Automatic Classification Using DDC on the Swedish Union Catalogue

Koraljka Golub, Johan Hagelbäck, Anders Ardö

19th European NKOS Workshop, 23rd TPDL
Oslo, 12 September 2019

Linnæus University

Contents

Purpose and aims

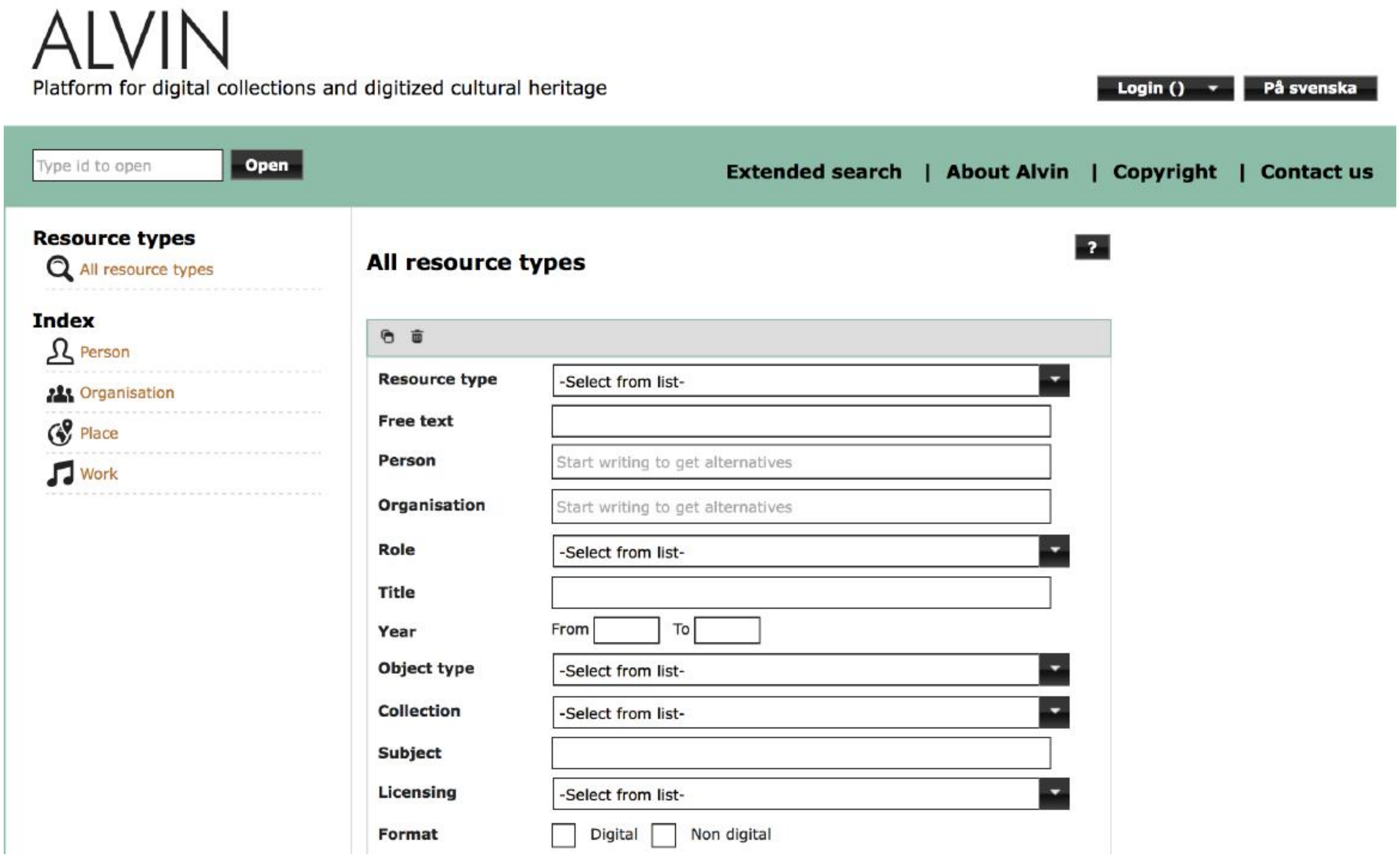
- To establish the value of automatically produced classes for Swedish digital collections

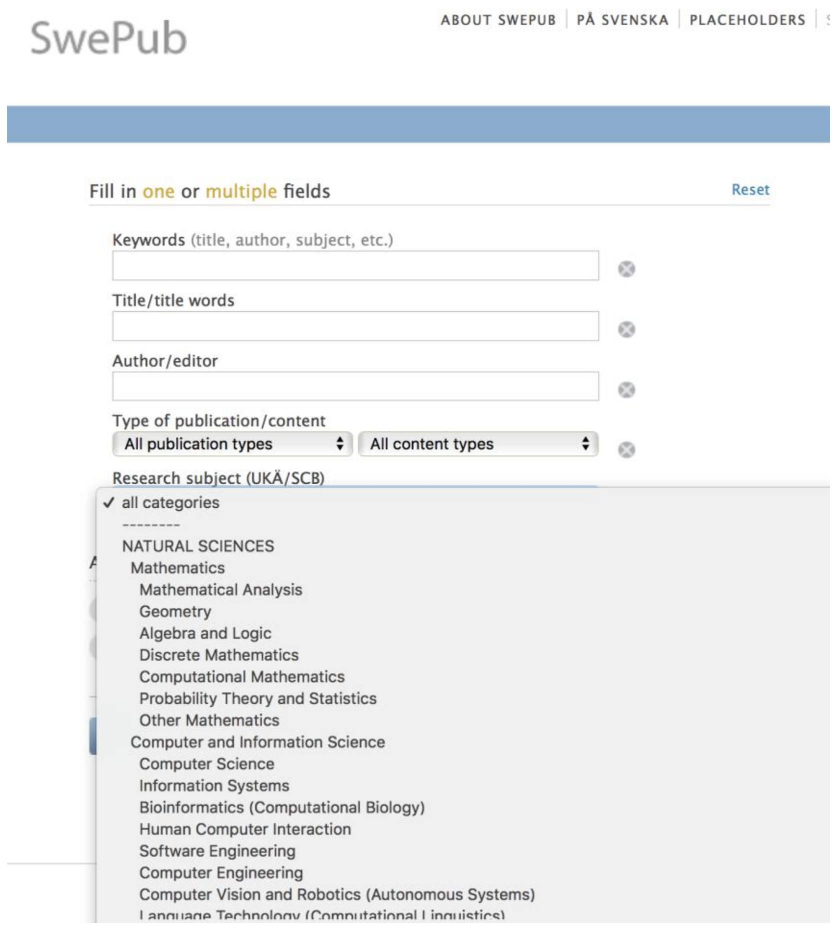
- Aims
  - Develop (and evaluate) automatic subject classification for Swedish textual resources from the Swedish union catalogue (LIBRIS)
    - http://libris.kb.se

  - Data set: 143,756 catalogue records containing DDC in LIBRIS
  - Using a machine learning approach
    - Multinomial Naïve Bayes (NB)
    - Support Vector Machine with linear kernel (SVM)

Rationale…

- Lack of subject classes and index terms from KOS in new digital collections

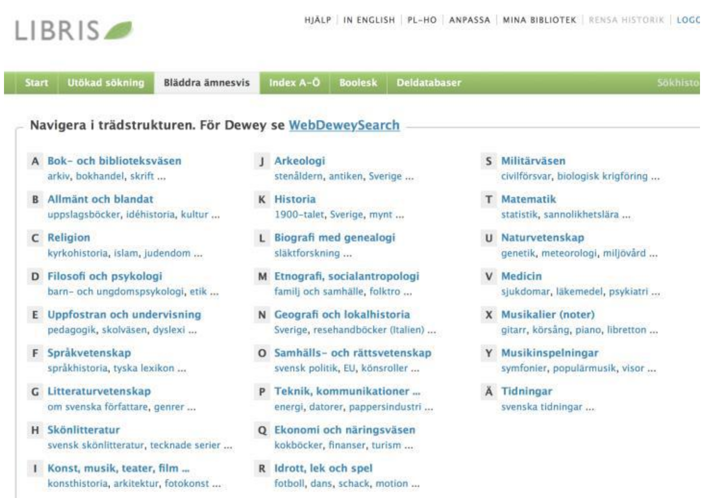

ALVIN
Platform for digital collections and digitized cultural heritage
Login ()
På svenska
Type id to open
Open
Extended search | About Alvin | Copyright | Contact us
Resource types
All resource types
Index
Person
Organisation
Place
Work
All resource types
?
Resource type
-Select from list-
Free text
Person
Start writing to get alternatives
Organisation
Start writing to get alternatives
Role
-Select from list-
Title
Year
From
To
Object type
-Select from list-
Collection
-Select from list-
Subject
Licensing
-Select from list-
Format
Digital
Non digital


SwePub
ABOUT SWEPUB | PÅ SVENSKA | PLACEHOLDERS
Fill in one or multiple fields
Reset
Keywords (title, author, subject, etc.)
Title/title words
Author/editor
Type of publication/content
All publication types
All content types
Research subject (UKÄ/SCB)
✓ all categories

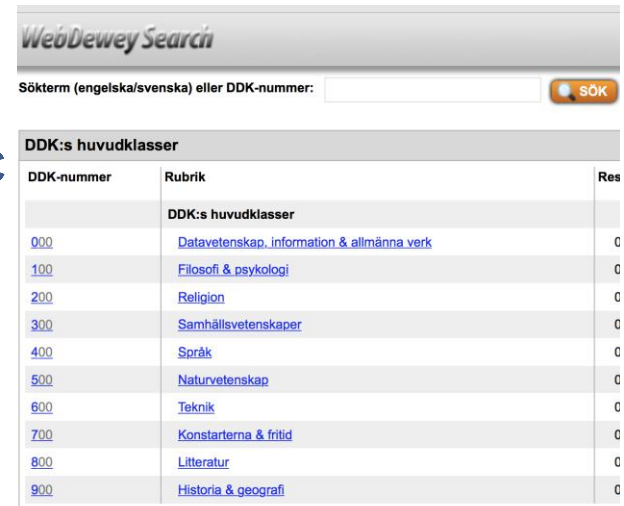
NATURAL SCIENCES
Mathematics
Mathematical Analysis
Geometry
Algebra and Logic
Discrete Mathematics
Computational Mathematics
Probability Theory and Statistics
Other Mathematics
Computer and Information Science
Computer Science
Information Systems
Bioinformatics (Computational Biology)
Human Computer Interaction
Software Engineering
Computer Engineering
Computer Vision and Robotics (Autonomous Systems)

… Rationale

- DDC chosen as a new national ‘standard’ in 2013



SAB → DDC



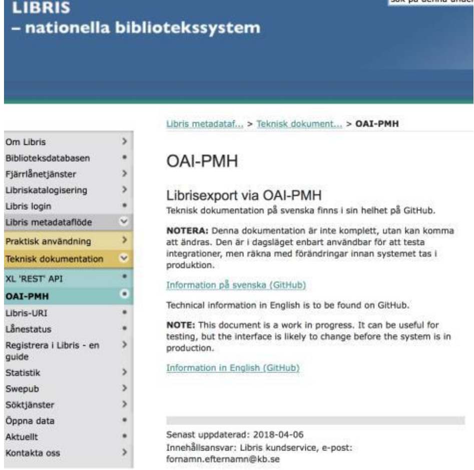
- LIBRIS has a large collection of resources with DDC assigned to Swedish resources to train on
- Explore automatic classification on Swedish DDC → interoperability, cross-search, multilingual, international…

Contents

DDC

- 23rd edition, MARCXML format
- 128 MB → relevant info extracted into MySQL database, total of 14,413 classes

  - Class number (field 153, subfield a);
  - Heading (field 153, subfield j);
  - Relative index term (persons 700, corporates 710, meetings 711, uniform title 730, chronological 748, topical 750, geographic 751; with subfields);
  - Notes for disambiguation: class elsewhere and see references (253 with subfields);
  - Scope notes on usage for further disambiguation (680 with subfields); and,
  - Notes to classes that are not related but mistakenly considered to be so (353 with subfields).



Data collection

- LIBRIS: 143,838 catalogue records in April 2018
  - Using OAIPMH protocol, MARCXML format
  - All LIBRIS records with 082 MARC field for DDC class
  - Relevant info extracted into MySQL:
    - Control number (MARC field 001), unique record identification number;
    - Dewey Decimal Classification number (MARC field 082, subfield a);
    - Title statement (MARC field 245, subfield a for main title and subfield b for subtitle); and,
    - Keywords (a group of MARC fields starting with 6*), where available -- 85.8% of records had at least one keyword.
  - DDC classes truncated to 3-digit codes, to maximise training quality

Training problem: imbalance between classes

- The most frequent class is 839 (Other Germanic literatures) with 18,909 records
- In total 594 classes have less than 100 records (70 of those have only 1 single record)

→ A dataset called “major classes” containing only classes with at least 1,000 records:

- 72,937 records spread over 29 classes

  (60,641 records spread over 29 classes when selecting records with keywords)

The different datasets generated from the raw LIBRIS data

Dataset	ID	Records	Classes
Titles	T	143,838	816
Titles and keywords	T_KW	121,505	802
Keywords only	KW	121,505	802
Titles, major classes	T_MC	72,937	29
Titles and keywords, major classes	T_KW_MC	60,641	29
Keywords only, major classes	KW_MC	60,641	29

Classifiers

- Pre-processing
  - Bag-of-words approach (stop-words retained) → over 130,000 unique words
  - Unigrams and 2-grams
  - TF-IDF scores

- Multinomial Naïve Bayes (NB) and Support Vector Machine with linear kernel (SVM) algorithms
  - Both have been used in text classification numerous times with good results
  - SVM typically better results than NB, but slower to train
  - NB can be trained incrementally, i.e. new training examples can be added without having to retrain the model with all training data

Evaluation measure

- Accuracy
- Amount of correctly classified examples

$$\text{Accuracy} = \frac{\text{Correctly classified examples}}{\text{Total number of examples}} \%$$

Contents

Major results

- SVM better than NB on all classes
  - On test set, best result **81.4%** accuracy for classes with over 1,000 training examples, or **58.1%** accuracy for all classes
    - When using **both titles and keywords**, unigrams and 2-grams
- Features
  - Number of training examples significantly influences performance
  - Keywords better than titles, keywords + titles best
  - Stemming only marginally improves results

NB

Dataset	Accuracy, unigrams		Accuracy, unigrams + 2-grams	
	Training set	Test set	Training set	Test set
T	83.54%	34.89%	95.82%	34.15%
T_KW	90.01%	55.33%	98.14%	55.45%
KW	75.28%	59.15%	84.95%	58.11%
T_MC	90.83%	54.21%	98.63%	50.51%
T_KW_MC	95.42%	76.52%	99.66%	75.96%
KW_MC	86.94%	77.25%	94.24%	77.09%

SVM

Dataset	Accuracy, unigrams		Accuracy, unigrams + 2-grams	
	Training set	Test set	Training set	Test set
T	93.74%	40.91%	99.59%	40.45%
T_KW	97.50%	65.25%	99.90%	66.13%
KW	83.09%	64.02%	92.38%	64.09%
T_MC	93.95%	57.99%	99.62%	57.80%
T_KW_MC	97.89%	80.75%	99.93%	81.37%
KW_MC	90.58%	79.56%	96.30%	80.38%

Top two levels, all examples from all classes

- Accuracy increased from 58.1% (three digits, 802 classes) to 73.3% (two digits, 99 classes)

Input data:	Title + subtitle + keywords									
			Naive Bayes		Naive Bayes (ngram=1,2)		Linear SVC		Linear SVC (ngram=1,2)	
Dataset	Examples	Categories	Training set	Test set	Training set	Test set	Training set	Test set	Training set	Test set
T_KW_stm_2D	121505	99	87,40%	65,64%	93,18%	67,79%	90,60%	72,68%	96,23%	73,32%
T_KW_2D	121505	99	88,26%	64,78%	93,55%	66,92%	91,21%	72,14%	95,48%	73,24%

Input data:	Keywords only									
			Naive Bayes		Naive Bayes (ngram=1,2)		Linear SVC		Linear SVC (ngram=1,2)	
Dataset	Examples	Categories	Training set	Test set	Training set	Test set	Training set	Test set	Training set	Test set
KW_2D	121505	99	78,36%	68,12%	82,53%	67,94%	81,75%	71,86%	86,18%	71,96%

Stopwords and less frequent words

- For major classes

- Removed stopwords (_sw) → reduced accuracy in most cases
- Removed less frequent words from the bag-of-words (_rem) → increased accuracy from 81.8% to 82.2%

Input data:	Title + subtitle + keywords, remove less frequent words									
			Naïve Bayes		Naïve Bayes (ngram=1,2)		Linear SVC		Linear SVC (ngram=1,2)	
Dataset	Examples	Categories	Training set	Test set	Training set	Test set	Training set	Test set	Training set	Test set
T_KW_MC	60641	29	95,42%	76,52%	99,66%	75,96%	97,89%	80,75%	99,93%	81,37%
T_KW_MC_rem	60641	29	90,17%	76,79%	93,25%	78,21%	92,51%	80,94%	95,02%	81,83%
T_KW_MC_stm	60641	29	94,32%	76,36%	99,59%	76,36%	97,21%	81,07%	99,91%	81,80%
T_KW_MC_stm_rem	60641	29	89,62%	76,26%	92,95%	78,27%	92,18%	81,34%	94,89%	82,20%
T_KW_MC_sw	60641	29	95,50%	76,46%	99,64%	76,62%	95,44%	80,98%	98,48%	81,24%
T_KW_MC_sw_rem	60641	29	90,28%	77,09%	92,33%	78,60%	92,46%	81,04%	94,30%	82,13%
T_KW_MC_sw_stm	60641	29	94,49%	76,59%	99,53%	76,95%	94,87%	81,40%	98,72%	81,24%
T_KW_MC_sw_stm_rem	60641	29	89,79%	76,36%	91,96%	78,90%	92,17%	81,54%	94,16%	81,90%

Word embeddings

- Word embeddings combined with different types of neural networks:
  - Simple linear network (Linear)
  - Standard neural network (NN)
  - 1D convolutional neural network (ConvNet)
  - Recurrent neural network (RNN)

Input data:	Keras embedding, 128 fts									
			NN		ConvNet		Linear		RNN	
Dataset	Examples	Categories	Training set	Test set	Training set	Test set	Training set	Test set	Training set	Test set
T_KW_MC	60641	29	96,19%	79,40%	95,33%	79,92%	97,17%	79,99%	92,76%	78,70%
KW_MC	60641	29	90,54%	78,23%	90,39%	79,15%	91,30%	78,41%	88,03%	78,74%
T_KW_MC_stm	60641	29	95,92%	79,57%	94,60%	80,38%	96,90%	80,81%	92,38%	79,16%

- Worse results than NB/SVM, but very close (80.8% compared to 82.2%)
  - Advantage of word embeddings is having a smaller representation size (then the stored data takes less space)

Common misclassifications

- Whole dataset:
  - Class 3xx (Social sciences, sociology & anthropology)
    - Other classes often misclassified as belonging to 3xx
    - 3xx often misclassified as other classes
    - Most misclassifications between 3xx and 6xx (Technology)

- Major classes dataset:
  - Fiction – mostly based on language and country
    - 823 (English fiction) misclassified as 839 (Other Germanic literatures)
    - 813 (American fiction in English) misclassified as 823 and 839
  - 306 (Culture and institutions) misclassified as 305 (Groups of people)

820 English & Old English literatures
821 English poetry
822 English drama
823 English fiction
824 English essays
825 English speeches
826 English letters
827 English humor and satire
828 English miscellaneous writings

Contents

Try improve algorithm performance…

- More training examples
  - Through linked open data and URIs from elsewhere?
  - Include records with SAO and LCSH without DDC, and through the files with mappings of SAO and LCSH to DDC, try use them as training documents?
  - Norwegian / other catalogues in DDC?

…Try improve algorithm performance…

- Take advantage of DDC

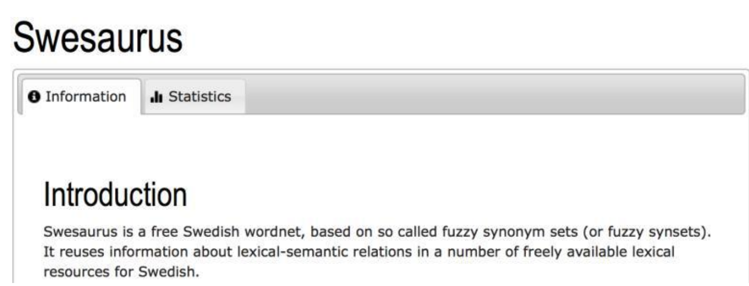
    - Class number (field 153, subfield a);
    - Heading (field 153, subfield j);
    - Relative index term (persons 700, corporates 710, meetings 711, uniform title 730, chronological 748, topical 750, geographic 751; with subfields);
    - Notes for disambiguation: class elsewhere and see references (253 with subfields);
    - Scope notes on usage for further disambiguation (680 with subfields); and,
    - Notes to classes that are not related but mistakenly considered to be so (353 with subfields).

    - Establish how these contribute to classification accuracy

…Try improve algorithm performance

- Evaluate ensemble learners combining different types of algorithms
    - String matching in the lack of training examples
        - Maui software http://www.medelyan.com/software
        - Scorpion approach https://www.oclc.org/research/activities/scorpion.html
        - Enrich with Swesaurus for more mappings and disambiguation

https://spraakbanken.gu.se/resource/swesaurus



Evaluation

- Test for all levels of classes
- Test with algorithms outputting more than one class
- Include misses in evaluation using measures like F-measure combining precision and recall
- Manual evaluation to identify causes for successes and failures
- Evaluate in the context of retrieval in real IR tasks

New forum for automatic indexing / classification

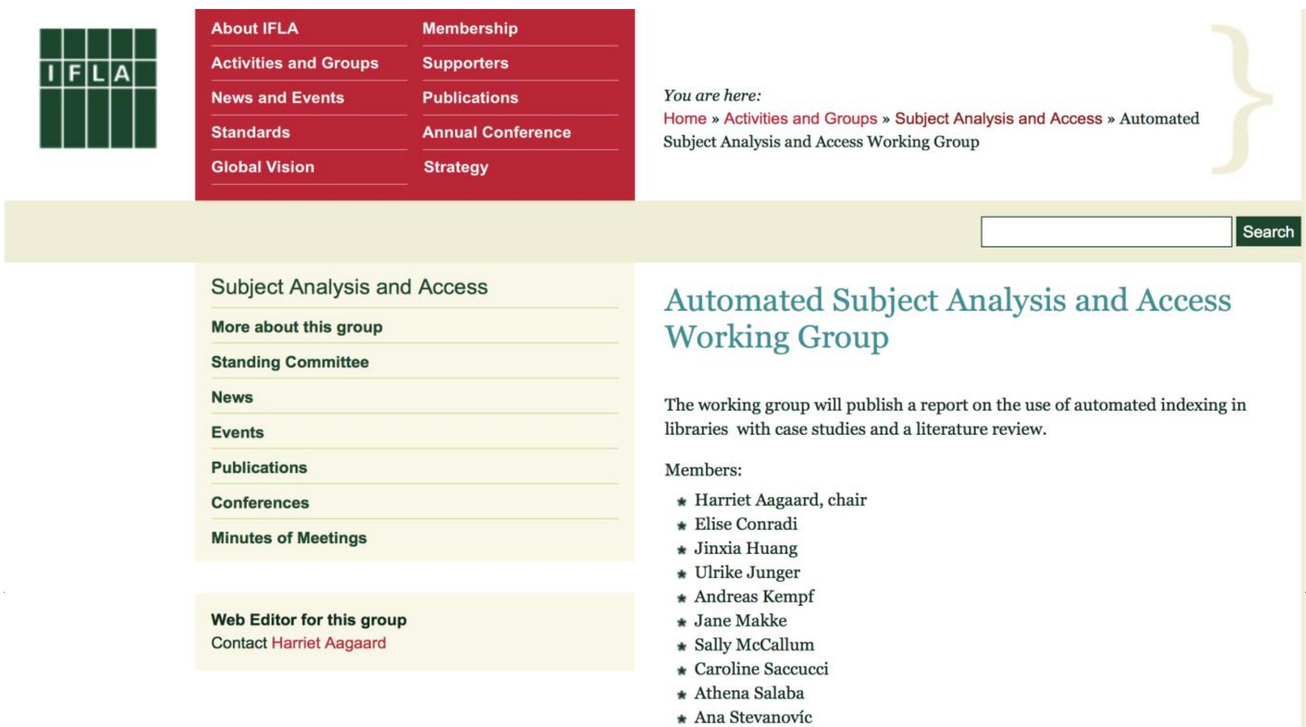
- DCMI Automated Subject Indexing IG

http://www.dublincore.org/groups/automated_subject_indexing_ig/

  - Open to all
  - Place where we could collaborate?
  - Create open source solutions?
    - Annif (http://annif.org)

New IFLA WG

- https://www.ifla.org/subject-analysis-and-access
  - Automated Subject Analysis and Access Working Group
    - https://www.ifla.org/node/92551



Thank you for your attention!

- Questions? Feedback?

- What does the practice want to see?
    - For which applications: Web Archives, repositories, CH collections, cross-search…?

- Contact: koraljka.golub@lnu.se