MICHEL BIEZUNSKI, INFOLOOM

TOPIC CURATION: MANUAL VS. AUTOMATED

Gilbane Digital Content Conference, Boston, November 29, 2017

CONTENTS

I. Conceptual Introduction

Topic, Curation, and Cleaning

II Two real world applications:

Tax Information and Academic Books

III A Practical Approach

I. CONCEPTUAL INTRODUCTION

BIG DATA

- Data Analytics:
 - Text, Voice, Image,
 Sentiment Analysis
- Search Engine Proprietary Technologies
- Artificial Intelligence,
 Machine Learning

- You are lucky to find what you need.
- There is so much out there that you practically always find something.
- Cleaning not relevant.

SMALL DATA

- Information owners know their content.
- They want their users to make the best of it.
- Good content is key to success.

- Data usually structured
- Sometimes unstructured

STRUCTURED DATA

- Databases
- XML Schemas
- Taxonomies
- Ontologies
- Controlled Vocabularies
- Authority Keywords

- Agreements required.
- Anything that doesn't fit can't be entered.
 - => Data cleaning mandatory.

CORPORATE INFORMATION CHALLENGES

 Information models are often too rigid. Creating new fields may be either 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 ABOUT DATA THAT DOESN'T FIT?

- Misspelled
- Uses nonauthoritative terms
- Foreign

- New, unheard of
 Ambiguous
- Innovative
- Creative
- Original
- Unexpected

- Complex
- Nuanced
- Implicit
- Controversial

HOME SWEET HOME

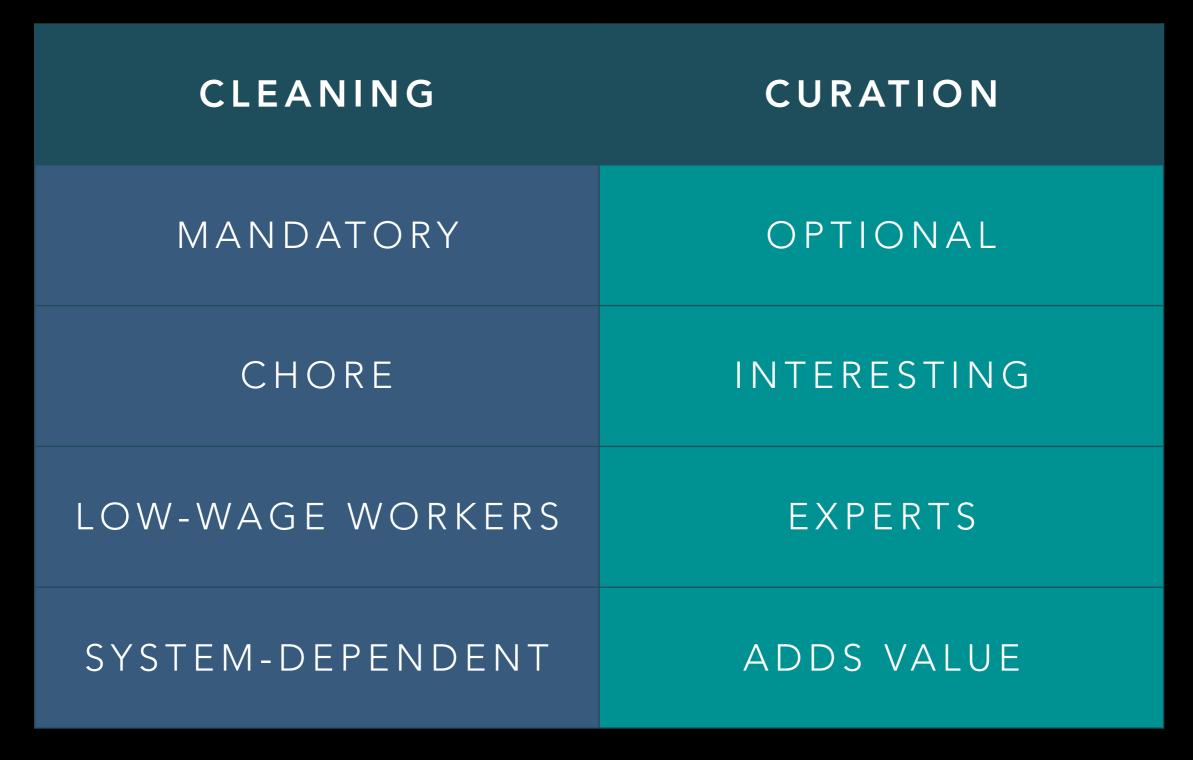
- New Year Resolution Syndrome: When entering a new house, we organize our stuff into drawers, shelves, armoires, closets. Everything fits.
- As time goes by, mess starts to accumulate. Things are not put where they should be, there is a little bit of everything everywhere.

WHAT IF MESS IS A FEATURE, NOT A BUG?

- Information is diverse, and constantly changing.
- Individuals have different opinions, and deserve to be heard.
- We need hospitable, flexible, adaptable information systems.

Graph Databases now trending vs. Relational/ Object Databases.

CLEANING VS. CURATION



Michel Biezunski, Infoloom. Gilbane Digital Content Conference, Boston, November 29, 2017.

TOPIC MAPS

- "Topic": Computer representation of a unique meaning.
- "Map": Graph-based navigation between topics.
- Standards: ISO 13250, RDF, Open Linked Data, DITA, OpenCypher...

II. TWO REAL-WORLD APPLICATIONS

CURATING COMPLEX DATA SETS

- IRS Tax Map: Latest Available Tax Information for Taxpayers
- NYU Enhanced Networked Monographs: Humanities Books' Indexes

THE IRS TAXMAP STORY



- Started as a way to help call centers' assistors to provide more accurate responses to taxpayers' questions.
- Still up, 15 years later, after 3 attempts at taxonomies, and web site overhaul
- Secret: It doesn't have to be 100% cleaned.
- Available at: https://taxmap.irs.gov



Tax Map Search:

GO

Search Help Navigation Help

Tax Map Index ABCDEFGHI JKLMNOPQR STUVWXYZ#

International Tax Topic Index

Affordable Care Act Tax Topic Index

Exempt Organization Tax Topic Index

> FAQs Forms Publications Tax Topics Worksheets

Comments
About Tax Map

IRS.gov Website

Form 941--Employer's Quarterly Federal Tax Return

View Forms Publications Tax Topics

Useful Links

Popular Publication Links

Reporting Adjustments to Form 941 or Form 944 Filing Form 941 or Form 944

Related Topic

Employment Taxes

Prior Year Form 941

Forms and Instructions

Form 941 Employer's Quarterly Federal Tax Return
Instructions for Form 941, Employer's Quarterly
Federal Tax Return

Related Forms

Statements

Form W-2 Wage and Tax Statement (Info Copy Only)

Instructions for Forms W-2 and W-3, Wage and
Tax Statement and Transmittal of Wage and Tax

Form 720 Quarterly Federal Excise Tax Return

Instructions for Form 720, Quarterly Federal
Excise Tax Return

Form 720-X Amended Quarterly Federal Excise Tax Return

Form 940 Employer's Annual Federal Unemployment (FUTA) Tax Return

Instructions for Form 940, Employer's Annual

Related Topic Links

- Employer
- ➤ Employment Tax
- Federal Insurance Contributions Act (FICA)
- ➤ Federal Tax Deposit (FTD)
- ➤ Filing Form 941 or Form 944
- Medicare
- ➤ Prior Year Form 941
- > Quarterly Tax Payments
- Reporting Adjustments to Form 941 or Form 944
- Social Security and Medicare Taxes
- > Withholding of Tax

MORE

MORE

MORE

MORE

TAXMAP CURATION

- Application of automatic rules
 - Containment, 3 words, remove topics with no occurrences and no relations.
- Maintenance of a knowledge base
 - Merge topics by setting synonyms, Related topics,
 Rejected topics
- Currently done using custom configuration settings.

BOOK INDEXES: A STORY

- Indexes: A must for professionals.
 - Human work: author or professional indexer
- Can we create indexes through algorithms?
 - Yes, if entries are occurrences of words in text, or if information is very strictly structured.
 - No, in any other case.

INTEGRATING BOOK INDEXES

- Idea: Integrate indexes of many books into a master index
- Goal: Provide the readers with direct access to any index entry.
- 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
```

American sulture. Invide contribution to A

	Uninstian amenument (to Constitution)							
48426 names found	□ constitution ⊘ Ø							
	□ Constitution ⊘ Ø							
☐ American Association of University Professors (AAUP)	□ Constitutional analysis of gender ⊘ Ø							
American Association of University Women (AAUW), report on sexual	□ Constitutional change							
☐ American Automobile Association ♥ Ø	□ Constitutional Convention ⊘ ⊘							
☐ American Bankers Association	□ Constitutional Convention ambivalence toward presidential povential							
☐ American Bar Association ⊘ ⊘	□ Constitutional Convention of 1787							
☐ American Bar Association (ABA)	□ Constitutional Convention of 1787 and John Rutledge							
	□ Constitutional Convention of 1787 and Oliver Ellsworth							
American Bible Society	□ Constitutional Courant ⊘ ⊘							
☐ American Blood (Nicholls)	□ Constitutional formalism							
☐ American Broadcasting Company	□ Constitutionalism							
□ American cartoons ⊘ ⊘	□ Constitutionalism and compromise ◊ ◊							
	□ Constitutionality ◊ ◊							
American civilization	□ Constitutional reform							
☐ American Civil Liberties Union	□ Constitutional reform and John Jay							
☐ American Civil Liberties Union (ACLU)	□ Constitutional reform and separation of powers							
☐ American civil religion ⊘ ⊘	□ Constitutional reform and weakness of federal judiciary							
□ American Civil War⊘ ⊘	□ Constitution and democracy ⊘ Ø							
	□ Constitution Christian amendments (proposed)							
American colonies	□ Constitution Color-blindness							
☐ American Committee to Save Bosnia	□ Constitution Constitutional Convention							
☐ American Committee to Save Bosnia (ACSB)	☐ Constitution Copyright ⊘ ⊘							
☐ American Constitution	□ Constitution Founding ideals							
☐ American Council for Judaism♡ ⊘	□ Constitution Freedom of Imagination							
	□ Constitution Segregation							
 American Council of Learning Societies, Commission on Cyberinfra 	□ Constitution, United States: 1787							

Constitution II 9-/2 /

HOW DID WE WORK?

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

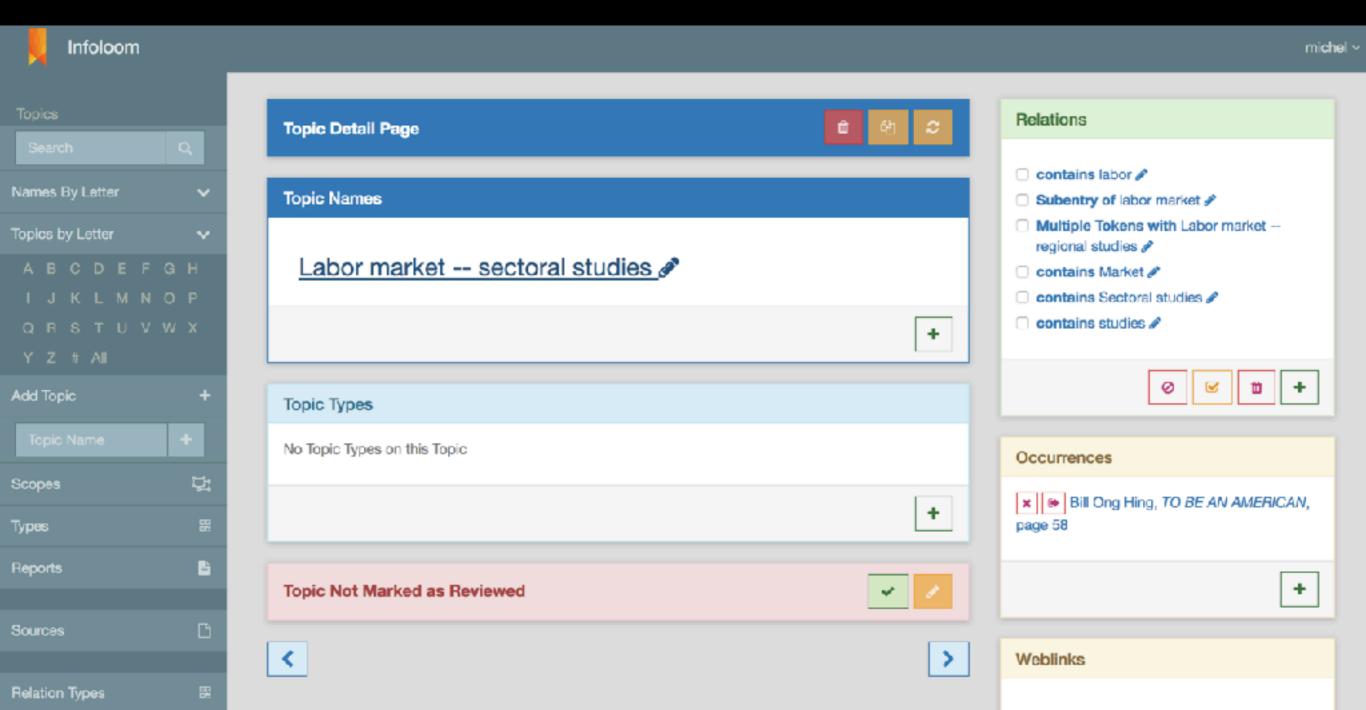
INTEGRATION OF INDEXES

- For ~80 books: yielded 48,000 entries.
 - 15,000 topics were curated.
 - 3,000 entries were merged. Currently, 45,000 topics.

AUTOMATIC PROCESSING TECHNIQUES

- Slugification
- Lemmatization
- Use of specialized dictionary containing predefined synonyms
- Automatic relationships:
 - See/See also, Entry/Subentries, Containment, 3 tokens

EDITORIALINTERFACE



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 had faith negotiators, and that my "courage" in

Topics at This Location

- Faculty -- and graduate student unionization >
- 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

Topic curation workflows

DC-2017, October 28 I @AlexandraAlisa



Topic Curation Toolkit automatically:

- merges topics
- creates relations between topics



Human editor *manually*:

- fixes processing errors
- removes certain relations
- merges synonyms and splitting up homonyms
- adds relations tool did not find
- adding new data, sometimes from external sources like Wikidata







DC-2017, October 28 I

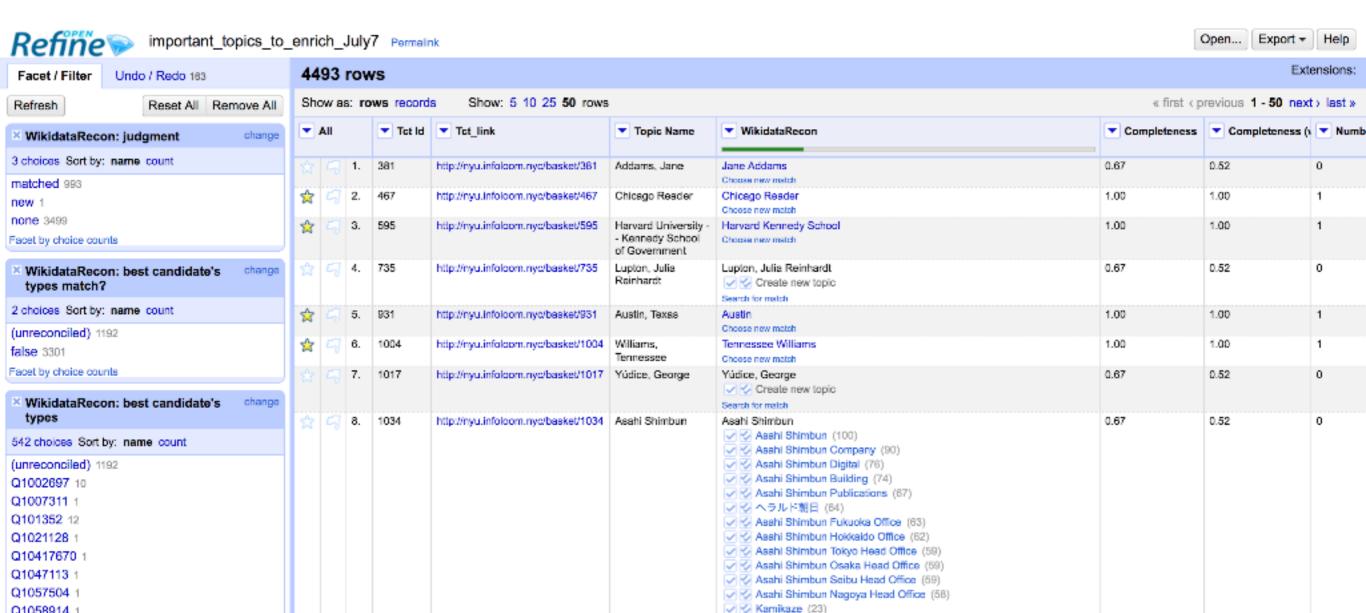
able si	howing topic co	mpletene	ss and link	scores								Overall link score	
		Completeness		Number of relations	Number of occurrences	Number of books	Number of words in top	Link score (relations)	Link score (occurrences)		Overall link =	0%	679
14034	Freud, Sigmund	1.00	1.00	203	105	22	2	0.35	0.89	0.79	67%	Overall link score	
9640	Foucault, Michel	1.00	1.00	11	73	28	2	0.02	0.62	1.00	55%	096	679
11928	United States of America	1.00	1.00	468	56	7	4	0.80	0.56	0.25	54%	0	
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	4996	Number of occurrence	005
826	The New York Times	1.00	1.00	35	89	18	4	0.06	0.75	0.64	49%	0	118
25141	Clinton, William Jefferson	1.00	1.00	259	61	12	3	0.44	0.52	0.43	45%	0	
7907	race	1.00	1.00	165	99	7	1	0.28	0.84	0.25	46%		
62	culture(s)	1.00	1.00	224	56	8	1	0.38	0.47	0.29	38%	Number of distinct bo	oks in which topic app
11204	gender	1.00	1.00	146	63	9	1	0.25	0.53	0.32	37%	0	28
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	35%	Number of relations	
277	globalization	1.00	1.00	25	97	G	1	0.04	0.82	0.21	35%	0	505
3131	World War I	1.00	1.00	25	59	15	3	0.04	0.50	0.54	35%	Ğ	
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%	Number of words in to	opic name
17941	Kent Immanue	1.00	1.00	13	67	12	2	0.02	0.57	0.43	34%	4	26
871	racism	1.00	1.00	98	62	9	1	0.17	0.53	0.32	34%	ā	
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%		
8541	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	0.08	0.04	31%		
18416	Law, the	1.00	1.00	360	15	5	2	0.62	0.13	0.18	31%		
7777	state, the	1.00	1.00	368	26	2	2	0.63	0.22	0.07	31%		

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









Create new topic

✓ ✓ Batman (100) Batman in film (100)

Search for match.

Batman



9.

1049

http://nyu.infolcom.nyc/basket/1049 Batman

Q1058914 1

Q1060829 1

Q1061197 1

O106259 4



0.52

0.67

III. A PRACTICAL APPROACH TO TOPIC CURATION

Ingest - Curate - Publish

INGEST - CURATE - PUBLISH

- Ingest Information Automated with customized configuration
- Curate Information Manual with User interface
- Publish Information Automated with customized configuration

HUMAN CURATION

- Merge (Synonyms)
- Detach (Homonyms)
- Categorize (Taxonomy)
- Connect with controlled vocabularies
- Relate Topics
- Map Topics to Sources

EXAMPLE OF CURATED TOPICS

- A name does not uniquely identify a topic.
- Example: "New York, New York"

10 names found

Movie New York, New York

New York

New York

New York

New York City

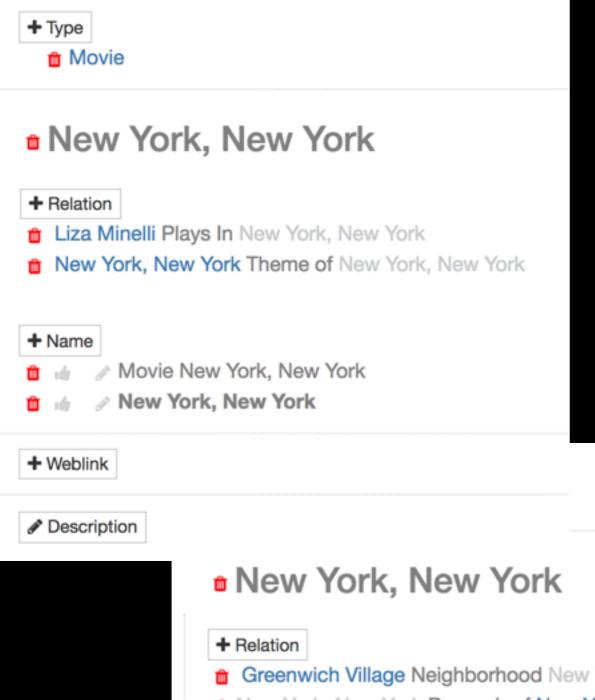
New York, New York

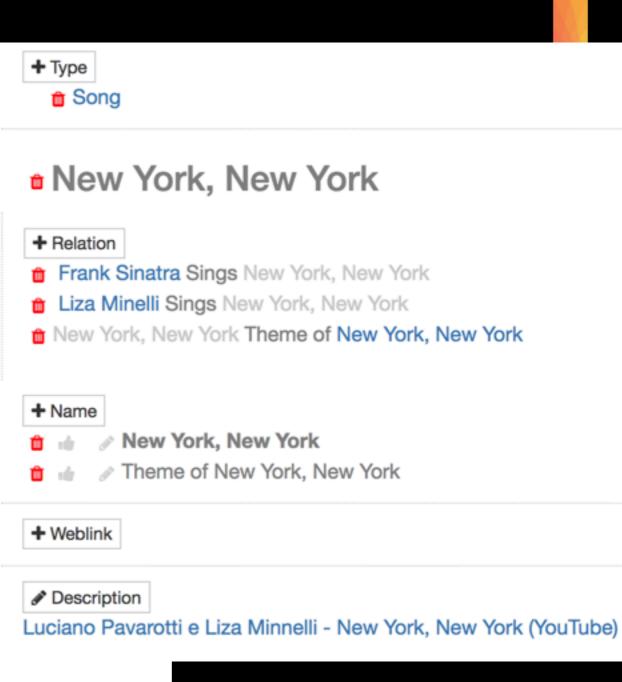
New York, New York

New York, New York

New York State

Theme of New York, New York





- m New York, New York Borough of New York
- Upper East Side Neighborhood New York, New York
- Upper West Side Neighborhood New York, New York
- Name Manhattan New York
- New York, New York

THE TAKEAWAY

- Curation is optional.
 It improves quality and accuracy.
- Curation is human intelligence used to improve results provided by artificial intelligence.



Q & A

MICHEL BIEZUNSKI MB@INFOLOOM.COM