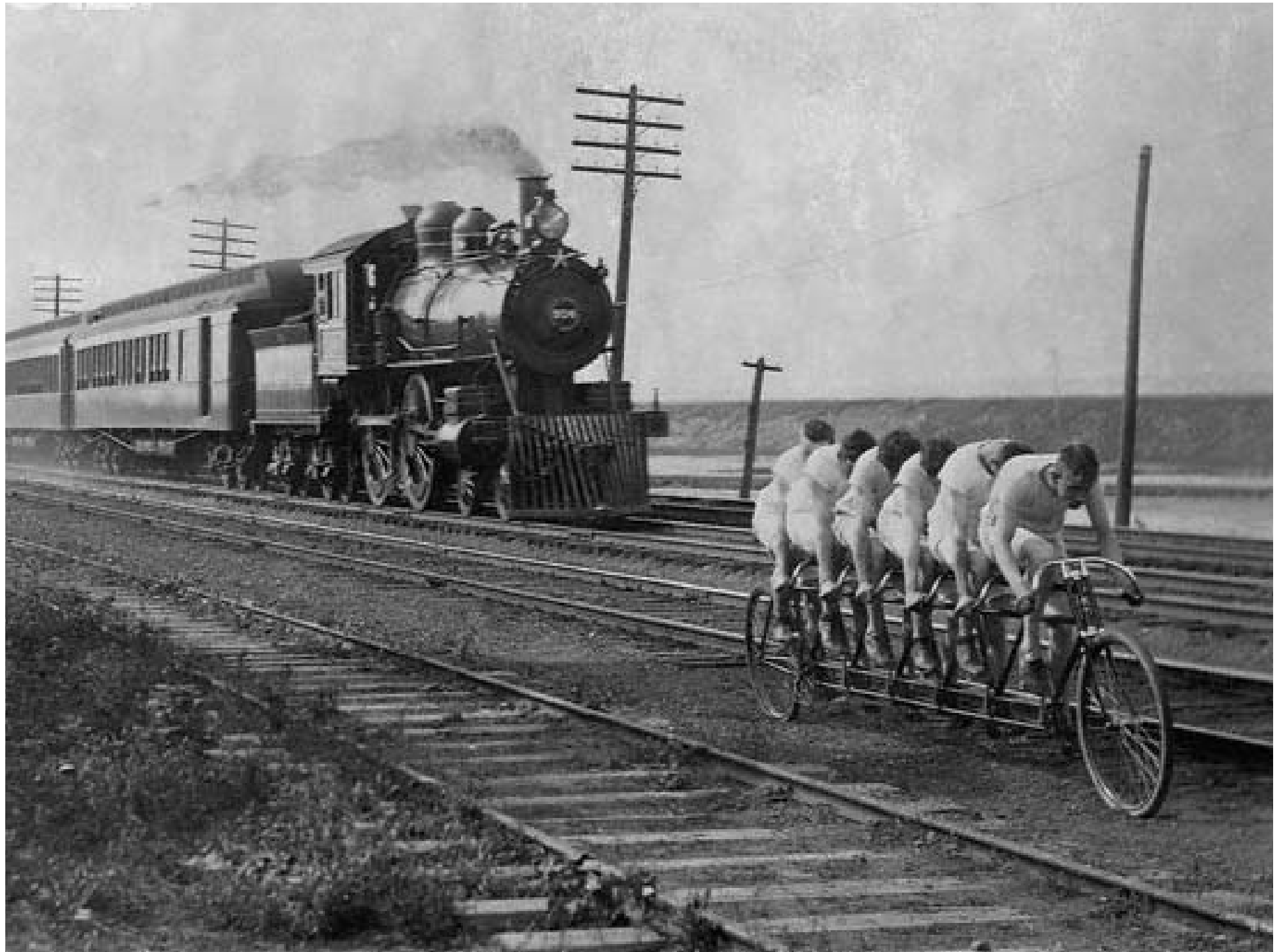




Desperately Seeking House Elves

Problems from Computational
Folkloristics

Timothy R. Tangherlini
U.C.L.A.





A story...

It was the old counselor from Skaarupgaard who came riding with four headless horses to Todbjærg church. He always drove out of the northern gate, and there by the gate was a stall, they could never keep that stall door closed.

They had a farmhand who closed it once after it had sprung open. But one night, after he'd gone to bed, something came after the farmhand and it lifted his bed straight up to the rafters and crushed him quite hard. Then the farmhand shouted and asked them to stop lifting him up there. "No, you've tormented us, but now you'll die..."

I heard that's how two farmhands were crushed to death. He wanted to close the door and then they never tried to close it again.



Some meta-data

- Told by Margrete Jensdatter
 - Weaver woman
 - 1813-1902
- Told
 - in 1889
 - in Mejlby
- Mentions
 - Todbjærg (Todbjerg)
 - Skaarupgaard (Skårupgård)
 - The old counselor
- Classified as
 - DS IV: Manor lords, ladies and mistresses

What can I do with this story?

- Some standard questions:
 - ☐ Role of ghosts in late 19th century Denmark?
 - ☐ Origins of the story?
 - ☐ Structure of the story?
 - ☐ Who, what, where of this story?
- With all these great questions, is there really a need for computational folkloristics?





How do I find a story?

How do I find similar stories?

- Collections and archives
 - Largely printed or hand written
 - Problems of access
- Collection indices
 - Largely idiosyncratic
 - Domain specific
- Type indices
 - Not uniformly applied / applicable
 - Emic – Etic categories
 - One text – one classifier

Evald Tang Kristensen Collection

- Evald Tang Kristensen (1843-1929)
 - Actively collected from 1865-1923
- Stories collected
 - from ~4500 informants
 - in ~2200 places (mostly in Jutland)
 - mentioning ~12,000 places (also mostly in Jutland)
- 24,000 manuscript pages
 - Covering ~210 fieldtrips
 - 6,000 additional loose leaf pages
 - 2,000 additional pocket diary pages
- 80 published volumes
 - ~100,000 entries
 - 12 separate indices





Connections...

- In 2000, none of this data was digitized
 - Now ~80% of the data is digitized
- Most of this data is “unconnected”
 - Indices to individual collections:
 - Informants and stories
 - Stories and places mentioned
 - Stories by topics
 - Topics are idiosyncratic and not consistently applied
 - Each published collection has its own index
 - Inconsistent orthography
 - Incomplete
 - Published collections are organized largely by genre (Danske sagn, Folkeæventyr fra Jylland, Gamle folks fortællinger om det jyske almueliv, etc)

E. F. Mac
J 5, I
Hans Mad
Hans Krist
Lærer Jen.

Koldi
lund,
282, 4
1743,
1904,
2173,
1892,
360, 4
VII 5
1196,

J. Madsen,
Lærer J. P.
Elev Jens I
Klavs Kris
Klavs Pete

Kristen M
2309.
Lærer L. M
Lars Krist
VI 77

Mads Kris
Marie Mai
Mikkel Ma
Morten M

1700,
Cand. the
præst
Niels Mad

riksh.
IV 14
697, 8
Niels Mad
1270.

Peder Mac
Lærer P. I
Lærer P.
mand

173,
415,
204,
240,
113,
547,

Vejstrupgård [1099] II A 42.
Vejstrup, Sønderj. [2393]
IV 909.
Vejvad, Vinding v. Holstebro
[2519] I 1128.
Velberstoft, Ty [1379] I 896.
Lille-Velbølle, Vedsted, Sønderj.
[3046] V 160.

Vellerup, Sønderj. [3074]
II I 378, 406.

Vellev v. Ulstrup [1835]
III 273, 498, IV 956, 1084,
V 1821.

Velling v. Ringkjøbing [2533]
I 536, III 179.

St.-Velling, Smedstrup v. Vejle
[2364] II E 60, IV 2309,
V 389, 1904, VII 1291.

V.-Velling v. Randers [1779]
II G 349, J 19, III 808,
903-905, IV 967.

Ø.-Velling, Randers [1782]
I 50, 1360.

Ø.-Velling skov [1782] III 1996.
Vellinggård, Bræstrup v. Frede-
ricia [2354] VII 1283.

Vellingshøj v. Hjerring
[1202] VI 778.

Vellumgård, Håsum, Salling
[1684] III 1319.

Vemmelev v. Slagelse [436]
I 10, 209, 839, II B 31, E 49,
F 82, IV 1132, VI 1210.

Vemmetofte skov v. Fakse [544]
III 98.

Veng v. Skanderborg [2161]
II A 152, H 263, III 528,
V 29, 305.

Veng sø [2161] III 528.

Vengegård [2161] III 983.

Venning, Alum v. Randers
[1764] I 1144, VI 945.

Venning bakker [1764]
III 106, 1175.

Vennergård, Velling v. Ring-
kjøbing [2533] V 1258, 1322,
1594.

Verninge v. Tommerup [875]

Vesterborg Sø. III 1123, 1124, 1125,
1154, 1155.

Vesterby, Darum S. VI 486.
Vesterbæk. II H 305.

Vesterbølle S., Rinds H. I 339, (913?),
IV 461, 462, 540, 606, 624, 689.

Vesterengene. II H 489.
Vestergaard, Ajstrup S. II I 190.

Vestergaard, Fjerritslev, Kollerup S. II
A 55.

Vestergaard, Søndbjærg S. II B 153.
Vestervig. II G 280, III 1279, IV 1073,
1074, 1075, 1076.

Vester-Klit. II A 81, IV 1028, 1029, 1030,
1031, 1032, 1033, 1034, 1035, V 227,
228.

Vesterris. II I 292.
Vestervig S., Refs H. I (120), 417, (796),
931, II B 123, 183, G 279, I 235, III

530, 557, 609, 958, IV 330, 449, 555,
736, 816, V 143, 411, 497, 498, 551,
711, 1135, VI 340, 642, 653, 654.

Vestero S., Læso. I 379, 473, 626, II A
88, B 125.

Vestarp. II H 468.
Vetterslev S., Ringsted H. I 64.

Vevlinge S., Skovby H. I 701, II H 10,
71, V 275, 620, 621.

Viborg Amt. III 1103.
Viborg. II G 255, H 425, I 201, Tillæg
17, III 134, 160, 287, 522, 560, 1044,
1045, 1222, IV 787, V 257, 1000.

Vibstrup Vej. V 568.

Viby S., Bjærg H. II H 34, 152, I 207,
III 697, 1228, IV 109, 110, 732, V 61,
VI 716.

Viby S., Ning H. III 1345, VI 36.
Viby, Lille-, Drigstrup S. IV 182.

Vibæk. IV 348, VI 799.
Viffertsholm, Gl.-. VI 317.

Vigerslev S., Skovby H. I 701, II C 18,
62, H 71, III 600, V 91, 971.

Vigerslev, Hvidovre S. V 191.
Vihojgaard. I 305.

Vilbjærg S., Hammerum H. IV 500.
Vildmose, Lille-. III 1192.

Vildmose, Store-. II E 71.
Vile S., Harre H. I 23, 83, 217, 413,
420, 670, II C 70, H 474, I 189, 288.

Villands H., Skaane. I 142, 320, 358,
382, 383, 384, 385, 386, 400, 507, 515,
571, 617, 644, 689, 850, 945, II D 31.

Villegården. III 1180.

Vennergård, Velling v. Ringkjø-
bing. Tb. II 125, 259.

Verninge v. Odense. Tb. VI 1201.
Vestbirk v. Horsens. Tb. VI 773.

Vestbjærg, Sulsted, Vendsyssel.
Tb. VI 123.

Vesterbæk, Sønderjylland. Tb.
III 1842.

Vesterskov, Almtoft v. Kjellerup.
H. IV 244.

Vestervig, Ty. H. II 58, V 278,
768, VI 133, 272. Tb. IV 361,
VI 230, 1111, 1140, 1198, 1203.

Vestervig kirke. H. II 286.
Vestervig kloster. H. II 58, 286

—289, VI 318. Tb. II 131, V 347.
Vestkær, Ølgod. Tb. V 473.

Vestrup, Albæk v. Randers. Tb.
III 273.

Vestrupgård, Saltum v. Åbybro.
H. III 45.

Vevlinge v. Bogense. Tb. III
1714.

Vibholm, Tim, Hing-h. H. VI
650.

Viborg. H. I 361, II 116, 303,
III 71, IV 394, V 19, 69, 229,
279, 358, 446, 476, VI 383, 410,
481, 493, 917, 924, 925. Tb. I

83, 1306, 1742, II 207, III 58,
IV 50, V 6, 35, 47, 68, 87, 261,
702, 727, 893.

Viborg-huset, Viborg. H. III 246.
Tb. V 313, VI 935.

Viborg krat. H. I 357.
Viborg-egn. H. VI 632. Tb. VI
175.

Viby (Lille-), Drigstrup v. Kjærte-
minde. Tb. V 262.

Vidstrup v. Hjerring. H. I 323.
Viffertsholm skov, Solbjærg v. Bæ-
lum. Tb. V 39.

Vildbjærg v. Herning. Tb. VI 877,
886.

Vildmose (Store-), Vendsyssel. H.
I 320. Tb. I 1417.

Vile, Salling. H. III 259. Tb. III
504, IV 169, V 170.

Vilhelmsborg, Mårslet v. Århus.
Tb. II 5.

Villendrup, Halling v. Hadsten.
Tb. VI 865.

Villerslev, Ty. H. V 205, VI 702.
Tb. III 513.

Villestrup, Astrup, Himmerland.
H. I 338, II 335. Tb. II 102,
244, 258, V 371, 468, VI 143,
144.

Vilslev v. Ribe. Tb. II 43, III 31,
65, 113, V 136, 139, 140, VI
598.

Vilslev mose. Tb. V 137.

Vilsted v. Logstor. H. I 25, II
323, III 93, 300, IV 43, 303, VI
651.

Vilsted mølle. Tb. II 255.

Vilstedgård, Vilsted. H. I 100.

Vind v. Holstebro. H. V 380. Tb.
I 1631, III 41, V 270.

Vind kirke. Tb. II 23.

Vindblæs v. Hadsund. H. IV 13,
VI 267.

Vindblæs, Himmerland. H. VI
585. Tb. VI 651.

Vinde v. Skive. H. V 702, 708.

Vindelev v. Vejle. Tb. VI 1305.

Vindelsbæk, Elsborg v. Rødkjær.
H. V 69.

Vinderslev v. Kjellerup. H. I 107,
316. Tb. IV 222, 288. Tb. VI
473.

Vinderslev hede. Tb. VI 137.

Vinderslevgård. H. II 319, V 68.

Vinding v. Holstebro. H. V 714.

Tb. I 88, 1630, 1631, V 49, 214,
VI 402, 470.

Vinding v. Bræstrup. H. IV 358.

Tb. III 508.

Vinding v. Nyborg. H. III 380.

Vinding v. Roskilde. H. VI 134.

Vinding v. Vejle. H. VI 292. Tb.
IV 372, V 180.

Vinding skov. H. I 365.

Vindum v. Bjerringbro. H. I 141,
VI 37, 643, 729.

Vindum hede. H. III 26.

Vindum skov. H. I 228, VI 95,
921.

Vindum-overgård. H. II 106, 112,
113, 293, III 132. Tb. I 1330.

Vinge (Nørre-), Sønderlyng-h. H.
II 420, V 158. Tb. VI 727.

Vingegård, N.-Vinge. Tb. VI 233.

Vingsted mølle, Bræsten v. Vejle.
Tb. VI 167.

Vinkel (Sønder-) v. Viborg. H.
I 13, 36, 129, 227, 228. II 106,
112, 113, 282, 352, 353, III 221,



Archival finding aids

- Archival finding aids are hand-written
 - The master index at the Danish folklore archive is entirely on hand-written cards
 - Cannot be easily OCR'ed
- There is no clear or consistent geographic coding of these materials

Several index cards...

Mikkelsen, Nils Mikael, Sognefoged Lyngby
 Ad. 404 b - o & b. (Top. 1249).

f. 22/ 1 1838.

Strandfogedga

(17/ 2 1844

Via

den overgaar

selv 1925

DS. III 6

6104,

M. O. T.

Ericksen, Kristense, Plet. (Top. 2433), aflyglskme.

(f. 30/11/1897 x 1393? Hurup, kom som 19-årig til at tjene
 i gård i Lyngum. G. med Ole Peter Mogensen fra
 Plet. - modtil Maren Olesdatter, Lyngum, se denne).
 of Johanne Marie Olesdatter
 Kristense (død 1893)

Da. 5708 b - 10a. 2191? Alv. 1497; 1504 - ~~08~~ ~~1508~~ 1508.

Ref. III 2: 174, 269.

J F. XI (1890), 18a, 20a, 21, 22B, 28a, 40a, 43, 46F, 47.
 J F. V 776. 49c, 57, 59B, 64a, 85.

2191, (1890)

† Mo. III 303, 306 J F XI, 342. J F VI 308.



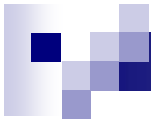
Memoirs as resource

- Tang Kristensen's memoir, *Minder og Oplevelser* acts as a 4 volume "Rosetta Stone" to the archive
 - Has no index
 - Was not machine readable
- Solution:
 - We OCR'ed and hand corrected the 1200 page memoir
 - We added tags for personal names and place names
 - We added tags for end and start points of field trips
 - Currently:
 - aligning personal names with informant index
 - aligning place names with gazetteer
 - extracting field trip itineraries and aligning with field diary pages



Goal: Use all the information

- Develop biographical information about storytellers
 - Gender
 - Education and Occupation (from MO)
 - Places of residence (from MO)
 - Family information (from MO, Church records and Census records)
 - Dates of birth, confirmation, death (from MO, Church records and Census records)
 - Economic information (from MO, fire insurance protocols and probate records)
- Connect storytellers to places
- Connect storytellers to stories
- Connect stories to places
- Connect storytellers to fieldtrips
 - Gives a date for collection
 - Gives a range of field diary pages for each storyteller
 - Provides an informant index for the field diaries ➔ Connect storytellers to field diary pages
- Connect stories to field diary pages
- These relations allow us to generate adjacency matrices very quickly based on all or subsets of the data



So that...

- given a particular research question, one can
 - rapidly find and retrieve resources that can help answer that question
 - visually browse the archive on multiple criteria
 - drill-down to underlying resources when one finds something of interest
- existing barriers to the type of research questions that can be posed the archive are removed



Problems, problems, problems

Non-standard resources

- Scanning, OCR, chunking
 - These are not terribly interesting
- Disambiguation / deduplication
- Geocoding challenges
- Connecting everything together
 - Editing / adding material



Disambiguation / Deduplication

- The problem of disambiguation is considerable
 - Duplication of informants and place names across collection indices
 - Informant and Place names with variant spellings
 - Multiple places with same name
 - Multiple people with same name

Microsoft Excel

File Edit View Insert Format Tools Data Window Help AbleBits.com

Type a question for help

A1 dupe_index

published_indices.xls:1

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
2208	3017-3018	Kristiane Larsdatter	Kristiane	Larsdatter						0	Rærup ved Rærup Nørre-Nebel	Rærup Nørre-Nebel	2598	55.796	8.357003	I 210 V 692 1197 VII 145 H I	
2209	2197-2196	Thomas Jensen	Thomas	Jensen	Husmand					0	Henne ved Henne Nørre-Nebel	Henne Nørre-Nebel	2599	55.73561	8.217435	III 2072 V 1288 VII 474 6 H I	
2210	193	Lærer Andersen		Andersen	Lærer					0	Lønne ved Lønne ved Varde	Lønne ved Varde	2601	55.80348	8.237117	III 2066 VII 1457	
2211	4157-4156	Kristen Olesen	Kristen	Olesen						0	Al ved Varde	Al ved Varde	2602	56.49679	8.809487	III 677 IV 647 1639 V 172 H I	
2212	4938	Jes Jørgen Schultz	Jes Jørgen	Schultz						0	Markskjæl Markskjælbæk Al ved Varde	Markskjælbæk Al ved Varde	2602	56.49679	8.809487	IV 1775	
2213	4076-4075	Hans Nielsen Nørgård	Hans Nielsen	Nørgård						0	Sig ved Va Sig Varde	Sig Varde	2611	55.66314	8.564768	IV 1005 V 777 VII 578	H I
2214	5100-5101	Ebbe Alexandersen Smed	Ebbe Alexandersen	Smed		Født 1809 i Lønborg [2575]	MO II MO III 254	Torstrup ve	Torstrup ved Varde	Torstrup ved Varde			2611	55.66314	8.564768	I 1333 III 1299 1389 1783 H I	
2215	1660	Frederik Horsbøl	Frederik	Horsbøl						Yderik	Torstrup	Yderik Torstrup	2611	55.66314	8.564768		H I
2216	1662-1661	Horsbøl Frederik og Niels	Frederik og Niels	Horsbøl	Gårdmand					0	Yderik Tor	Yderik Torstrup ved Varde	2611	55.66314	8.564768	III 458 1300 16 IV 508	H I
2217	1366-1365	Iver Hansen	Iver	Hansen						0	Hulvig ved Hulvig ved Varde	Hulvig ved Varde	2613	55.69758	8.648629	I 79 IV 1484	H I
2218	219-218	Nikoline Kristine Andersen	Nikoline Kristine	Andersen	Møllerkone	Født 1835 se: Letbæk mølle	MO III 250	Letbæk m	Letbæk mølle ved Varde	Letbæk mølle ved Varde			2614	55.71879	8.604901	II H 527 IV 2040 2155 223 H I	
2219	2852	Lærer Kristensen		Kristensen	Lærer			0	Tirstrup ve	Tirstrup ved Varde			2614	55.71879	8.604901	II H 271	
2220	5477	Johanne Marie Thomasdatter	Johanne Marie	Thomasdatter		Huskone MO III 250	MO III 250	Tirstrup ve	Tirstrup ved Varde	Tirstrup ved Varde			2614	55.71879	8.604901	I 719 1248 II E 158 IV 669 V	
2221	5592-5593	Niels Kristensen Tønning	Niels Kristensen	Tønning		MO III 250	MO III 250	Tirstrup ve	Tirstrup ved Varde	Tirstrup ved Varde			2614	55.71879	8.604901	IV 1562 V 1469 VI 495 V H I	
2222	1333-1332	Hansen Ane Margrete	Ane Margrete	Hansen		Født 1854 Havlund	MO III 249 I	Havlund Ø	Havlund Ølgod	Havlund Ølgod			2615	55.80775	8.617866	I 323 983 II J I 1920	H I
2223	3172-3173	Jørgen Andreas Lavrsen	Jørgen Andreas	Lavrsen	Husmand	Skrædder husmand MO III 249	MO III 249	Havlund Ø	Havlund Ølgod	Havlund Ølgod			2615	55.80775	8.617866	I 40 II B 162 J 111 III 419 H I	
2224	4998-4997	P Sindberg	P	Sindberg	Lærer			0	Ølgod [261	Ølgod			2615	55.80775	8.617866	III 813	H I
2225	5089-5088	Frands Nielsen Slot	Frands Nielsen	Slot				0	Ølgod [261	Ølgod			2615	55.80775	8.617866	V 216	H I
2226	5698-5699	Mette Marie Vig	Mette Marie	Vig				0	Ølgod [261	Ølgod			2615	55.80775	8.617866	I 1254 II G 185 I 131 III 24 H I	
2227	446	A Birch	A	Birch	Stationsforstander				Ølgod	Ølgod			2615	55.80775	8.617866		
2228	1029	Frifelt Salomon	Salomon	Frifelt		Forfatter			Ølgod	Ølgod			2615	55.80775	8.617866		II Tillæg 8
2229	1343	Frederik Hansen	Frederik	Hansen	Maler				Ølgod	Ølgod			2615	55.80775	8.617866		H I
2230	5598	Torlund Mads	Mads	Torlund					Ølgod	Ølgod			2615	55.80775	8.617866		III 690 V 281 6
2231	444	Den k...											2620	55.6484	8.60735		III 645 V 1497

published_indices.xls:2

	A	B	C	D	E	F	G	H	I	J
17	0	J P Åbo	J P	Åbo	Lærer					
18	0	Abraham Kirsten	Abraham					Engesvang ved Silke Engesvang Moselu		1
19	0	Abrahamsen Fru	Abrahamsen					Em Vendsyssel [12 Em		1
20	2	Karl Abrahamse Karl	Abrahamsen					Mølholm Vejle		
21	2	Karl Abrahamse Karl	Abrahamsen					Kildebrønde Sjælland		
22	2	Lærer M Abrah Lærer M	Abrahamsen		Født 1/9 11			0 Ukendt [0000] Ukendt		
23	0	Peder Jørgensen Peder Jørg	Åby			Manuskript		0 Kildebrønde Sjæll [Kildebrønde Sjæll		
24	0	Achthon	Achthon							
25	0	Adamsdatter Maren	Adamsdat	ter				Magleby Ellidshøj		
26	0	Adersen Karl	Karl	Adersen				Lydersholm		
27	1	Anders Agård	Anders	Agård	Husmand			Fjølstervang Vorgod		
28	1	Anders Agård	Anders	Agård		Født o 1820 MO III 228		Fjølstervang ved He Fjølstervang ved He		2
29	1	Agerskov L	L	Agerskov		Lærer		Lem ved Ringebøing		

published_indices.xls:3

	A	B	C	D	E	F	G	H	I	J
1	dupe_index	Full Name	First Names	Last Name	Occupation	DS_Notes	DSnr_notes	JA_notes	MO_refs	Collection Place
2	1	"Den rette Ordinants"		"Den rette Ordinants"						
3	2	Aabo J P	J P	Aabo		Lærer				
4	3	Aagaard Hedevig Kirstine	Hedevig Kirstine	Aagaard						
5	4	Aakjær Jeppe	Jeppe	Aakjær		(se ogsaa Jensen Jeppe Aakjær)				
6	5	Aale Stephan Olufsen	Stephan Olufsen	Aale						
7	6	Aas		Aas		Lærer				
8	7	Aas		Aas						
9	8	Aase R P	R P	Aase		Musiker				

Ready

NUM

start

Inbox - Microsoft...

4 Microsoft Offi...

Clipboard01 - Trf...

published_indice...

published_indice...

published_indices...

BookDrive Editor ...

Removable Disk (...)

EN

11:15 PM



Geocoding challenges

- Place names in the collection:
 - when indexed use an idiosyncratic topographical index used only by the Danish Folklore Archive (Skjelborg 1967)
- The topographic index does not easily align with existing gazetteers
 - Orthographic differences
 - Different names for same place
- Historical organization no longer in use / cannot rely on contemporary resources
 - Current Kommune organization different from Amt/Herred/Sogn divisions of the 19th century
 - Boundaries have shifted dramatically
- Existing Danish gazetteers use Easting/Northing
 - Requires conversion for use with most web-based mapping

Microsoft Excel

File Edit View Insert Format Tools Data Window Help AbleBits.com

Type a question for help

K1 Unnormalized

df1_gazetteer_reconciled.xls:1

	A	B	C	D	E	F	G
2548	15858	Tjæreghaver	6114254	573175	55.16948365	10.14879592	
2549	8304	Tjæreby	6137240	640691	55.36127931	11.2195616	Dupe
2550	8106	Tjæreby	6124520	650149	55.24427097	11.36183632	Dupe
2551	7504	Tjæreby	6068922	689370	54.73178805	11.94121179	Dupe
2552	4670	Tjæreby	6138581	687212	55.35775702	11.95342557	Dupe
2553	8784	Tjæreby	6165852	695603	55.59916875	12.1047843	Dupe
2554	9262	Tjæreby	6203450	701520	55.93404362	12.22630388	Dupe
2555	18925	Tjæreborgvad	6146900	474577	55.46759112	8.597879158	
2556	21261	Tjæreborg Nørremark	6148003	473354	55.47743704	8.578429533	
2557	20217	Enge	6144513	473651	55.44609373	8.583458941	
2558	12412	Tjæreborg	6146615	473590	55.46497798	8.582295104	
2559	5932	Tjærebybæk	6259240	567251	56.4728251	10.09168775	
2560	9707	Tjæreby	6259055	568359	56.47100385	10.10962169	
2561	5783	Tjærebybæk	6244367	552608	56.34107841	9.851034256	Dupe

Disambiguated / Lost_provinces / Topo_index-lookup / Danish_100k / Convert / Sheet1 / Sheet3 /

df1_gazetteer_reconciled.xls:3

	A	B	C	D	E	F	G	H	I	J	K	L
1	Topograp	Place na	Resolve: Y	Z	Lat	Long	Gaz_Dup	Dupe: Checked	Unnormalized			
2	1	Danmark	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	Danmark			
3	2	Sjælland	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	Sjælland			
4	3	København	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A				
4	4	Frederiks borg	København	6176746	723364	55.68488	12.55339	Dupe x				
5	5	Helsingør	Helsingør	6213648	722720	56.01611	12.57344					
6	6	Hillerød	Hillerød	6203490	706486	55.93229	12.30569					
7	7	Frederiks sund	Frederikss	6192287	692376	55.83765	12.07222					
8	10	Kronborg	Kronborg	6216328	725583	56.0388	12.62152	Dupe				
9	11	Tikøb	Tikøb	6213794	715114	56.02091	12.45177					
10	12	Hornbæk	Hornbæk	6221114	714870	56.08666	12.45374	Dupe				
11	13	Hellebæk	Hellebæk	6220056	720312	56.07469	12.54015	Dupe				
12	14	Asminderød	Asminderød	6207969	712881	55.96966	12.41141					
13	15	Grønholt	Grønholt	6205112	711333	55.94472	12.38441	Dupe				
14	16	Karlebo	Karlebo	6202035	713070	55.91636	12.40974					
15	17	Blovstrød	Blovstrød	6196656	711775	55.86868	12.38485					
16	18	Birkerød	Birkerød	6194135	714775	55.84474	12.43071					
17	19	Hørsholm	Hørsholm	6197935	718740	55.87703	12.49698					
18	20	Holbo	Holbo									
19	21	Esbønderup	Esbønderup									

df1_gazetteer_reconciled.xls:2

	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH
22619	Avlum	n	32	6236061	487760	55.26922	8.802368739	56.16	9.183652																						
22620	Idum	n	32	6243662	468530	55.33662	8.490976394	56.20	11.83633																						
22621	Narre Omme	n	32	6225956	470824	55.17769	8.530034018	56.10	39.63963																						
22622	Torsted	n	32	6228733	463911	56.20216	8.46696445	56.12	7.778951																						
22623	Holmsland Høj	n	32	6220824	449192	56.12977	8.182597453	56.7	47.17992																						
22624	Gl	n	32	6217668	450891	56.1016	8.210508883	56.6	5.748409																						
22625	Vorgod	n	32	6216017	481913	56.08893	8.702898	56.5	20.14455																						
22626	Borris	n	32	6201562	478183	55.95891	8.650557976	55.57	32.09128																						
22627	Oddum	n	32	6192258	475001	55.87516	8.600453989	55.52	30.59109																						
22628	Torstrup	n	32	6186875	472620	55.65314	8.564768031	55.39	47.30762																						
22629	Starup	n	32	6186965	487876	55.6494	8.807350079	55.38	54.22666																						
22630	Nykirke	n	32	6152049	473026	55.51407	8.667859407	55.30	50.847																						
22631	Sneum	n	32	6146476	476743	55.46389	8.522173867	55.27	50.00038																						
22632	Anst	n	32	6148680	514973	55.48402	8.236329088	55.29	2.460876																						
22633	Holsted	n	32	6152050	494846	55.5145	8.918381418	55.30	52.20245																						
22634	Vejrup	n	32	6152824	486537	55.5213	8.786763458	55.31	16.66991																						
22635	Astrup	n	32	6158603	489137	55.57329	8.82776318	55.34	23.84242																						
22636	Fanup	n	32	6135313	483040	55.36384	8.732442738	55.21	49.82492																						
22637	Melholt	n	32	6334763	585934	57.14906	10.42030589	57.8	53.02771																						
22638	Blåhøj	n	32	6192229	502177	55.87555	9.034734097	55.52	31.96941																						
22639	Tjebor	n	32	6280666	450751	56.70705	8.195587953	56.42	25.38377																						
22640	Studså	n	32	6216329	494342	56.09206	8.39063842	56.5	31.40243																						
22641	Esbønderup	n	32	6216787	708521	56.05068	12.3485145	56.3	2.432555																						
22642	Ude Sundby	n	32	6192235	632195	55.82818	12.06809635	55.50	17.44985																						
22643	Oppe Sundby	n	32	6190715	693656	55.82303	12.0915197	55.49	22.90068																						
22644	Avnslev	n	32	6136389	607996	55.36192	10.70372579	55.21	42.92287																						
22645	Skærbølev	n	32	6087900	612240	54.92544	10.7545954	54.55	31.60087																						
22646	Gerum	n	32	6362588	585270	57.39807	10.4189149	57.23	53.03561																						
22647	Ræer	n	32	6328315	480111	57.09775	8.671735723	57.5	51.90268																						
22648	Hjærmål	n	32	6323970	488577	57.05901	8.81862189	57.3	32.42223																						
22649	Haderslev	n	32	6122989	529708	55.2625	9.46735715	55.16	8.983306																						
22650	Starup	n	32	6121708	534370	55.24068	9.540538663	55.14	26.45486																						
22651	Bjerring	n	32	6129135	531600	55.3078	9.497810732	55.18	27.37327																						
22652	Tyrstrup	n	32	6134730	530647	55.35794	9.49340987	55.21	28.57832																						
22653	Jegerup	n	32	6126702	518852	55.281	9.296785629	55.16	51.61037																						
22654	Maugstrup	n	32	6127071	521823	55.28959	9.34363219	55.17	22.5151																						
22655	Slogs	n	32	6091317	508597	54.9687	9.13428974	54.58	7.331418																						
22656	Højst	n	32	6095144	502704	55.00316	9.042274109	55.0	11.73988																						
22657	Medelby	n	32	6073867	514777	54.8183	9.178581667	54.48	42.58815																						
22658	Tårnby	n	32	6142120	706685	55.38168	12.26272123	55.22	54.03531																						
22659	Dråby	n	32	6192228	687465	55.83905	11.99387683	55.50	20.59685																						
22660	Søby	n	32	6124579	565324	55.26335	10.02735254	55.15	48.05816																						
22661	Åsted	n	32	6367284	584449	57.44039	10.40685789	57.26	25.40312																						
22662	Hvidbjerg	n	32	6288953	478934	56.744	8.622284048	56.44	38.3915																						
22663	Ellstrup	n	32	6368900	559596	57.45885	9.933316171	57.27	31.84608		</																				



Some fixes

- Using NLP (Mallet and Lingpipe) to recognize names in unindexed resources
- Using ddupe to disambiguate names
 - See Getoor et. al. 2006
 - Working to use Geo-ddupe to speed process (Getoor et. al. 2007)
 - Find it [here](#)
- Figuring out processes to combine indices by using information about:
 - Places
 - People
- Relatively fast workflow / High accuracy combination of the indices



Adding more information

- Keyword indexing
 - Extract keywords from texts
 - Requires language expertise
 - Named Entity Detection
 - Indices are incomplete / additional place names and personal names
- Shallow ontology
 - Devise a two-level ontology for collection (Paneva, Rangochev, Luchev 2007; AFS Ethnographic Thesaurus 2007)
- Genre classification
 - Rapid classification based on Tang Kristensen's collection ordering
- Topic modeling (experimental)
 - Supervised learning – statistical classifiers
 - LSI and LDA



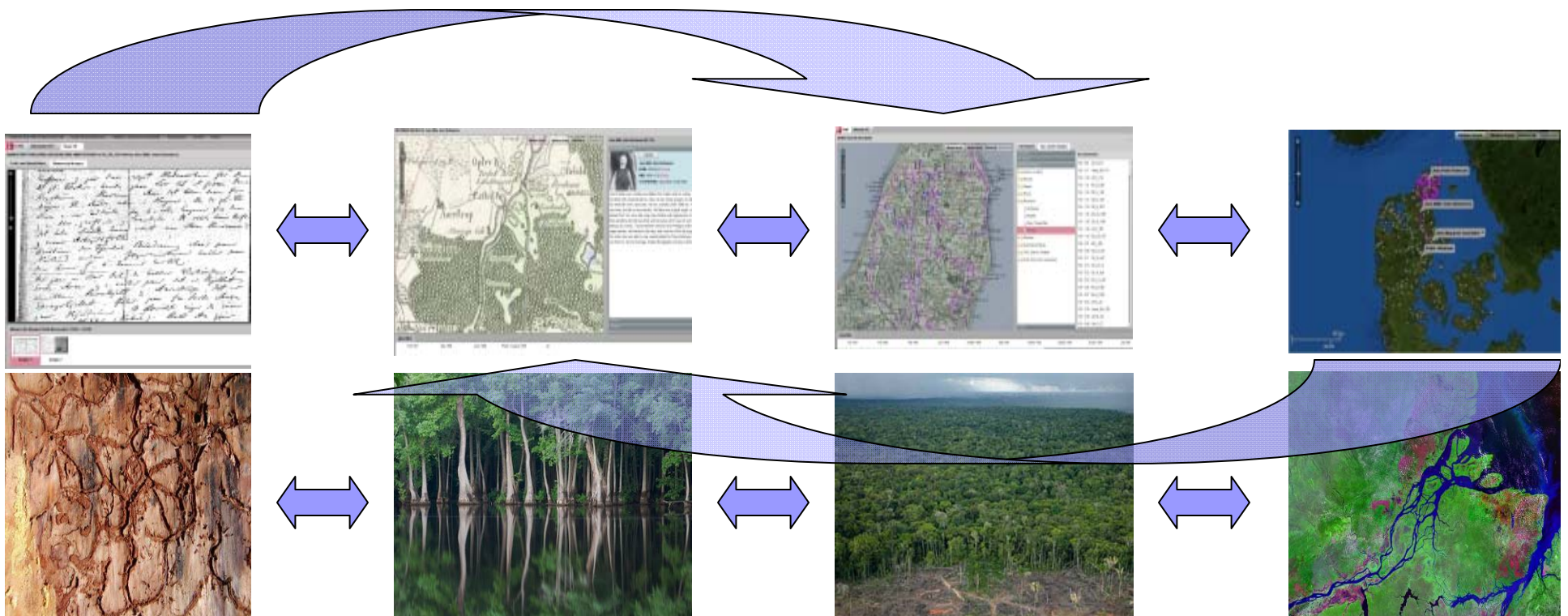
The Pilot Project:


Can one do this in a useful way?

- Five main informants
 - Told approximately 400 stories
 - Added an additional 500 “variant” stories
 - Project aims to
 - Present stories in rich manner
 - Allow for drill-down
 - Connect collection to maps
 - Contemporary
 - Satellite
 - Historical
 - Offer multiple paths for navigation through the collection
 - By person
 - By place
 - By story (keyword, topic, ontology, genre)
 - By time (fieldtrip)

Distant Reading to Micro Reading

- Danish folklore data nexus
- Allow for multiple views on the same archival resources





An (embedded) network model of Folklore

- Folklore, in this model,
 - emerges from the dialectic tension between the individual and tradition
 - and is expressed within and across social networks
- As such, it is also embedded in time and space



A very brief digression

- Rumor as a “hyperactive” transmissionary state of legend
 - Often does not include a strategy for dealing with threat
 - Pushes call for action out of narrative world into real world
- Rumors in social media
 - Hurricane Katrina
 - Political unrest and upheavals (with Roja Bandari):
 - Iran
 - Egypt
 - Tunisia
- “Rumor monger”
 - Rumor discovery on Twitter, Facebook, Blogs
 - Rumor / disinformation
 - Signal profiles in the network?
 - Burstiness



Using a network model

- For understanding connections
 - Between people, places and texts
- For discovering patterns that might be of interest given a particular research question
 - Homophily and similarity – text discovery
- For generating new research questions



How classifiers have failed us...

- Expensive
- Predicated on a particular research question
 - E.g. The Types of International Folk Tales
- One text ↔ One classifier
 - Frequent misclassification
 - Catch-all categories



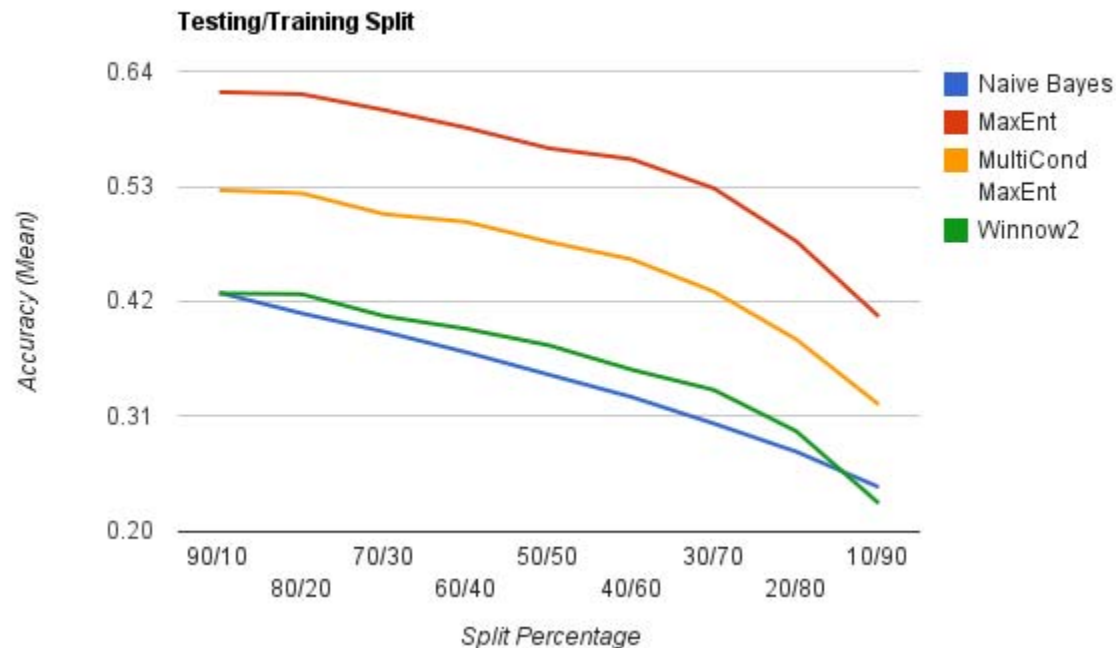
Some experiments with supervised learning

- NB and Max Entropy classifiers

- ☐ Used 12,000 texts from the *Danske Sagn* collection (approximately 1/9 of the entire collection)
- ☐ Dirty OCR
- ☐ Chunked into individual stories
- ☐ Only looking at text
 - Can we get the machine to classify texts in the same manner as Tang Kristensen did?

[illegible]

Some comparisons of classifiers and splits (training target)





So?

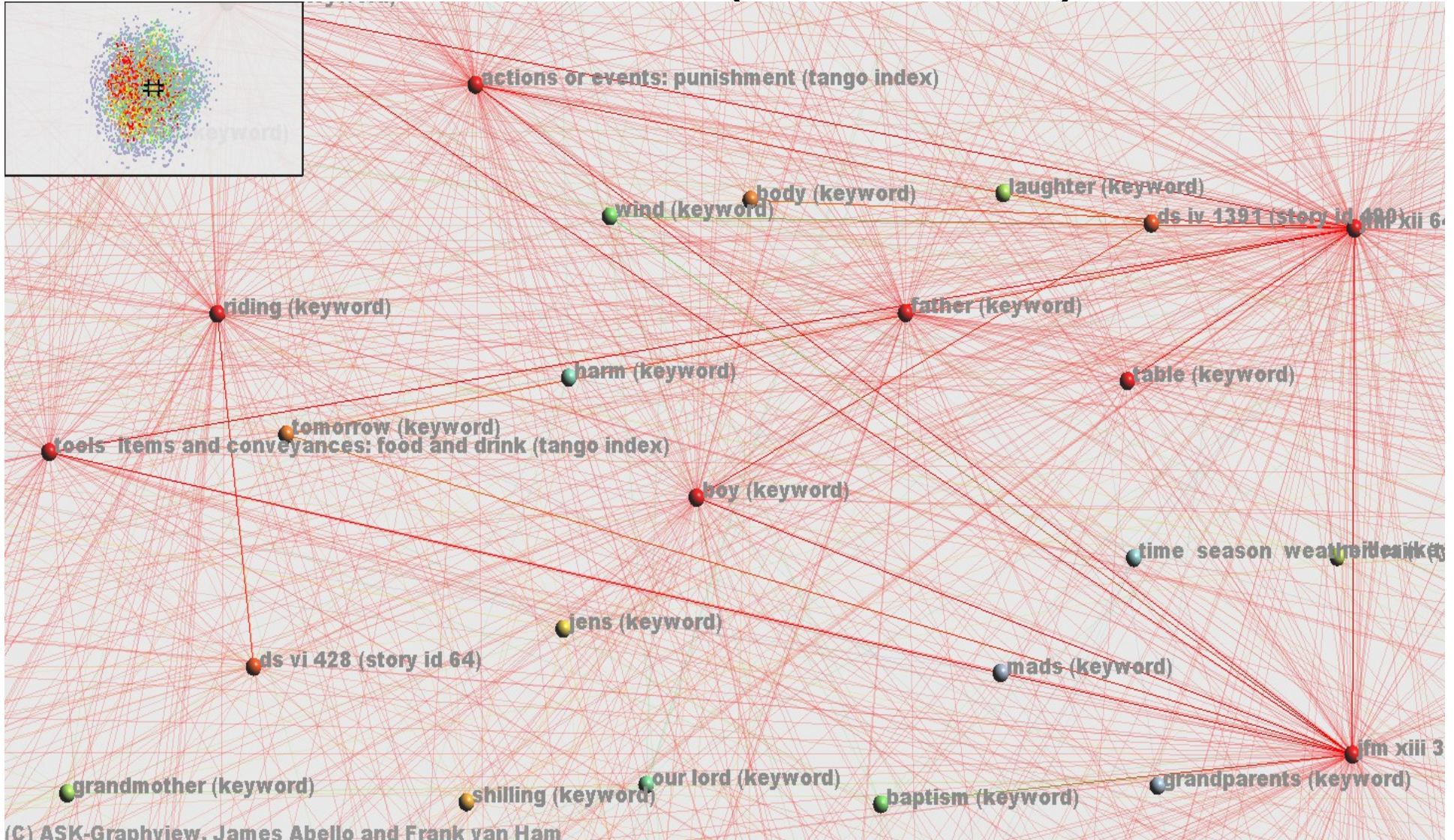
- Certain categories are easily recognized by the text classifier
- Some categories are “catch-all”
 - Hekse og deres idrætter
 - Witches and their sport
 - Forsjellige spøgeri og gjenfærds nedmaning
 - Different types of ghosts and their conjuring
 - Gjenfærd på forskjellige steder
 - Revenants in different places
 - Bjærgmænd
 - Mound dwellers
- Can I use the network model
 - to refine these classifications?
 - discover new areas of interest?



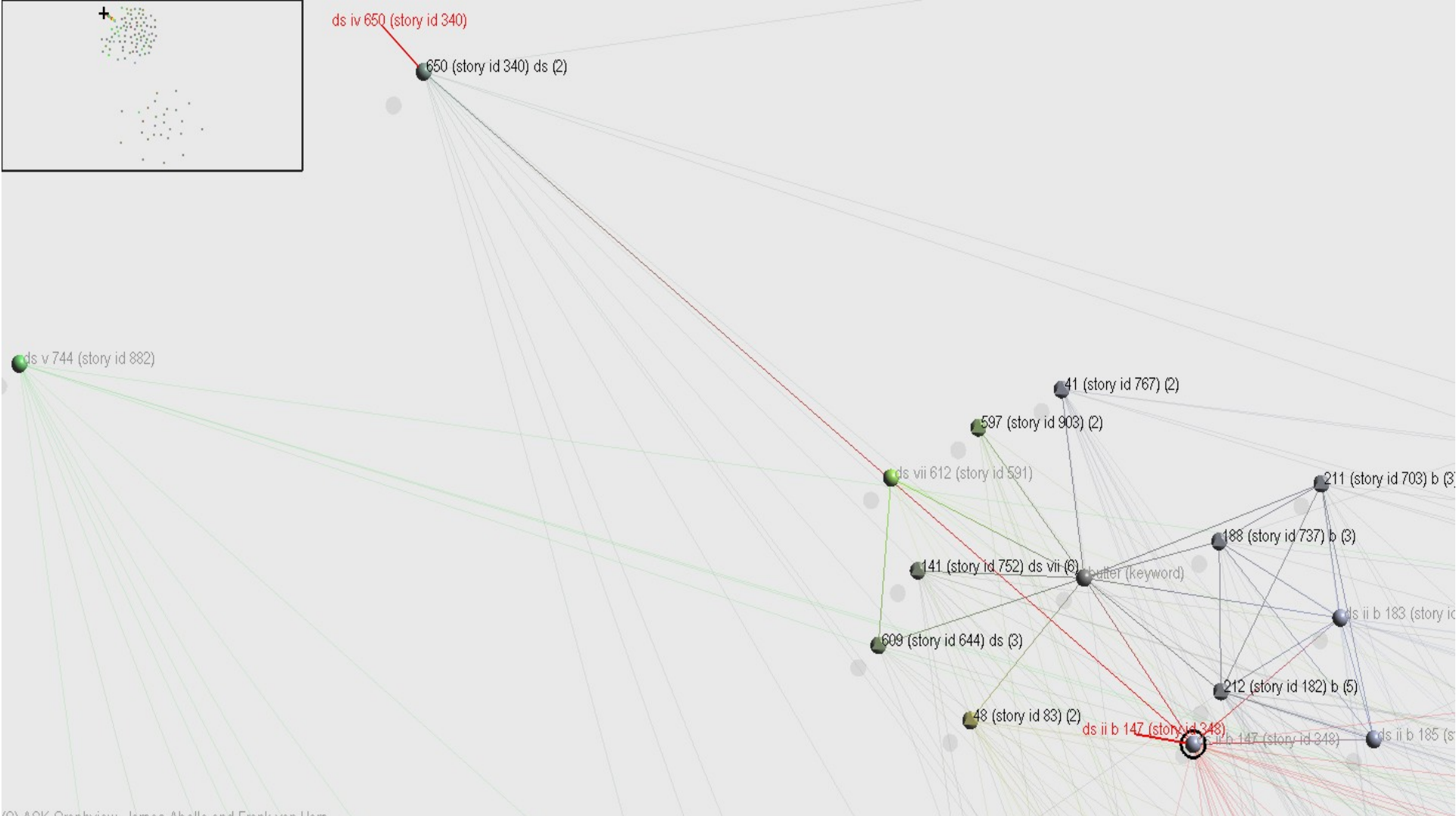
Using a network map to help with text classification

- Limit the edgelist to story data
 - ☐ Keywords
 - ☐ Topic Index
 - ☐ Shallow Ontology
 - ☐ Named entities
- Sort the network
- Find my target story
- Look for stories in the immediate neighborhood
 - ☐ layout is no longer random

Using a network map to help with text classification (unsorted)



Government	Percentage
Current government	85%
Previous government	15%





Networks as text classifier

- Places the text in a neighborhood of like minded stories (stories about ghosts)
- Oddly places the text in a neighborhood of stories that mention butter
- Worth a look...



Remember our ghost story?

- DS IV 650
- Classified as a story about manor lords, not ghosts!
- Impossible to find in the archive if one is looking for a ghost story
- Found it in the network neighborhood along with DS II B 147



DS II B 147

- A story about a house elf at a farm in Egå...
- Ends as follows:
 - When they got home, the farmhand was happy because now he'd gotten something to use for feed, and afterward *nis* could go and feed just as much as he wanted to. Then they got another farmhand, and he didn't want to let him go on like that. But he got lifted up in his bed and all the way up to the rafters, so he lay there dead when people got up the next morning.



More questions

- Is this just an idiosyncratic illustration?
- How can we make use of all of the other network data in a meaningful way?
- Can we use the folklore data for more than just the study of folklore?



Folklore and Google Books

- Can we discover folklore topics in the Nordic literary corpus, 1800-1923, based on our knowledge of folklore topics in Tang Kristensen?
- Can an awareness of the connections between authors (mined from literary histories) help us trace influence?



A little experiment

- Create a series of topics modeled on the Tang Kristensen collection, using Latent Dirichlet Allocation (Blei et al. 2003)
 - 50 topics
- Chunk the short stories of St. Steensen Blicher
 - author known for using folkloric topics and topics from everyday life
- Map the alignment between folklore topics and Blicher



Possible directions

- Use knowledge of Blicher from literary histories to increase the target group (iterative)
- Map drop-off in the folkloric across the literary landscape



Thanks / Funding

■ Thanks to:

- Peter Leonard (UCLA)
- Dr. Peter Broadwell (UCLA)

■ Funding

- Google Books Humanities Grants
- NSF #IIS-0970179 “Network Pattern Recognition for the Humanities”
- NEH #HT5001609 “Network Analysis for the Humanities”