Decentralized subject indexing: a case from television

BBK colloquium, IBI Humboldt

Nils Pharo

16.10.2018

OSLO METROPOLITAN UNIVERSITY
STORBYUNIVERSITETET

Program

- OsloMet ABI
- Decentralised indexing
- Method
- Findings

Oslo Metropolitan University

- Previously Høgskolen i Oslo og Akershus
- Engineering, health, education, journalism, social work, economics, LIS
- 20 000 students, 2200 employees (1400 academic staff members)
- Bachelor, master and PhD programmes

ABI – Institutt for arkiv-, bibliotek- og informasjonsfag

- Established 1940
- Bachelor programmes in Library and Information Science (130 students/year) and Archival Science (40 students/year)
- Master and PhD programmes in LIS
- Co-organize iConference 2020 (in Borås, Sweden)

Research at ABI

- Three research groups (LittKult, InfoSam, MetaInfo)
- MetaInfo – Metadatabased information systems
  - Research on information behaviour, system interaction, metadata generation and utilization, metadata modelling, automatic knowledge extraction and system infrastructures.
  - Several projects, including TORCH – aim: support the conversion of the program description archives of the Norwegian Broadcasting Corporation (NRK)

Decentralised indexing

- NRK changed their indexing practice
  - Previously: metadata experts performed descriptive and content indexing of all TV programs
  - From 2012: production staff (content experts) performed indexing, supervised by experts from archive & research department

Vocabulary

- Controlled vocabulary → 2012
- 2012: semicontrolled vocabulary. Suggested terms, but possible to overrule. Post-indexing control

Research question

- Master thesis by Veslemøy Søbak → JASIST article
- *What consequences does decentralized indexing have on the indexing vocabulary?*
  - RQ1: What characteristics of television programs are indexed?
  - RQ2: How does decentralized indexing influence indexing consistency?
  - RQ3: How do indexers' practice and motivation affect indexing quality?

Some tables and illustrations from Søbak, V., & Pharo, N. (2017). Decentralized subject indexing of television programs: The effects of using a semicontrolled indexing language. *Journal of the Association for Information Science and Technology*, *68*(3), 739-749.

Indexing policy

- Tags should cover the “who,” “what,” and “where” of all indexed program parts. “When” is included when necessary.
- Tags are written in bokmål (one of the two standard forms of the Norwegian language).
- Single terms and concepts should be used, not sentences.
- Use common terms and concepts.
- Well-known names on incidents should be used.
- Abbreviations are used if these are better known.
- Common synonyms should be added.
- Terms must be precise, but more general terms can be added as well.
- Ambiguous terms should not be used.
- Tags should always be chosen with caution.

Method

- One week of TV programs (October 2012) – 249 programs (823 parts)
- Content analysis of all “tags” used
- Interviews with five production staff members

Overview of programs and tags

Overview	8 oct	9 oct	10 oct	11 oct	12 oct	13 oct	14 oct	Total
No of programs	33	43	39	38	35	31	30	249
Reruns	2	5	5	5	3	17	12	49
Regional news	2	2	2	2	1	0	0	9
Programs w/o tags	21	23	25	22	21	10	13	135
Analyzed programs	8	13	7	9	10	4	5	56

Distribution of tags in program parts

Overview	8 oct	9 oct	10 oct	11 oct	12 oct	13 oct	14 oct	Total
No of program parts	124	193	145	143	147	35	36	823
Duplicates	0	4	1	0	0	0	0	5
No of parts w/o tags	61	91	83	85	94	11	22	447
Analyzed program parts	63	98	61	58	53	24	14	371
No of tags	277	488	284	314	263	126	76	1828

RQ1: What characteristics of television programs are indexed?
(facetted scheme)

- Name (person, item, organization)
- Time (year period)
- Place (city, county, municipality, area, landscape, room)
- NRK jargon
- Form (genre, format)
- Language
- Subject (product, activity, role, case, condition, animal, attribute, phenomenon…)

Facet distribution

- Subjects 61.5 %
- Names 28.5 %
- Place 8 %
- Form, Time, Language, NRK jargon 1 % each

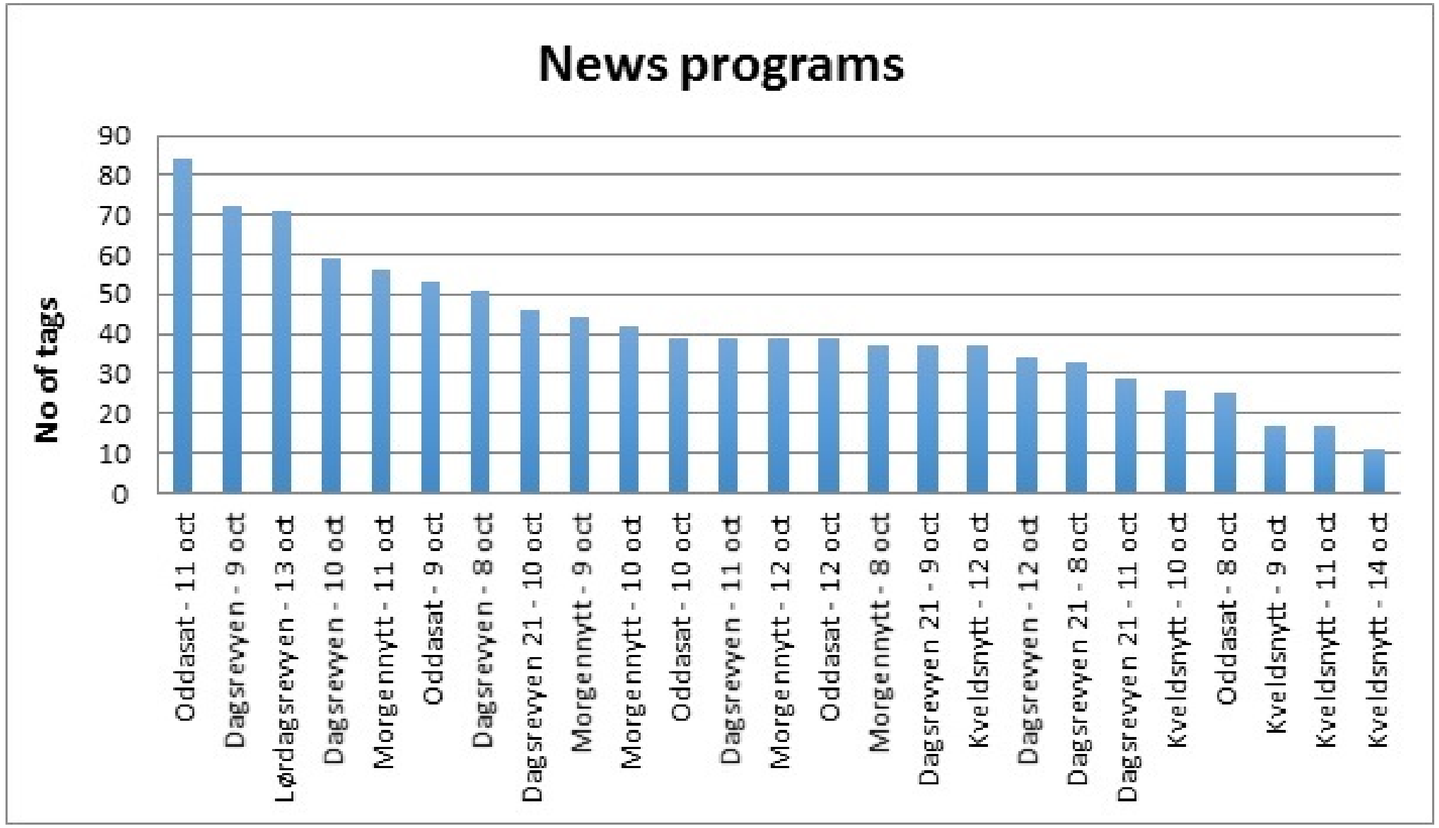
RQ2: How does decentralized indexing influence consistency?

- Consistency in the selection of terms
- More specifically we analysed consistency with respect to exhaustivity – operationalized as number of tags per program

Tag categories in editorial offices

Editorial office	Subject	Place	Name	Time	Form	NRK jargon	No. cat.	Total
News	60.9 (461)	12.3 (93)	23.5 (178)	0.9 (7)	0.8 (6)	0.3 (2)	1.3 (10)	100 (757)
Culture and entertainment	67.1 (367)	5.9 (32)	18.3 (100)	1.3 (7)	4 (22)	2.2 (12)	1.3 (7)	100 (547)
Sport	36.7 (104)	1.8 (5)	59.3 (168)	1.8 (5)	—	—	0.4 (1)	100 (283)
Documentaries and fact	77.7 (150)	7.8 (15)	13 (25)	—	1 (2)	0.5 (1)	—	100 (193)
Health, consumer and life style	83.9 (26)	6.5 (2)	—	—	—	9.7 (3)	—	100 (31)
Children	100 (17)	—	—	—	—	—	—	100 (17)
Total	61.5 (1,125)	8 (147)	24.8 (471)	1 (19)	1.6 (30)	1 (18)	1 (18)	100 (1,828)

News programs



Tag frequency

Frequency	No of tags	Total freq	Percentage
1	1015	1015	55.5 %
2	206	412	22.5 %
3	68	204	11.2 %
4	9	36	2.0 %
5	5	25	1.4 %
6	7	42	2.3 %
7	2	14	0.8 %
8	1	8	0.4 %
9	3	27	1.5 %
10	2	20	1.1 %
11	1	11	0.6 %
14	1	14	0.8 %
Total		1828	100.0 %

Other consistency issues

- Spelling errors
- Foreign language (not bokmål)
- Grammatical form (plural singular)
- Use of sentences
- Capital letters

RQ3: How do indexers’ practice and motivation affect indexing quality?

- Five informants (from ½ year to 15 years of experience, 2 daily, 1 weekly 2 irregularly indexing)
- No metadata background
- Four directly involved in TV production, one in editorial office
- Two of them coordinate indexing

Interviews

- A lot of time spent ”nagging” journalists to index
- Tagging is considered a duty “forced upon them”
- Time spent indexing: from 5 to 45 minutes per program
- The list of tags is used to varying degree: “use it actively” “find it irritating”
- Tags were targeted for the external audience

Conclusion

- Facet distribution similar to previous work on folksonomy analysis; subject being the dominant facet. Sport programs have an overrepresentation of name tags
- Tag consistency: only 28 % of the programs are indexed. Very varying indexing exhaustivity. Many tags only used once. Terminological and indexing political errors.

Limitations

- Only one week of data
- Data from early stage under new regime
- Not compared with indexing practice by metadata experts