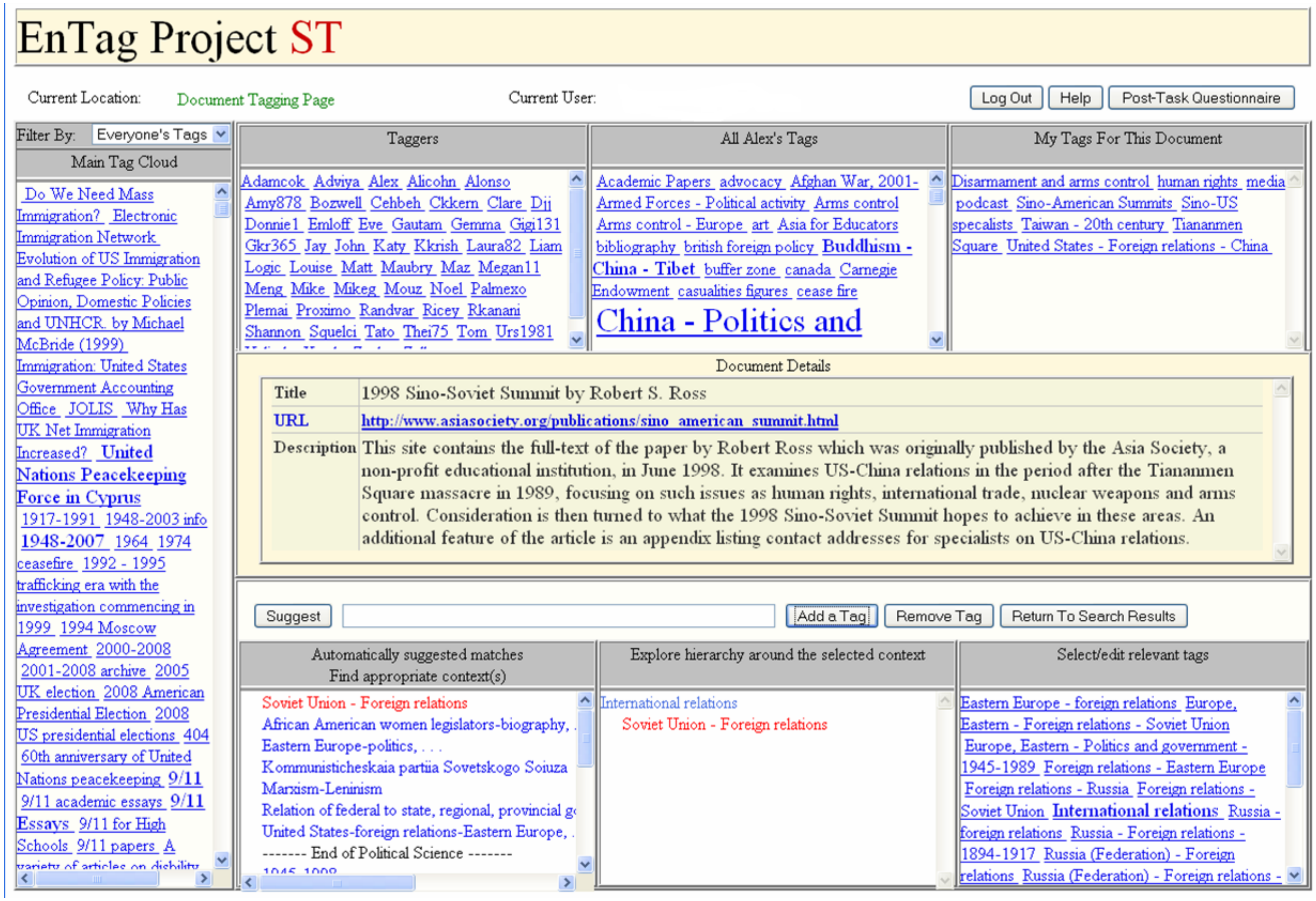
End-user indexing

- Author keywords or social tags in Web 2.0 services also provided by library catalogues
- Cheaper, provide additional perspectives like new scientific terms
- But, no control of word forms, homonymy, polisemy or synonymy

"*…the cost savings made in the provision of low-quality indexing are cancelled out by the high costs incurred by searchers who fail either to find everything that they want (low recall) or, often more frustratingly, to avoid everything that they do not want (low precision)…*"

(Furner 2010, 1861)

Offering users to choose from KOS



Current Location: Document Tagging Page

Current User:

Filter By: Everyone's Tags

Main Tag Cloud

Do We Need Mass Immigration? Electronic Immigration Network Evolution of US Immigration and Refugee Policy: Public Opinion, Domestic Policies and UNHCR by Michael McBride (1999) Immigration: United States Government Accounting Office JOLIS Why Has UK Net Immigration Increased? United Nations Peacekeeping Force in Cyprus 1917-1991 1948-2003 info 1948-2007 1964 1974 ceasefire 1992 - 1995 trafficking era with the investigation commencing in 1999 1994 Moscow Agreement 2000-2008 2001-2008 archive 2005 UK election 2008 American Presidential Election 2008 US presidential elections 404 60th anniversary of United Nations peacekeeping 9/11 9/11 academic essays 9/11 Essays 9/11 for High Schools 9/11 papers A variety of articles on disbility

Taggers

Adamcok Adviya Alex Alicohn Alonso Amy878 Bozwell Cehbeh Ckkern Clare Dij Donnie1 Emloff Eve Gautam Gemma Gigi131 Gkr365 Jay John Katy Kkrish Laura82 Liam Logic Louise Matt Maubry Maz Megan11 Meng Mike Mikeg Mouz Noel Palmexo Plemai Proximo Randvar Ricey Rkanani Shannon Squelci Tato Thei75 Tom Urs1981

All Alex's Tags

Academic Papers advocacy Afghan War, 2001- Armed Forces - Political activity Arms control Arms control - Europe art Asia for Educators bibliography british foreign policy Buddhism - China - Tibet buffer zone canada Carnegie Endowment casualities figures cease fire China - Politics and

My Tags For This Document

Disarmament and arms control human rights media podcast Sino-American Summits Sino-US specalists Taiwan - 20th century Tiananmen Square United States - Foreign relations - China

Document Details

Title	1998 Sino-Soviet Summit by Robert S. Ross
URL	http://www.asiasociety.org/publications/sino_american_summit.html
Description	This site contains the full-text of the paper by Robert Ross which was originally published by the Asia Society, a non-profit educational institution, in June 1998. It examines US-China relations in the period after the Tiananmen Square massacre in 1989, focusing on such issues as human rights, international trade, nuclear weapons and arms control. Consideration is then turned to what the 1998 Sino-Soviet Summit hopes to achieve in these areas. An additional feature of the article is an appendix listing contact addresses for specialists on US-China relations.

Automatically suggested matches
Find appropriate context(s)

Soviet Union - Foreign relations
African American women legislators-biography, .
Eastern Europe-politics, . . .
Kommunisticheskaia partiia Sovetskogo Soiuza
Marxism-Leninism
Relation of federal to state, regional, provincial g
United States-foreign relations-Eastern Europe, .
------- End of Political Science -------
1945-1998

Explore hierarchy around the selected context

International relations
Soviet Union - Foreign relations

Select/edit relevant tags

Eastern Europe - foreign relations Europe, Eastern - Foreign relations - Soviet Union Europe, Eastern - Politics and government - 1945-1989 Foreign relations - Eastern Europe Foreign relations - Russia Foreign relations - Soviet Union International relations Russia - foreign relations Russia - Foreign relations - 1894-1917 Russia (Federation) - Foreign relations Russia (Federation) - Foreign relations -

Offering users to choose from KOS

- The importance of controlled vocabulary suggestions for indexing and retrieval:
  - To help produce ideas of which tags to use
  - To make it easier to find focus for the tagging
  - To ensure consistency
  - To increase the number of access points in retrieval

  - However, the value and usefulness of the suggestions proved to be dependent on the quality of the suggestions, both as to conceptual relevance to the user and as to appropriateness of the terminology

(Golub, Lykke, Tudhope 2014)

Automatic indexing…

3 major approaches

- Text categorization
- Document clustering
- String matching

…Automatic indexing…

Automatic indexing beneficial

- Address the scale and sustainability
- Enrich bibliographic records
- Establish more connections across resources

Reported success of automated tools

- Entirely replace manual indexing to machine-aided indexing (MAI)
  - MAI example: NLM´s Medical Text Indexer

MTI
NLM Medical Text Indexer
Providing Indexing Assistance
Since 2002
Biomedical Literature
MTI
MTIFL
MeSH Suggestions

…Automatic indexing

Evaluation problem (Golub et al. 2016)

- Research comparing automatic versus manual indexing is flawed (Lancaster 2003, p. 334)
  - Out of context, laboratory conditions
  - Few reports on indexing tools in operating information systems

Challenge A: relevance 1/2

- Purpose of indexing: making relevant documents retrievable
- Relevance
  - A complex phenomenon
    - Many possible document-query relationships
    - E.g., for children/scientists, query/information need/task…
  - Subjective
  - Multidimensional and dynamic (Borlund 2003)

Challenge A: relevance 2/2

In practice, evaluation of IR is based on pre-existing relevance assessments

- Initiated by Cranfield tests
- A gold standard
  - A test collection consisting of a set of documents
  - A set of 'topics'
  - A set of relevance assessments

- "*In spite of the dynamic and multidimensional nature of relevance, in practice evaluation of information retrieval systems has been reduced to comparison against the gold standard—a set of pre-existing relevance judgments which are taken out of context. An early study on retrieval conducted by Gull in 1956 powerfully influenced the selection of a method for obtaining relevance judgments. Gull reported that two groups of judges could not agree on relevance judgments. Since then it has become common practice to not use more than a single judge or a single object for establishing a gold standard.*"

(Saracevic 2008, 774)

Challenge B: indexing

Aboutness

- Dependent on factors like interest, task, purpose, knowledge…

Exhaustivity and specificity of indexing

- Related to indexing policies at hand
- A subject correctly assigned in a high-exhaustivity system may be erroneous in a low-exhaustivity system

Terms assigned automatically but not manually might be wrong or they might be right but missed by manual indexing

→ Not good to use just the existing classes as the gold standard

METHODOLOGY

Methodology …

- Based on Golub et al. (2016)
- Data collection: a subset of SMDB, SND, SwePub

- A comparison of assigned terms against a carefully crafted 'gold standard' and in the context of actual information retrieval

  - The '**gold standard**' developed through input of
    - Professional catalogue librarians
    - End users who are experts in the subject at hand,
    - End users who are inexperienced in the subject
    - Several automated subject indexing software applications

  - **Information retrieval** will involve end users conducting actual searching on the indexed collection of resources and marking how relevant each retrieved resource is

… Methodology

- Automated subject index terms derived from several algorithms will be compared against the ‘gold standard’ and in the retrieval test

- The analysis will also include looking at what caused the retrieval of the document at hand: a cataloguer’s term, subject expert’s term, inexperienced user’s term or an automated term

- Also log analysis and questionnaires to help contextualize the results

Automated indexers

- To be built/adjusted for DDC, SAO and the Swedish language

- String matching
- Machine learning
- Commercial: Data Harmony (rule-based)

SIGNIFICANCE

Significance…

The value of the professional, the automated and the end-user ways of creating subject index terms will be determined

→ Allowing for informed decisions on ensuring high quality subject access points as part of the Swedish library and information infrastructure

– Cheap assignment of controlled subject terms useful at various stages of the metadata creation workflow:

- By an author creating original index terms at the time of deposit;
- By a reader annotating (for colleagues/world or for recommendation for inclusion in a collection);
- By a cataloguer

…Significance

- Provision of (semi)-automated solutions for assigning DDC will enable:
  - Hierarchical browsing by subject
  - Retrievability of Swedish resources in multilingual systems
  - Integration of Swedish resources into the Semantic Web (as DDC is available as Linked Data)

- The resulting empirically tested comprehensive methodology framework will be of interest to other researchers

Partners

Linnaeus University
Lund University
University of South Wales

University of Aalborg
University of Buffalo
Charles Sturt University

Swedish National Library
Linnaeus University Library
Access Innovations
OCLC

References

Borlund, P. (2003). The concept of relevance in IR. *Journal of the American Society for Information Science and Technology 54(10)*, 913-925.

Golub, K., Dagobert, S., Buchanan, G., Tudhope, D., Lykke, M. & Hiom, D. (2016). A framework for evaluating automatic indexing or classification in the context of retrieval. *Journal of the Association for Information Science and Technology, 67(1),* 3-16.

Golub, K., Lykke, M. & Tudhope, Douglas (2014). Enhancing social tagging with automated keywords from the Dewey Decimal Classification. *Journal of Documentation, 70(5)*, 801-828.

Hjørland, B. (2002). Epistemology and the socio-cognitive perspective in information science. *Journal of the American Society for Information Science and Technology 53(4),* 257-270.

Lancaster, F. W. (2003). Indexing and abstracting in theory and practice. 3rd ed. Champaign: University of Illinois.

Saracevic, T. (2008). Effects of inconsistent relevance judgments on information retrieval test results: A historical perspective. *Library Trends 56(4)*, 763-783.

Soergel, D. (1994). Indexing and retrieval performance: The logical evidence. *Journal of the American Society for Information Science 45(8),* 589-599.