MICHEL BIEZUNSKI, INFOLOOM

DATA CLEANING VS. HUMAN CURATION

AI Open Camps, New York, Nov. 19, 2017

CONTENTS

- Concept: Cleaning and Curation
- Experiment: Integrating Indexes
- Curation Philosophy

BEYOND BIG DATA

- Powerful algorithms are used to make sense of unknown data
  - Text analysis, Image analysis, Voice analysis, Sentiment Analysis
  - Data analytics, AI.
- A world of known data

NEXT STEPS

- 1. Improve Technology
  - Develop more accurate algorithms: Improved pattern recognition, more data in, machine learning.
- 2. Involve Humans
  - Low paid/volunteers: crowd-sourcing, manual tagging
  - Subject matter experts: librarians, taxonomists, information architects.
  - Decision makers, Information asset managers

CLEANING DATA?

- To be processed into a computer system, data must be “accredited”. It must obey certain rules.
    - Fit into a database field, XML element, spreadsheet column, taxonomy classification.
    - May require use of controlled vocabulary
- Data that doesn’t fit must be cleaned to be entered.
- Process can be tedious and time-consuming.

WHAT ABOUT NON-CLEAN DATA?

- Ambiguous
- Complex
- Nuanced
- Foreign
- Controversial
- Original
- New
- Innovative
- Creative
- Implicit
- Unexpected
- Misspelled
- Uses non-authoritative terms

HOME SWEET HOME

- New Year Resolution Syndrome: When entering a new house, we organize our stuff into drawers, shelves, armoires, closets. Everything fits.

- As time goes by, mess starts to accumulate. Things are not put where they should be, there is a little bit of everything everywhere.

- A certain amount of mess is useful to feel at home. Too much mess can be stressful. It depends on each of us.

CORPORATE INFORMATION CHALLENGES

- Schema-driven information architectures impose every piece of information to fit somewhere.
- New data doesn't necessarily fit the existing locations. It is likely that some would need new fields, but creating them may be impossible, too costly, or too resource intensive.
- When data doesn't fit the existing silos, the whole silo is discarded and a new one is designed.
- The same story starts again….

WHAT IF?

- Mess is a feature, not a bug
- Information diverse, and constantly changing.
- Needs hospitable systems, accepting flexibility.
- Graph databases vs. Relational/Object Databases.

BOOK INDEXES: A STORY

- Powerful ways to quickly find information
    - Human work: author or professional indexer
- Can we create indexes through algorithms?
    - Yes, if entries are occurrences of a set of words.
    - Yes, if information complies with a well-defined structure
    - No, in any other case.

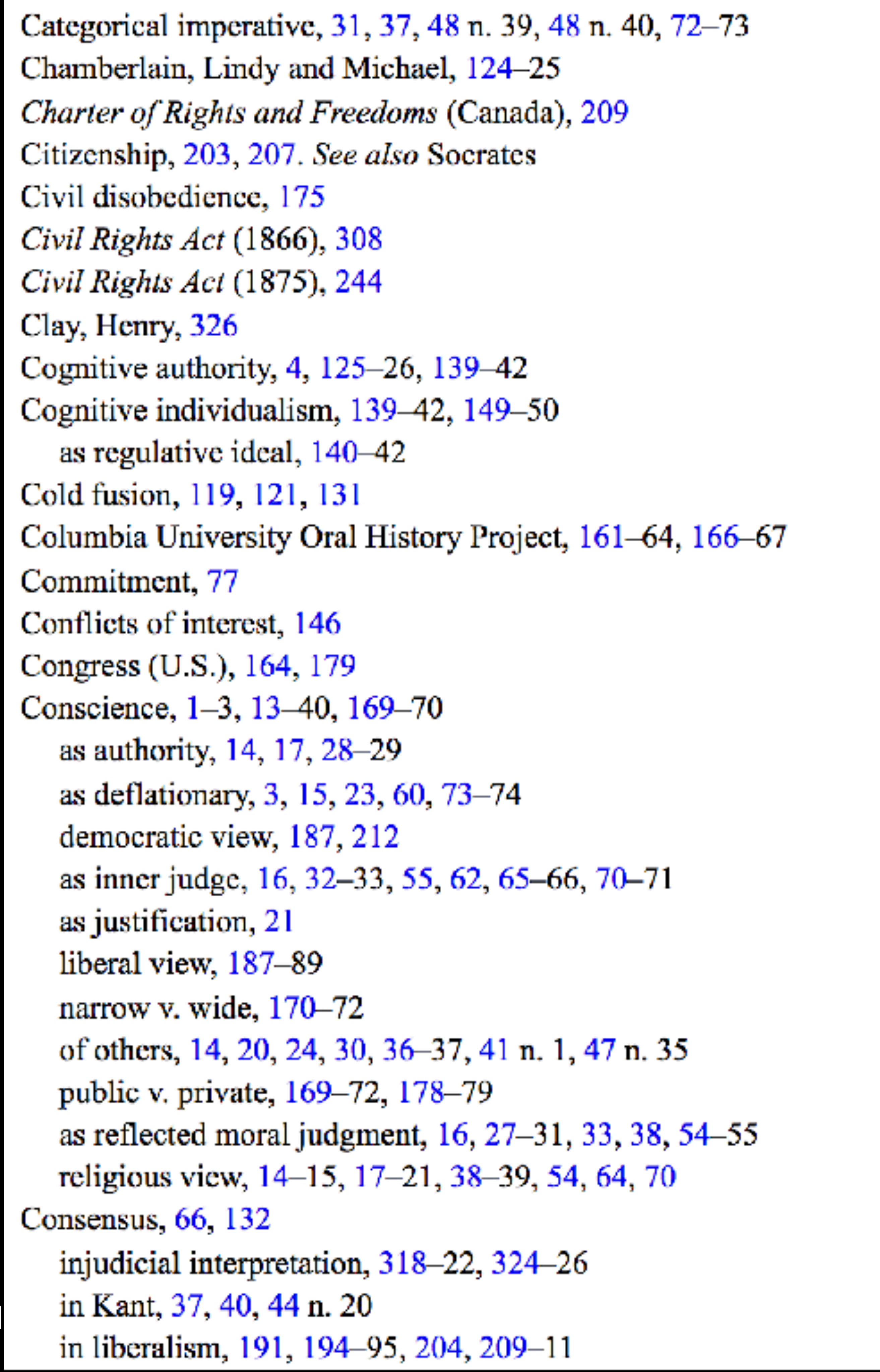
INTEGRATING BOOK INDEXES

- Idea: Why not integrate together indexes of books within a collection?

- Goal: Provide to the readers a way to directly access any entries of the books.

- Project with NYU Library, 3 academic publishers: NYU Press, U of Michigan Press, U of Minnesota Press. Founded by Carnegie Mellon. Still ongoing.

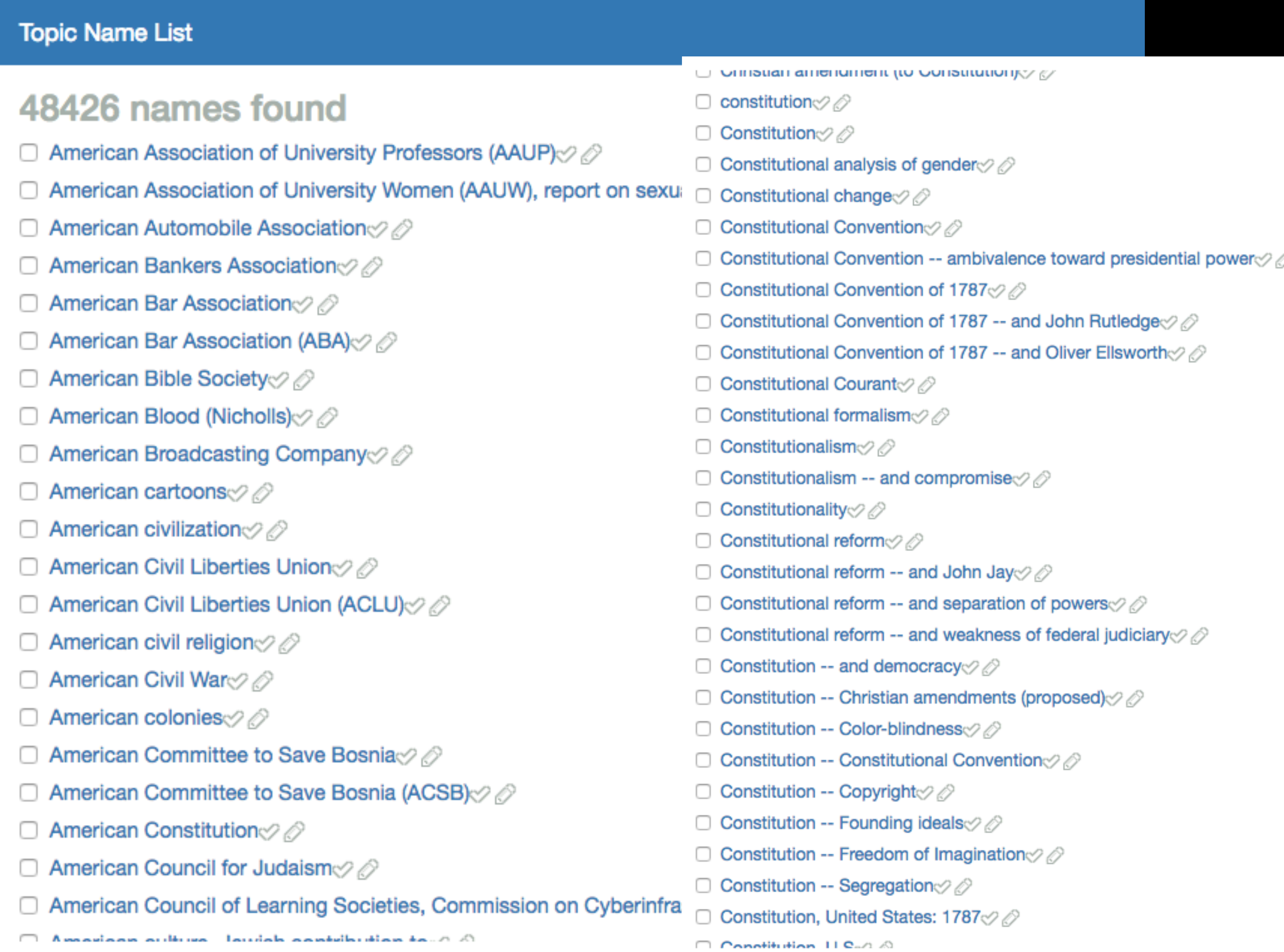
INDEXES ARE COMPLEX…



- Categorical imperative, 31, 37, 48 n. 39, 48 n. 40, 72–73
- Chamberlain, Lindy and Michael, 124–25
- *Charter of Rights and Freedoms* (Canada), 209
- Citizenship, 203, 207. *See also* Socrates
- Civil disobedience, 175
- *Civil Rights Act* (1866), 308
- *Civil Rights Act* (1875), 244
- Clay, Henry, 326
- Cognitive authority, 4, 125–26, 139–42
- Cognitive individualism, 139–42, 149–50
  - as regulative ideal, 140–42
- Cold fusion, 119, 121, 131
- Columbia University Oral History Project, 161–64, 166–67
- Commitment, 77
- Conflicts of interest, 146
- Congress (U.S.), 164, 179
- Conscience, 1–3, 13–40, 169–70
  - as authority, 14, 17, 28–29
  - as deflationary, 3, 15, 23, 60, 73–74
  - democratic view, 187, 212
  - as inner judge, 16, 32–33, 55, 62, 65–66, 70–71
  - as justification, 21
  - liberal view, 187–89
  - narrow v. wide, 170–72
  - of others, 14, 20, 24, 30, 36–37, 41 n. 1, 47 n. 35
  - public v. private, 169–72, 178–79
  - as reflected moral judgment, 16, 27–31, 33, 38, 54–55
  - religious view, 14–15, 17–21, 38–39, 54, 64, 70
- Consensus, 66, 132
  - injudicial interpretation, 318–22, 324–26
  - in Kant, 37, 40, 44 n. 20
  - in liberalism, 191, 194–95, 204, 209–11

INTEGRATION OF INDEXES

- Increases Complexity
- For 100 books yielded 48,000 names
  - 45,000 topics (3,000 names were merged into common topics)



Topic Name List

48426 names found

- □ American Association of University Professors (AAUP) ✓ ✎
- □ American Association of University Women (AAUW), report on sexu
- □ American Automobile Association ✓ ✎
- □ American Bankers Association ✓ ✎
- □ American Bar Association ✓ ✎
- □ American Bar Association (ABA) ✓ ✎
- □ American Bible Society ✓ ✎
- □ American Blood (Nicholls) ✓ ✎
- □ American Broadcasting Company ✓ ✎
- □ American cartoons ✓ ✎
- □ American civilization ✓ ✎
- □ American Civil Liberties Union ✓ ✎
- □ American Civil Liberties Union (ACLU) ✓ ✎
- □ American civil religion ✓ ✎
- □ American Civil War ✓ ✎
- □ American colonies ✓ ✎
- □ American Committee to Save Bosnia ✓ ✎
- □ American Committee to Save Bosnia (ACSB) ✓ ✎
- □ American Constitution ✓ ✎
- □ American Council for Judaism ✓ ✎
- □ American Council of Learning Societies, Commission on Cyberinfra
- □ American culture -- Jewish contribution to ✓ ✎

- □ Christian amendment (to Constitution) ✓ ✎
- □ constitution ✓ ✎
- □ Constitution ✓ ✎
- □ Constitutional analysis of gender ✓ ✎
- □ Constitutional change ✓ ✎
- □ Constitutional Convention ✓ ✎
- □ Constitutional Convention -- ambivalence toward presidential power ✓
- □ Constitutional Convention of 1787 ✓ ✎
- □ Constitutional Convention of 1787 -- and John Rutledge ✓ ✎
- □ Constitutional Convention of 1787 -- and Oliver Ellsworth ✓ ✎
- □ Constitutional Courant ✓ ✎
- □ Constitutional formalism ✓ ✎
- □ Constitutionalism ✓ ✎
- □ Constitutionalism -- and compromise ✓ ✎
- □ Constitutionality ✓ ✎
- □ Constitutional reform ✓ ✎
- □ Constitutional reform -- and John Jay ✓ ✎
- □ Constitutional reform -- and separation of powers ✓ ✎
- □ Constitutional reform -- and weakness of federal judiciary ✓ ✎
- □ Constitution -- and democracy ✓ ✎
- □ Constitution -- Christian amendments (proposed) ✓ ✎
- □ Constitution -- Color-blindness ✓ ✎
- □ Constitution -- Constitutional Convention ✓ ✎
- □ Constitution -- Copyright ✓ ✎
- □ Constitution -- Founding ideals ✓ ✎
- □ Constitution -- Freedom of Imagination ✓ ✎
- □ Constitution -- Segregation ✓ ✎
- □ Constitution, United States: 1787 ✓ ✎
- □ Constitution, U.S. ✓ ✎

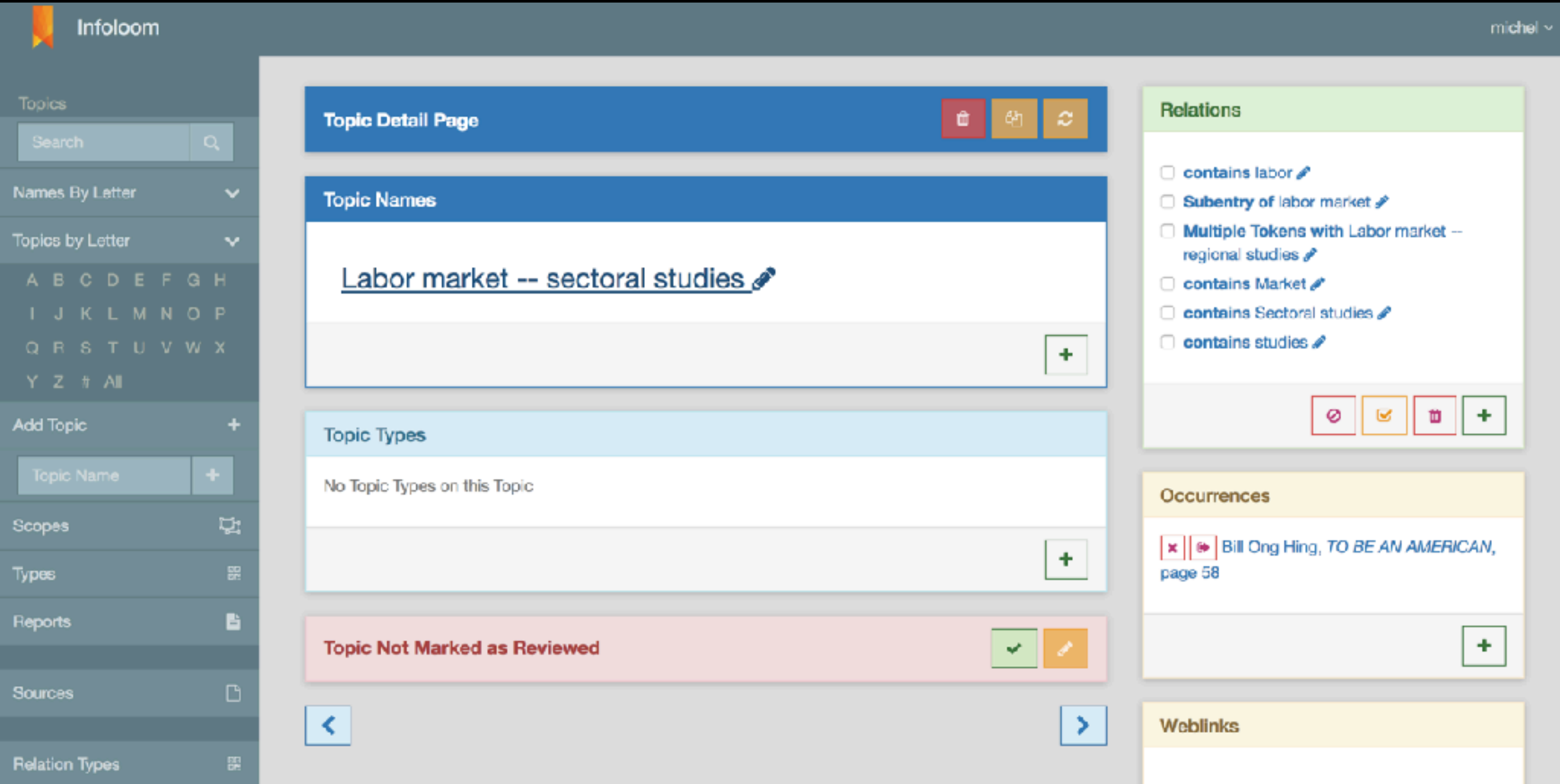
WHAT HAPPENED?

- We have built a tool to manage topic integration.
- NYU has hired a “Semantic Editor” who worked on cleaning the information and finding methods.

WHAT DOES "CLEANING" MEAN?

- Topics are grouped into semantic clusters sharing the same meaning.
- However, since their number has no limit, there is no constraint on how many topics there should be.
- Several strategies are possible:
    - Use automated processes to coalesce them: slugification, lemmatization, use of specialized dictionary
    - Manual curation for fine-tuning.
- In any case, the result is usable, even without cleaning.

EDITORIAL INTERFACE



Topic Detail Page

Topic Names

Labor market -- sectoral studies

Topic Types

No Topic Types on this Topic

Topic Not Marked as Reviewed

Relations

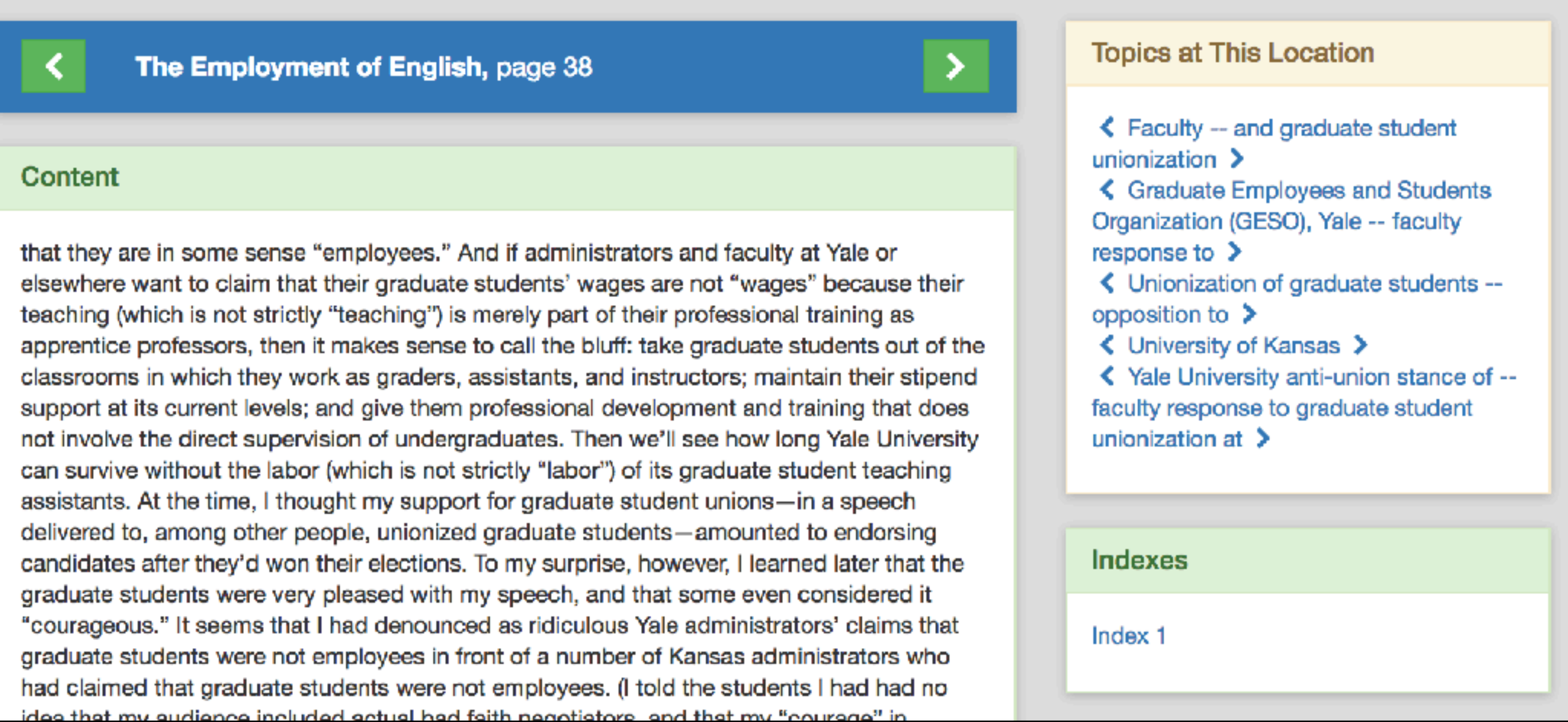
- □ contains labor
- □ Subentry of labor market
- □ Multiple Tokens with Labor market -- regional studies
- □ contains Market
- □ contains Sectoral studies
- □ contains studies

Occurrences

Bill Ong Hing, *TO BE AN AMERICAN*, page 58

Weblinks

EBOOK PAGE



The Employment of English, page 38

Content

that they are in some sense “employees.” And if administrators and faculty at Yale or elsewhere want to claim that their graduate students’ wages are not “wages” because their teaching (which is not strictly “teaching”) is merely part of their professional training as apprentice professors, then it makes sense to call the bluff: take graduate students out of the classrooms in which they work as graders, assistants, and instructors; maintain their stipend support at its current levels; and give them professional development and training that does not involve the direct supervision of undergraduates. Then we’ll see how long Yale University can survive without the labor (which is not strictly “labor”) of its graduate student teaching assistants. At the time, I thought my support for graduate student unions—in a speech delivered to, among other people, unionized graduate students—amounted to endorsing candidates after they’d won their elections. To my surprise, however, I learned later that the graduate students were very pleased with my speech, and that some even considered it “courageous.” It seems that I had denounced as ridiculous Yale administrators’ claims that graduate students were not employees in front of a number of Kansas administrators who had claimed that graduate students were not employees. (I told the students I had had no idea that my audience included actual bad faith negotiators, and that my “courage” in

Topics at This Location

- Faculty -- and graduate student unionization
- Graduate Employees and Students Organization (GESO), Yale -- faculty response to
- Unionization of graduate students -- opposition to
- University of Kansas
- Yale University anti-union stance of -- faculty response to graduate student unionization at

Indexes

Index 1

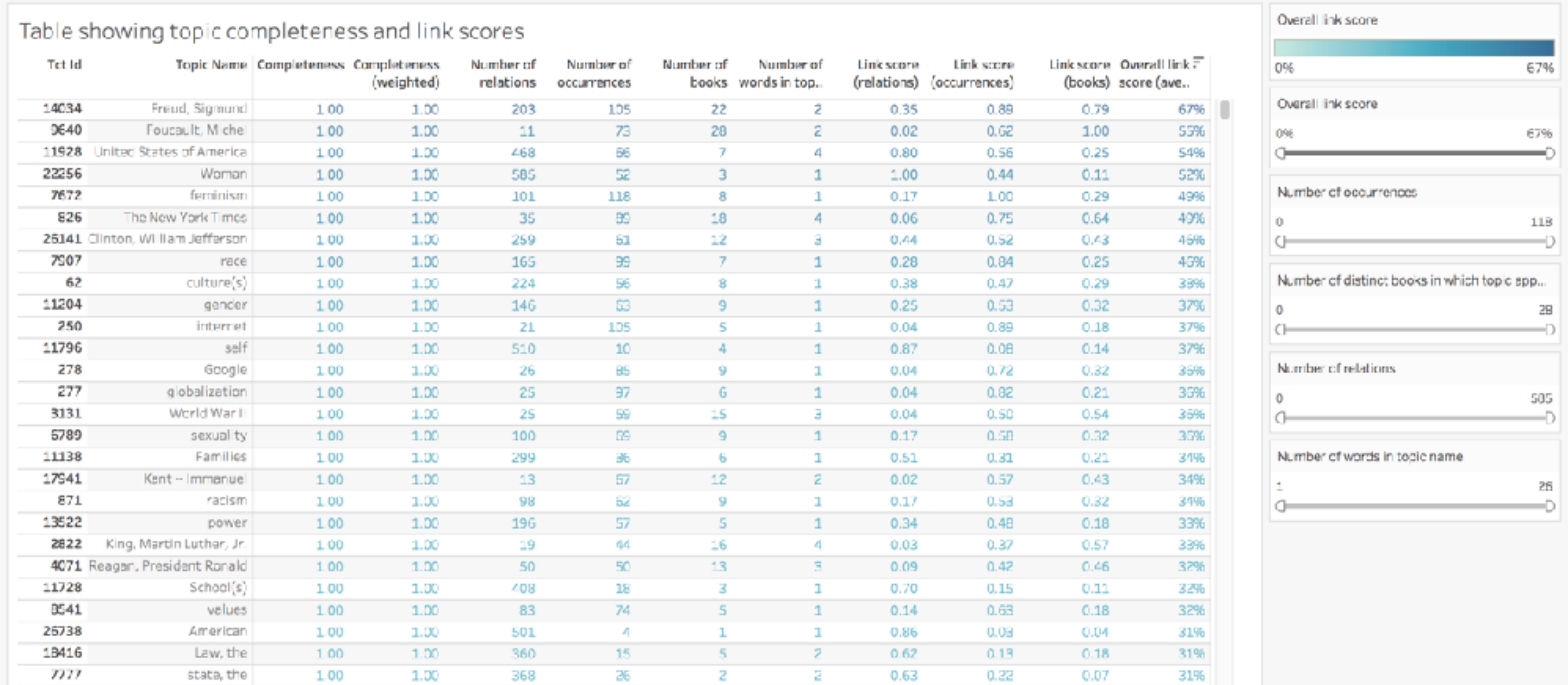


Table showing topic completeness and link scores

Tct Id	Topic Name	Completeness	Completeness (weighted)	Number of relations	Number of occurrences	Number of books	Number of words in top..	Link score (relations)	Link score (occurrences)	Link score (books)	Overall link score (ave..
14034	Freud, Sigmund	1.00	1.00	203	105	22	2	0.35	0.89	0.79	67%
9640	Foucault, Michel	1.00	1.00	11	73	28	2	0.02	0.62	1.00	55%
11928	United States of America	1.00	1.00	468	66	7	4	0.80	0.56	0.25	54%
22256	Woman	1.00	1.00	585	52	3	1	1.00	0.44	0.11	52%
7672	feminism	1.00	1.00	101	118	8	1	0.17	1.00	0.29	49%
826	The New York Times	1.00	1.00	35	89	18	4	0.06	0.75	0.64	49%
25141	Clinton, William Jefferson	1.00	1.00	259	61	12	3	0.44	0.52	0.43	46%
7307	race	1.00	1.00	165	99	7	1	0.28	0.84	0.25	45%
62	culture(s)	1.00	1.00	224	56	8	1	0.38	0.47	0.29	39%
11204	gender	1.00	1.00	146	63	9	1	0.25	0.53	0.32	37%
250	internet	1.00	1.00	21	105	5	1	0.04	0.89	0.18	37%
11796	self	1.00	1.00	510	10	4	1	0.87	0.08	0.14	37%
278	Google	1.00	1.00	26	85	9	1	0.04	0.72	0.32	36%
277	globalization	1.00	1.00	25	97	6	1	0.04	0.82	0.21	35%
3131	World War II	1.00	1.00	25	59	15	3	0.04	0.50	0.54	36%
5789	sexuality	1.00	1.00	100	69	9	1	0.17	0.58	0.32	35%
11138	Families	1.00	1.00	299	36	6	1	0.51	0.31	0.21	34%
17941	Kent -- Immanuel	1.00	1.00	13	67	12	2	0.02	0.57	0.43	34%
871	racism	1.00	1.00	98	62	9	1	0.17	0.53	0.32	34%
13522	power	1.00	1.00	196	57	5	1	0.34	0.48	0.18	33%
2822	King, Martin Luther, Jr.	1.00	1.00	19	44	16	4	0.03	0.37	0.57	33%
4071	Reagan, President Ronald	1.00	1.00	50	50	13	3	0.09	0.42	0.46	32%
11728	School(s)	1.00	1.00	408	18	3	1	0.70	0.15	0.11	32%
3541	values	1.00	1.00	83	74	5	1	0.14	0.63	0.18	32%
25738	American	1.00	1.00	501	4	1	1	0.86	0.03	0.04	31%
13416	Law, the	1.00	1.00	360	15	5	2	0.62	0.13	0.18	31%
7777	stata, the	1.00	1.00	368	26	2	2	0.63	0.22	0.07	31%

Overall link score
0% 67%

Overall link score
0% 67%

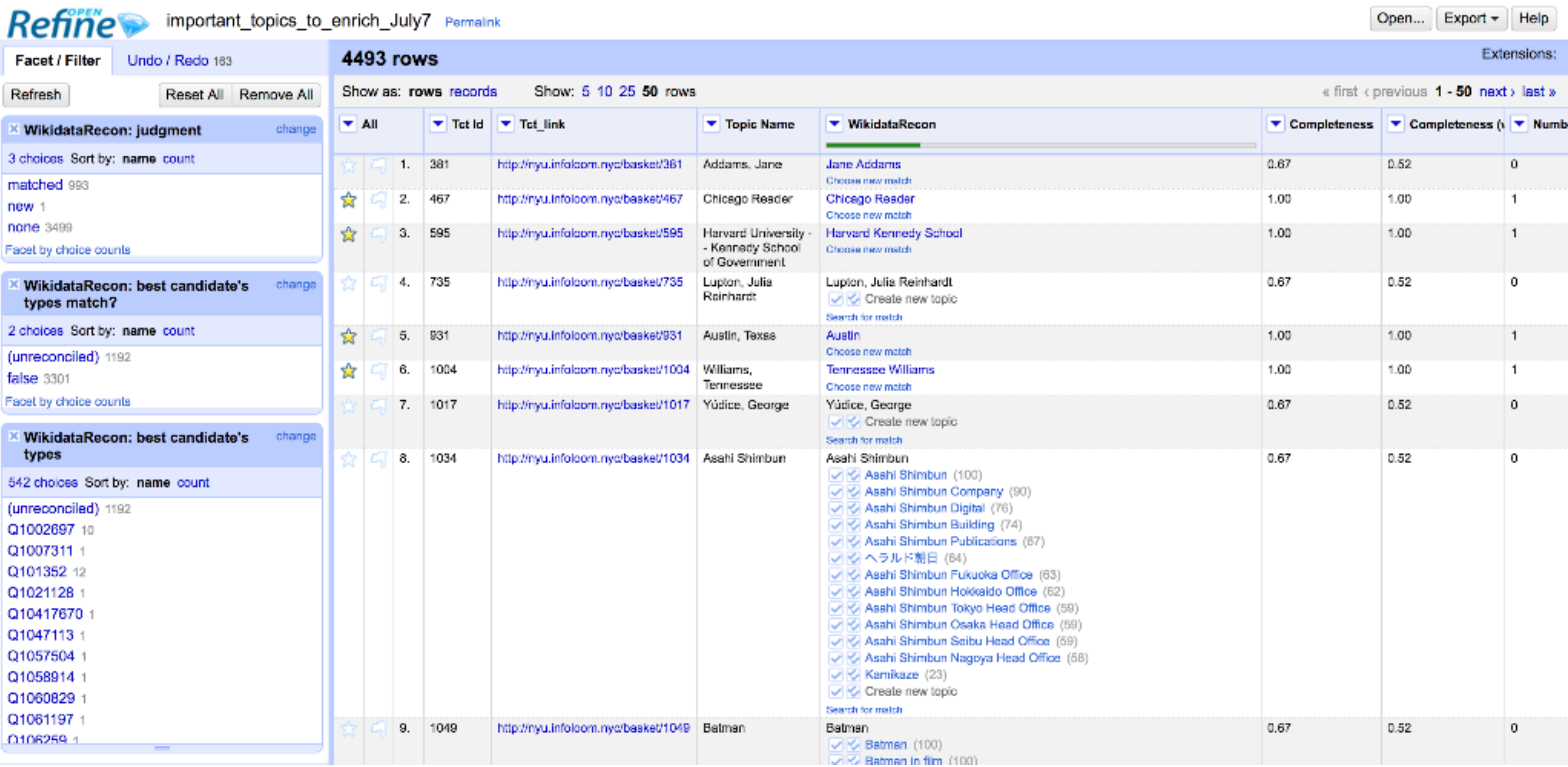
Number of occurrences
0 118

Number of distinct books in which topic app..
0 28

Number of relations
0 585

Number of words in topic name
1 26

Metric based on Ochoa, X., & Duval, E. (2009). Automatic evaluation of metadata quality in digital repositories.



important_topics_to_enrich_July7 Permalink

4493 rows

Show as: **rows** records   Show: 5 10 25 **50** rows   « first ‹ previous **1 - 50** next › last »

☒ **WikidataRecon: judgment** change

3 choices Sort by: **name** count

matched	993
new	1
none	3499

Facet by choice counts

☒ **WikidataRecon: best candidate's types match?** change

2 choices Sort by: **name** count

(unreconciled)	1192
false	3301

Facet by choice counts

☒ **WikidataRecon: best candidate's types** change

542 choices Sort by: **name** count

(unreconciled)	1192
Q1002697	10
Q1007311	1
Q101352	12
Q1021128	1
Q10417670	1
Q1047113	1
Q1057504	1
Q1058914	1
Q1060829	1
Q1061197	1
Q106259	1

All	Tct Id	Tct_link	Topic Name	WikidataRecon	Completeness	Completeness (v	Numb
☆ 1.	381	http://nyu.infoloom.nyc/basket/381	Addams, Jane	Jane Addams Choose new match	0.67	0.52	0
★ 2.	467	http://nyu.infoloom.nyc/basket/467	Chicago Reader	Chicago Reader Choose new match	1.00	1.00	1
★ 3.	595	http://nyu.infoloom.nyc/basket/595	Harvard University - - Kennedy School of Government	Harvard Kennedy School Choose new match	1.00	1.00	1
☆ 4.	735	http://nyu.infoloom.nyc/basket/735	Lupton, Julia Reinhardt	Lupton, Julia Reinhardt ☑ ☑ Create new topic Search for match	0.67	0.52	0
★ 5.	931	http://nyu.infoloom.nyc/basket/931	Austin, Texas	Austin Choose new match	1.00	1.00	1
★ 6.	1004	http://nyu.infoloom.nyc/basket/1004	Williams, Tennessee	Tennessee Williams Choose new match	1.00	1.00	1
☆ 7.	1017	http://nyu.infoloom.nyc/basket/1017	Yúdice, George	Yúdice, George ☑ ☑ Create new topic Search for match	0.67	0.52	0
☆ 8.	1034	http://nyu.infoloom.nyc/basket/1034	Asahi Shimbun	Asahi Shimbun ☑ ☑ Asahi Shimbun (100) ☑ ☑ Asahi Shimbun Company (90) ☑ ☑ Asahi Shimbun Digital (76) ☑ ☑ Asahi Shimbun Building (74) ☑ ☑ Asahi Shimbun Publications (67) ☑ ☑ ヘラルド朝日 (64) ☑ ☑ Asahi Shimbun Fukuoka Office (63) ☑ ☑ Asahi Shimbun Hokkaido Office (62) ☑ ☑ Asahi Shimbun Tokyo Head Office (59) ☑ ☑ Asahi Shimbun Osaka Head Office (59) ☑ ☑ Asahi Shimbun Seibu Head Office (59) ☑ ☑ Asahi Shimbun Nagoya Head Office (58) ☑ ☑ Kamikaze (23) ☑ ☑ Create new topic Search for match	0.67	0.52	0
☆ 9.	1049	http://nyu.infoloom.nyc/basket/1049	Batman	Batman ☑ ☑ Batman (100) ☑ ☑ Batman in film (100)	0.67	0.52	0

AGAINST CLEANING?

Against Cleaning” (Rawson & Muñoz, Curating Menus)

- Mess is part of life.
- We need to design systems that are open to it.
- Similar to democracy. It’s more difficult to accommodate various points of view, but it’s worth it.

WHAT IS DATA WORTH?

- If it’s just collected, it’s interesting for statistical usages.
- If it’s known, has been appropriately reviewed and curated, it has much more value.
- If your business/activity relies on providing good data, you’d better think about it!
- Technology is here to help humans, not to prevent us to do what we need to do!