Herausforderung Diskriminierung durch KI-Systeme: (Nicht nur) technische Ursachen und Lösungsansätze

Bettina Berendt

TU Berlin, Weizenbaum Institut für die Vernetzte Gesellschaft, und KU Leuven

Folien, Literatur und Bilder : https://people.cs.kuleuven.be/~bettina.berendt/Talks/berendt_2021_12_01.pdf

#vBIB21

1.12.2021

Was ist (algorithmische) Diskriminierung?

Computing and Information	Bibliographies	Library and Information Sciences	Dictionaries and Encyclopedias	Special Topics	Periodicals
116,597	21,600	25,784	10,009	565	3,842
000	001	002	003	004	005
General Works	Knowledge	History of the book	Systems Theory	Computer science	Computer programming, programs, data, security
791	15,071	1,294	2,816	26,002	50,258
001.0	001.1	001.2	001.3	001.4	001.5
Knowledge -- Subdivisions	Intellectual Life	Scholarship And Learning	Humanities	Research	[formerly : Cybernetics & related disciplines] [formerly : Information & communication]
559	107	337	541	3,038	1,318
001.50	001.51	001.52	001.53	001.54	001.55
	[formerly : Communication]		[formerly : Cybernetics]	[formerly : Communication through language]	[formerly : Communication through records]
20	304		329	207	107
001.530	001.531	001.532	001.533	001.534	001.535
		[formerly : Bionics]	[formerly : Self-organizing systems]	[formerly : Perception theory]	[formerly : Artificial intelligence]

3	4	5	6	7	8	9
Social Sciences	Language	Natural sciences and mathematics	Technology and Application of Knowledge	Arts and Recreation	Literature	History and Geography
1,108,788	170,451	362,953	804,760	803,871	1,146,568	674,630
33	34	35	36	37	38	39
Economics	Law	Public Administration, Military Science	Social problems and services; associations	Education	Commerce, Communications, Transportation	Customs, Etiquette, Folklore
192,492	126,967	53,635	140,489	145,691	42,269	59,919

300	301	302	303	304	305	306	307	308	309
Social Sciences	Sociology, Anthropology	Social Interaction	Social Processes	Factors affecting social behavior	Social groups ; Inequality	Culture and Institutions	Communities	[No longer used]	Social situation and conditions [No longer used]
11,801	26,824	21,053	23,248	9,713	55,521	56,710	11,521	762	3,244
305.0	305.1	305.2	305.3	305.4	305.5	305.6	305.7	305.8	305.9
		Age groups	By Gender	Women	Class	Religious groups	Language groups	Ethnic and national groups ; racism, multiculturalism	People by occupation and miscellaneous social statuses
375	6	8,162	3,244	14,633	4,745	1,039	60	17,822	2,236

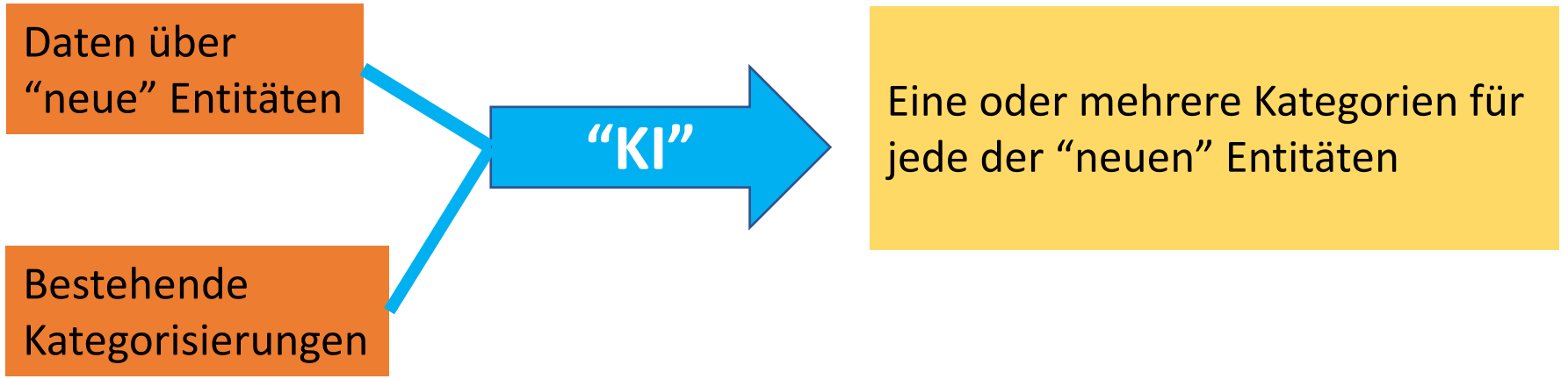
LibraryThing

Diskriminierung und Bias: Definitionen, Rechte, Computer

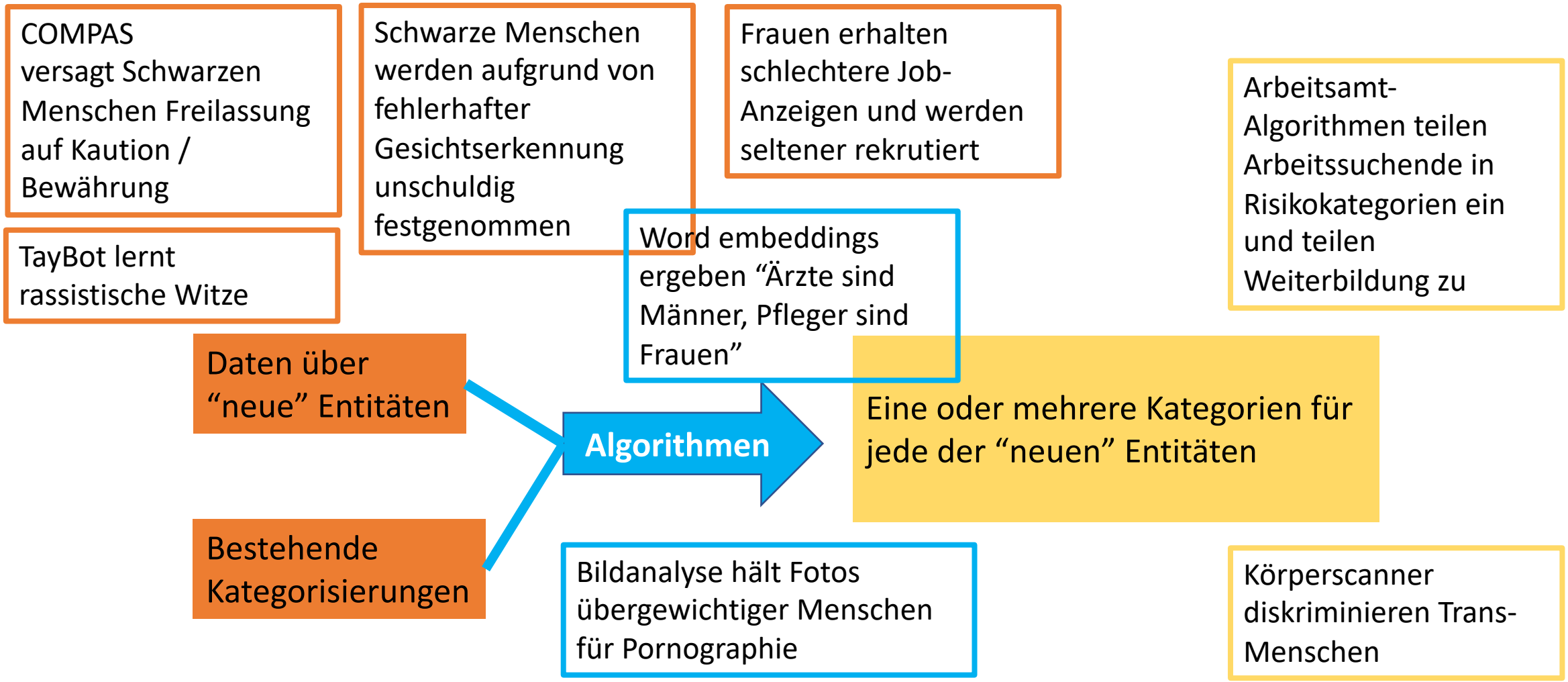
- Benachteiligungen aus Gründen der Rasse oder wegen der ethnischen Herkunft, des Geschlechts, der Religion oder Weltanschauung, einer Behinderung, des Alters oder der sexuellen Identität (§1 AGG)
- mittelbare Benachteiligung [:] dem Anschein nach neutrale Vorschriften, Kriterien oder Verfahren Personen [...] benachteiligen können, es sei denn, [...] rechtmäßiges Ziel sachlich gerechtfertigt und die Mittel [...] angemessen und erforderlich. (§3 (2) AGG)
- Niemand darf wegen seines Geschlechtes, seiner Abstammung, seiner Rasse, seiner Sprache, seiner Heimat und Herkunft, seines Glaubens, seiner religiösen oder politischen Anschauungen benachteiligt oder bevorzugt werden. Niemand darf wegen seiner Behinderung benachteiligt werden. (Art. 3 GG; s.a. Art. 21 EU-Charta)
- Der Genuß der in dieser Konvention anerkannten Rechte und Freiheiten ist ohne Diskriminierung insbesondere wegen [EN: such as] des Geschlechts, der Rasse, der Hautfarbe, der Sprache, der Religion, der politischen oder sonstigen Anschauung, der nationalen oder sozialen Herkunft, der Zugehörigkeit zu einer nationalen Minderheit, des Vermögens, der Geburt oder eines sonstigen Status zu gewährleisten. (Art. 14 EMRK)
- “A system discriminates unfairly if it denies an opportunity or a good or if it assigns an undesirable outcome to an individual or group of individuals on grounds that are unreasonable or inappropriate.”
- A computer system has bias if it “systematically and unfairly discriminates against certain individuals or groups of individuals in favor of others”.

(Friedman and Nissenbaum, 1996)

Komponenten maschinellen Lernens / Kategorisierung



Beispiele von Diskriminierung, die sich aus den Komponenten ergeben *(vereinfachte Zuordnung)*

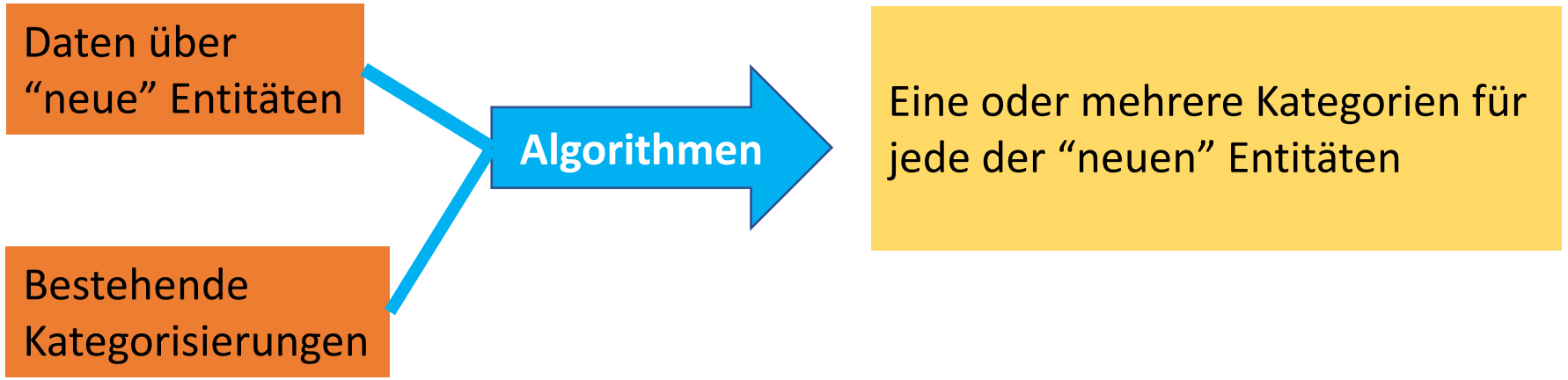


Gegenmaßnahmen aus der Informatik: Lösung, Techno-Fix, oder …?

Pre-processing:
z.B. *data sanitization*, repräsentativere Datensets

In-processing:
z.B. *de-biasing*

Post-processing:
z.B. Filtern von Resultaten



Betrifft das auch Bibliotheken?

Beispiel Automatische Inhaltserschließung: Lernen von Mappings: Text → Kategorien



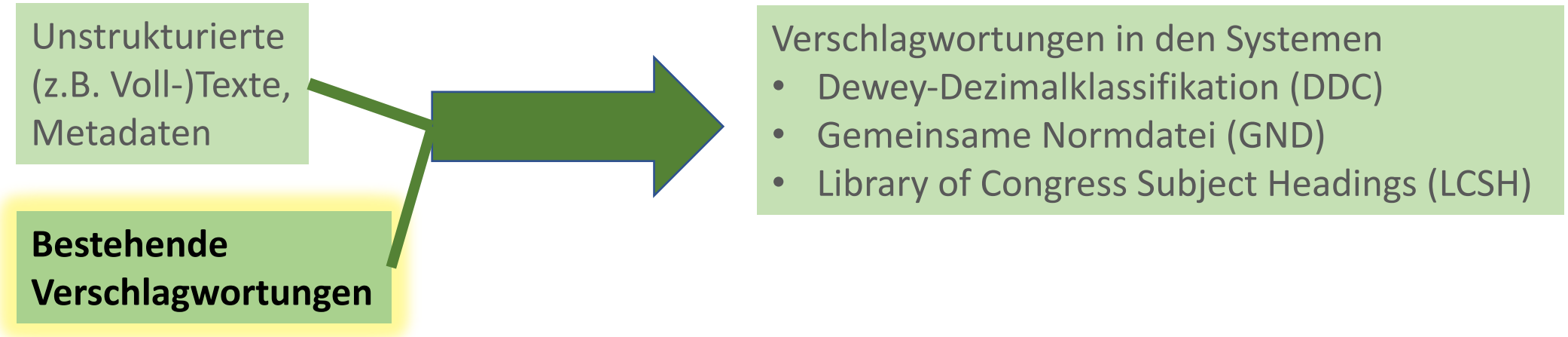
Unstrukturierte (z.B. Voll-)Texte, Metadaten

Bestehende Verschlagwortungen

Verschlagwortungen in den Systemen
- Dewey-Dezimalklassifikation (DDC)
- Gemeinsame Normdatei (GND)
- Library of Congress Subject Headings (LCSH)

Basierend auf https://wiki.dnb.de/display/FNMVE/Netzwerk+maschinelle+Verfahren+in+der+Erschliessung
Details zu aktuellen technischen Entwicklungen: https://wiki.dnb.de/display/FNMVE/Fachtagung+2021

Herausforderung 1



Probleme: altbekannt und doch hartnäckig …



All of the libraries that Porter consulted for guidance relied on the Dewey Decimal Classification. *“Now in [that] system, they had one number—326—that meant slavery, and they had one other number—325, as I recall it—that meant colonization,” she explained in her oral history. In many “white libraries,” she continued, “every book, whether it was a book of poems by James Weldon Johnson, who everyone knew was a black poet, went under 325. And that was stupid to me.”*

(Dorothy Porter, on the motivation for her re-cataloguing work starting in the late 1930s, reported by Nunes, 2018)

“Books that cover the history of the civil rights movement, immigrant histories, and women’s history were getting sent to the 300s [Social Science]”, not to the 900s [History]

(Jess de Courcy Hines on the situation in 2018, reported by Schwartz, 2018)

Herausforderung 2



Ein konstruiertes/fiktives, aber doch realistische Probleme demonstrierendes Bsp. – (1) die gute Nachricht

Title: Artificial intelligence and natural man

Author: Boden, Margaret A.

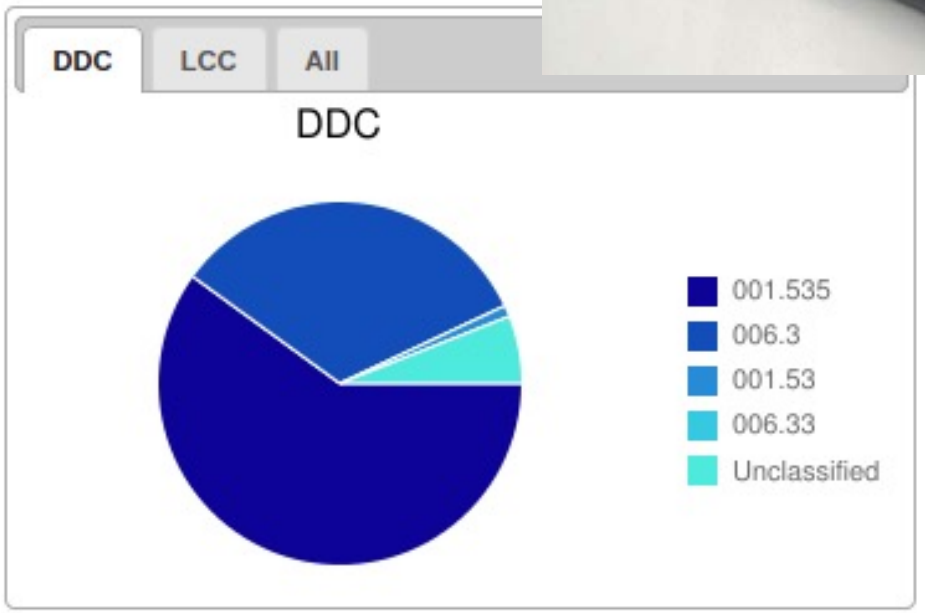
Formats: **Editions:** 60 **Total Holdings:** 1521

OCLC Work Id: 5898690

Record Link: http://classify.oclc.org/classify2/ClassifyDemo?owi=5898690



DDC:	Class Number	Holdings	Links
Most Frequent	001.535	918	Web Dewey
Edition: 22	006.33	6	
LCC:	**Class Number**	**Holdings**	**Links**
Most Frequent	Q335	1400	ClassWeb



Konstruiertes Bsp. – (2) Mögliche Effekte von vortrainierten Sprach-modellen

ingly written descriptive details to find them. Dr Boden assumes a minimum of background knowledge in her readers, and displays admirable gifts as a teacher, especially in her choice of similes and worked examples. I particularly enjoyed her comparison of a computer programme with a knitting pattern!

Text Synth

Model: GPT-J 6B top-k: 40 top-p: 0.9 temperature: 1.0 seed: 0

Select an example

Margaret Boden compares computer programs to knitting patterns

Completed Text:

Margaret Boden compares computer programs to knitting patterns and then uses them to create a novel.

If you've ever knitted a scarf, you'll understand what Margaret Boden is talking about. A simple pattern, often with just a few stitches, can turn out to be a completely different thing when knitted.

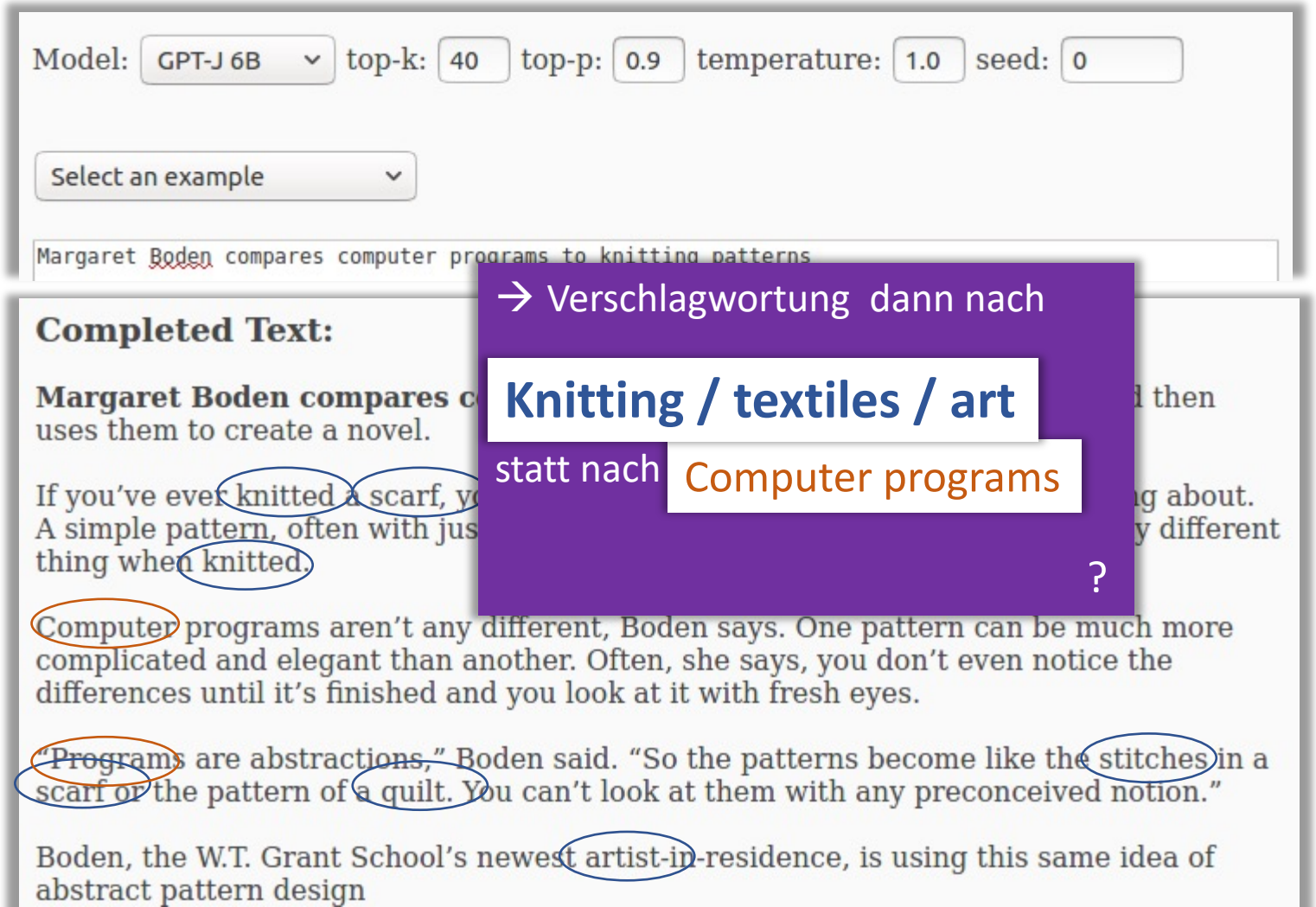
Computer programs aren't any different, Boden says. One pattern can be much more complicated and elegant than another. Often, she says, you don't even notice the differences until it's finished and you look at it with fresh eyes.

"Programs are abstractions," Boden said. "So the patterns become like the stitches in a scarf or the pattern of a quilt. You can't look at them with any preconceived notion."

Boden, the W.T. Grant School's newest artist-in-residence, is using this same idea of abstract pattern design

Konstruiertes Bsp. – (2) Mögliche Effekte von vortrainierten Sprach-modellen

ingly written descriptive details to find them. Dr Boden assumes a minimum of background knowledge in her readers, and displays admirable gifts as a teacher, especially in her choice of similes and worked examples. I particularly enjoyed her comparison of a computer programme with a knitting pattern!



→ Verschlagwortung dann nach

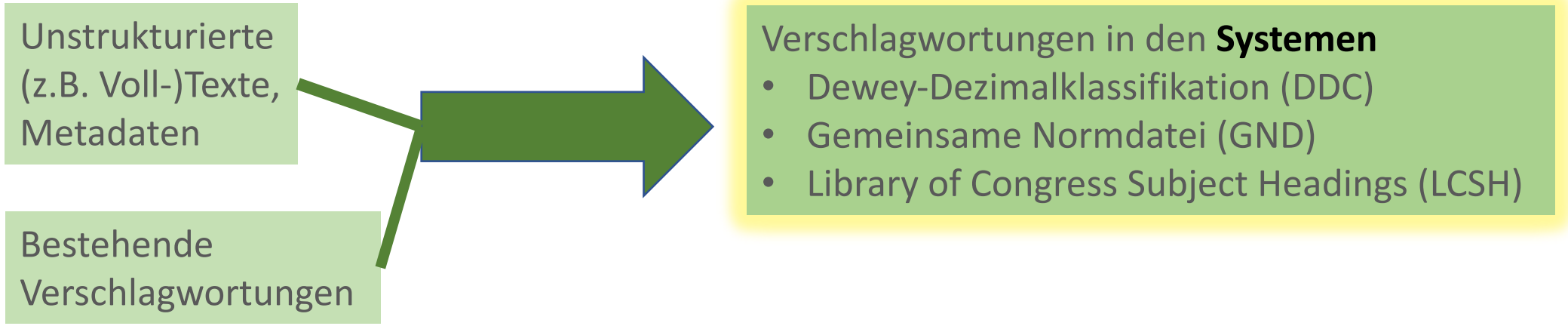
Knitting / textiles / art

statt nach Computer programs

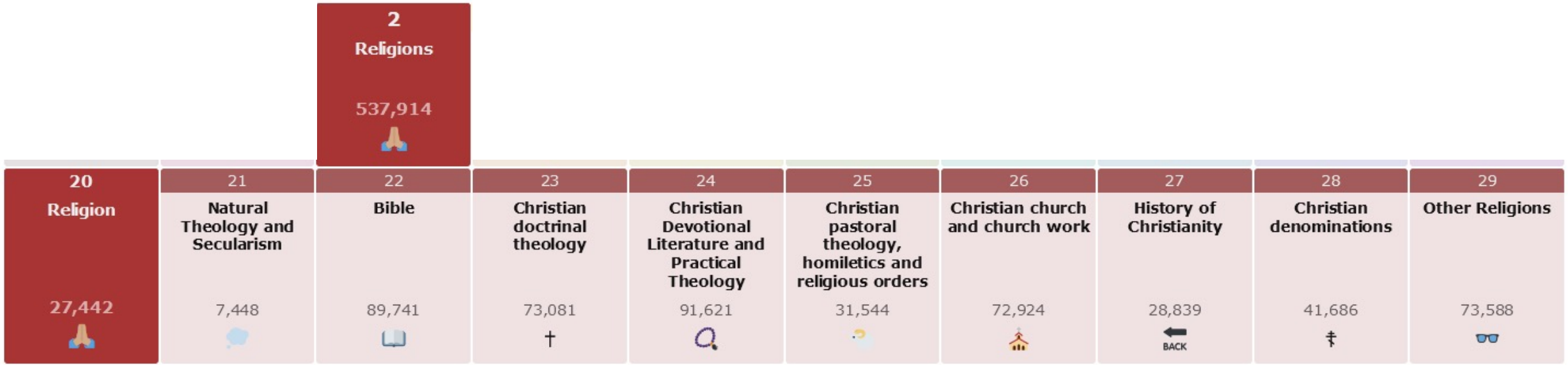
?

Text Synth

Herausforderung 3



Weitere Probleme von DDC u.a. Ontologien (1): Unterschiedliche “Dichte”



LibraryThing

Weitere Probleme von DDC u.a. Ontologien (2): Kategorien und ihre Baumstruktur

- Bsp. LGBTQ+:
  - ‘Abnormal psychology’ in 1932
  - ‘Social problems’ in 1989
  - *Today? You’ll find it at* ‘306.7 — Sexual orientation, transgenderism, intersexuality’

  To be clear, the current classification is an improvement but… I mean… LGBTQ+ people are [Groups of] PEOPLE (aka: 305), so finding LGBTQ+ related content "between prostitution and child trafficking on one side and fetishes and BDSM on the other” feels truly wrong. We can do better. (Care Harder, n.d. / O’Hara, 2018)
    - Anmerkung: Die Kategorien sind 2021 etwas, aber nicht substanziell anders
- Ontologien: “formal, explicit specification of shared conceptualisations”
- Ohne geht es nicht.
- Können Ontologien, die > 1 Schlagwort zuordnen, diese Probleme lösen?

Leider nein.

Wer definiert eigentlich, was diskriminierend ist? Und wie können relevante Stakeholder miteinbezogen werden?

LIBRARY OF CONGRESS SUBJECT HEADING CHANGE DOESN'T ADDRESS THE REAL ISSUE

Kelly Jensen | Nov 15, 2021

Last week, the Library of Congress (LOC) made a change celebrated by a wide range of organizations, including the American Library Association. After years of pushing to make changes to the cataloging subject headings "aliens" and "illegal aliens," the LOC replaced them with the terms "noncitizen" and "illegal immigration." The decision has been discussed since at least 2016, when Congress's conservative politicians intervened and determined the headings would stay as-is.

#DropTheIWord campaign specifically focuses on the ways media undermines progress by continuing to refer to immigrants, asylum seekers, and other undocumented individuals as anything less than a person. The Right began using the term "illegal" to topple immigration reform, and the continued use of the word has only led to more harm and marginalization by undocumented individuals.

The LOC subject change may remove the word "alien," but it retains the use of "illegal." The term is not only dehumanizing, it's inaccurate and it's racially charged.

Herausforderung 4



Unstrukturierte (z.B. Voll-)Texte, Metadaten

Bestehende Verschlagwortungen

Verschlagwortungen in den Systemen

- Dewey-Dezimalklassifikation (DDC)
- Gemeinsame Normdatei (GND)
- Library of Congress Subject Headings (LCSH)



Risiken der “Popularität” als Basis von Empfehlungen in der Interaktion mit Nutzern



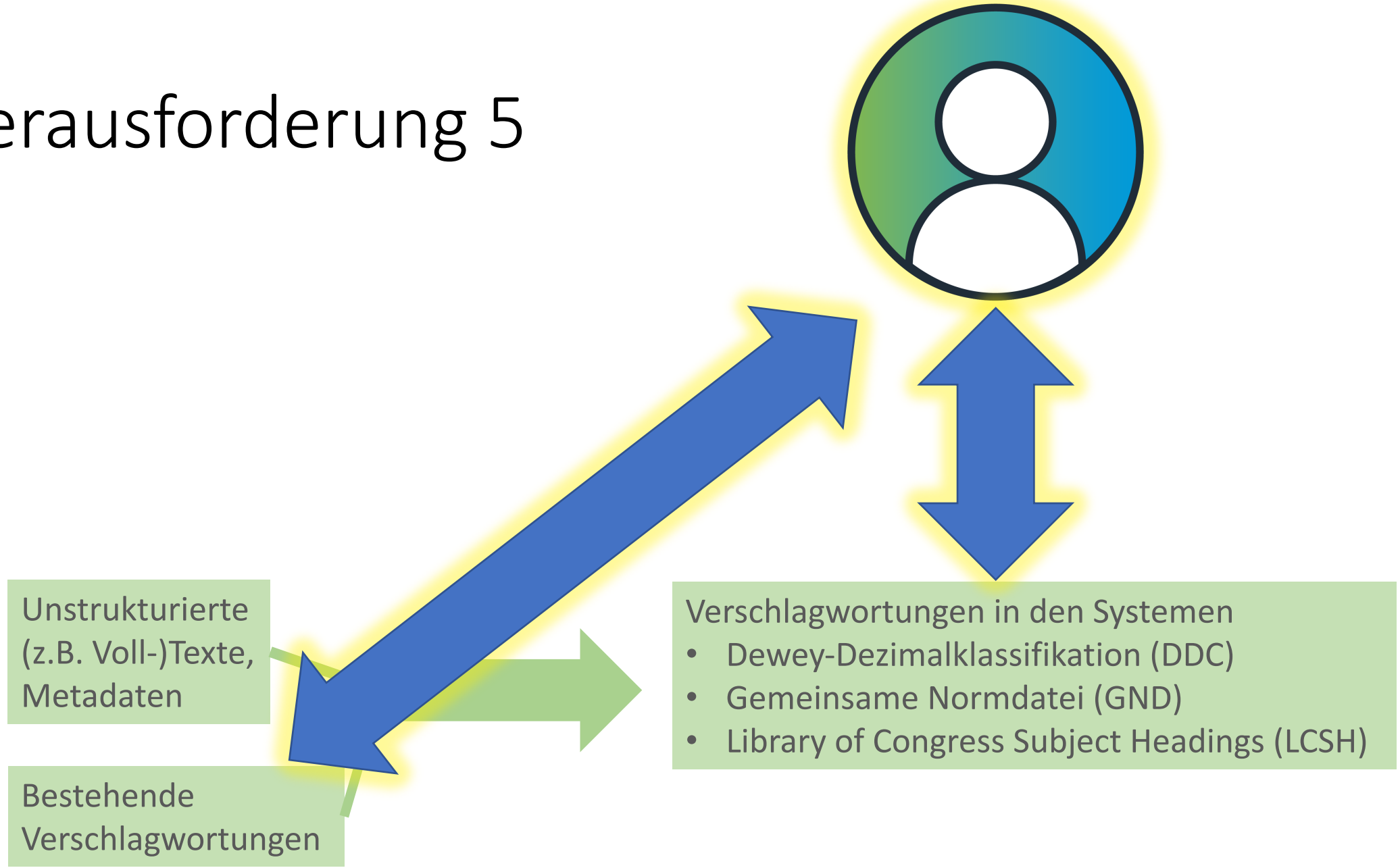


Diskriminerung vulnerabler Gruppen durch:

- Aussagen über sie
- Vorenthalten “besserer” Information

Risiken der “Popularität” als Basis von Empfehlungen in der Interaktion mit Nutzern

Herausforderung 5



Partizipatives Design

The Dewey Decimal Classification needs you!



Violet Fox • 21 May 2019 • Cataloging



The editors of the Dewey Decimal Classification system have always collaborated with librarians to ensure the classification is up to date. Today, we're excited to share changes that are making the editorial work on Dewey more transparent, inclusive, and responsive to community needs—and we need your help!

025.431: The Dewey blog

Everything you always wanted to know about the Dewey Decimal Classification® system but were afraid to ask ...

ENHAN

ARCHIVES

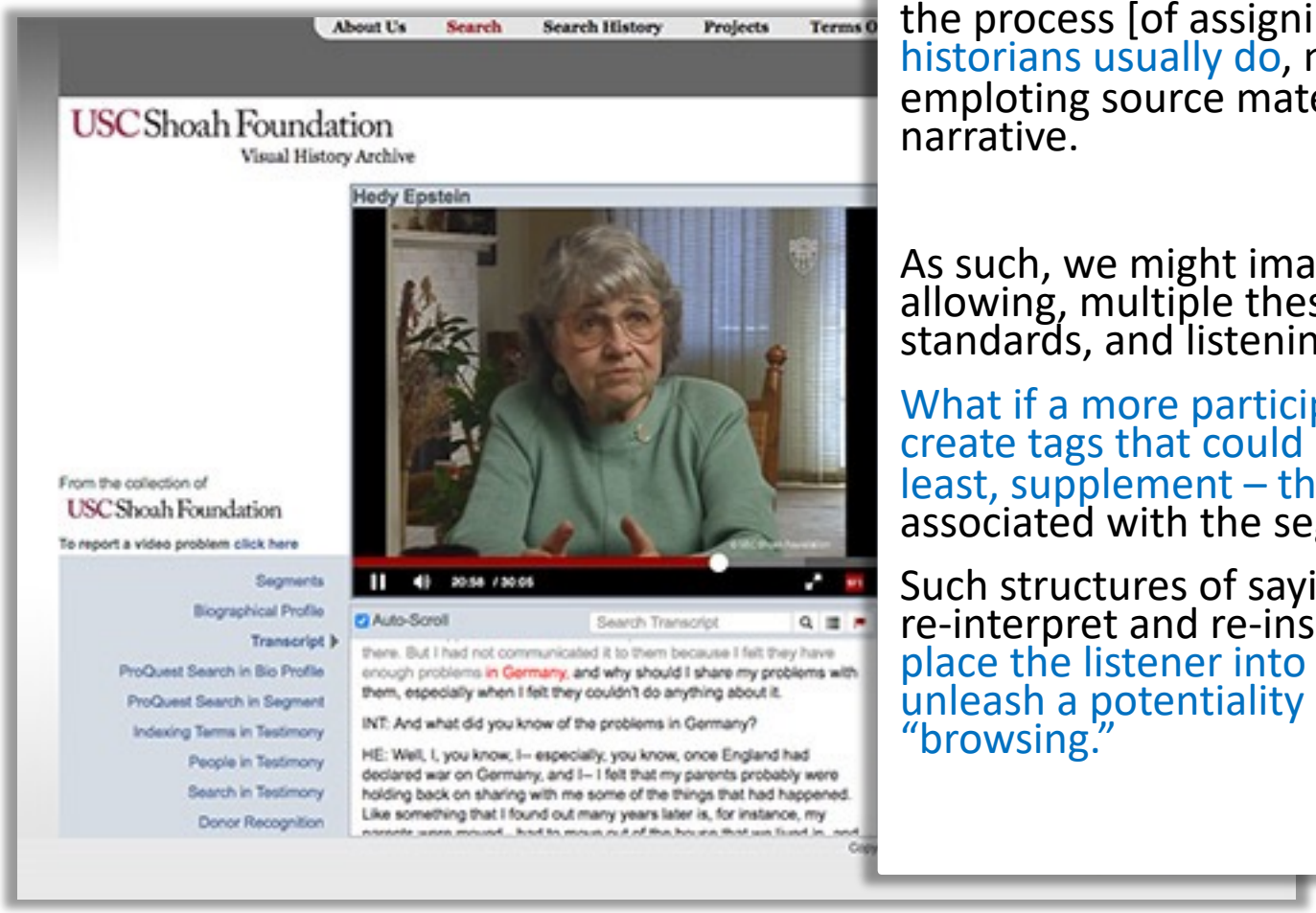
« What is an easy way to determine whether adding a standard subdivision to a specific
standard subdivision? | Main | Mapping Dewe

10 January 2019

Local option for classifying works about Indigenous peoples

"To be clear, libraries using Dewey are *always* welcome to implement local solutions to areas where standard notation does not serve their users well."

Partizipatives Design, (fach)kulturelle Praktiken, und Prosumers? – und: KI kann so viel mehr!



the process [of assigning index terms] is exactly the opposite of what historians usually do, namely to create narratives from data by emploting source material, evidence, and established facts into a narrative.

As such, we might imagine how a fluid data ontology might work, by allowing, multiple thesauruses that recognize a range of knowledge, standards, and listening practices […]

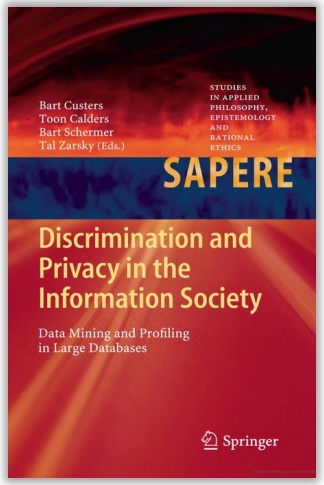
What if a more participatory architecture allowed for other listeners to create tags that could unsay the said, or in other words, undo – or, at least, supplement – the definitive indexing categories and keywords associated with the segmented testimonies? […]

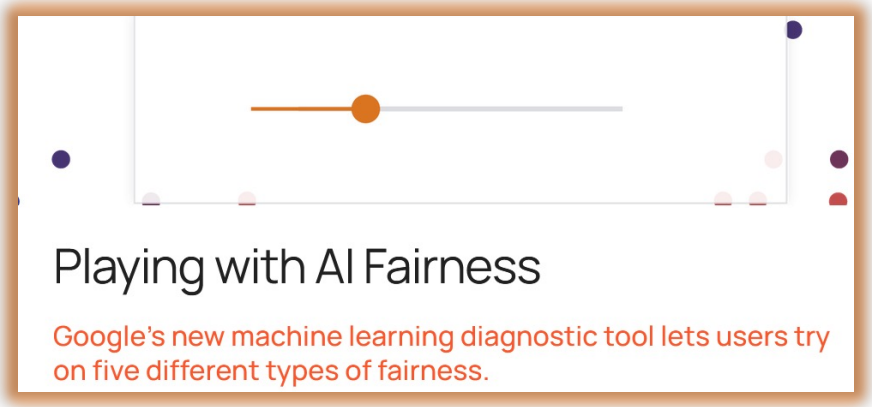
Such structures of saying and unsaying the database would constantly re-interpret and re-inscribe the survivors' stories in ways that not only place the listener into an active relationship of responsibility but unleash a potentiality of meaning in every act of "saying" and “browsing.”

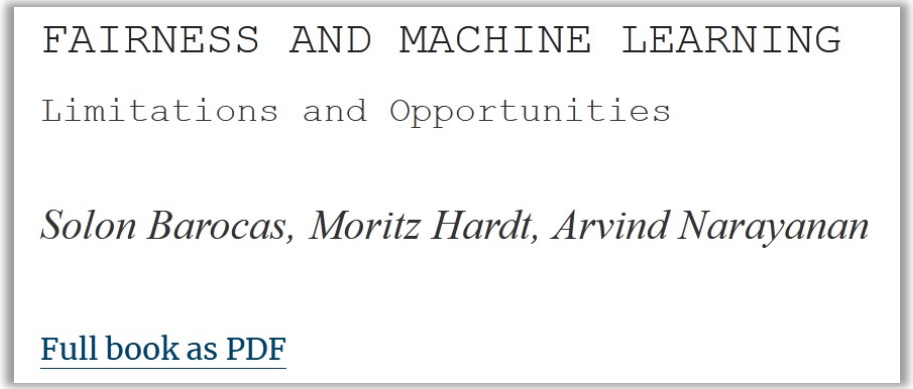
(Presner, 2016)

Was tun?

Was tun Forscher:innen, Entwickler:innen, Zivilgesellschaft, Gesetzgeber?







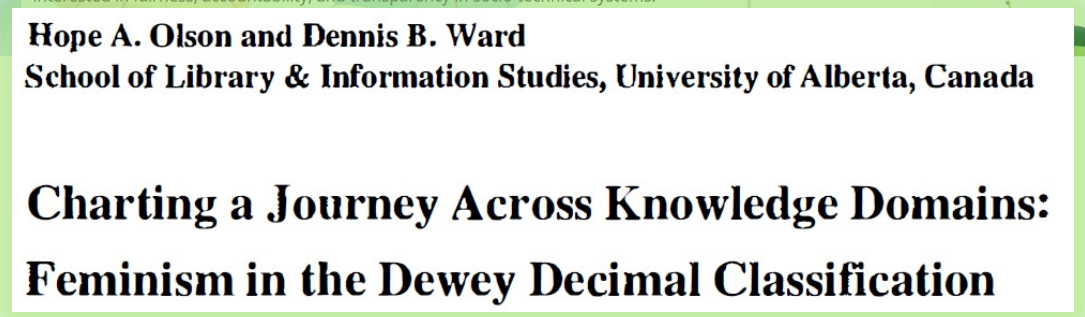
DSGVO

AI Act





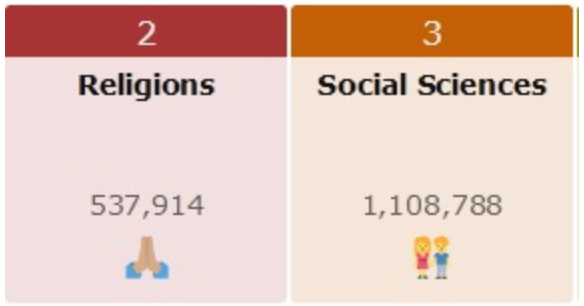
& https://facctconference.org/network/ & …



Was tun Bibliothekar:innen?



290:
Eid, Rosh Hashanah, Diwali



390:
Thanksgiving, Christmas, Easter

Howard's [a librarian at Fairhill Elementary School in Fairfax, VA] school has a diverse population, and she wants every student in the building who celebrates a religious holiday to find books about all of them in the same place.

The pandemic gave her the flexibility to devote time to the project this year, since the district allowed her to tackle tasks not related to student achievement.

(Joseph, 2021)

Computer kamen gerade rechtzeitig,
um soziale Erfindungen unnötig erscheinen zu lassen.

nach Joseph Weizenbaum, 1985

Algorithmische Diskriminierung
(+ ihre Techno-Fixes)

Computer kamen gerade rechtzeitig,

um soziale Erfindungen unnötig erscheinen zu lassen.

DANKE !

Literatur (1)

Überblicksdarstellung, auch hinsichtlich der Beziehungen zum Recht, mit ausführlicher Bibliographie:

Berendt, B. (in press). Algorithmic discrimination. To appear in G. Comandè (Ed.), *Encyclopedia of Law and Data Science.* Edward Elgar Publishing Ltd. (last author version: https://people.cs.kuleuven.be/~bettina.berendt/Papers/Algorithmic Discrimination BB 2020 10 13.pdf)

In diesen Folien zitierte Literatur

Solon Barocas, Moritz Hardt and Arvind Narayanan (2019). Fairness and Machine Learning. http://www.fairmlbook.org

ben-Aaron, D. (1985). Weizenbaum examines computers and society. The Tech 105(16). April 9, 1985. http://tech.mit.edu/V105/N16/weisen.16n.html

Burell, S. (2021). Die Diskriminierung steckt oft im Detail. https://www.spiegel.de/start/stellenanzeigen-werden-oft-fuer-maenner-formuliert-wie-frauen-trotzdem-den-job-bekommen-a-2ff0215c-009c-48b1-b045-a462ca808cd7

Care Harder (n.d.). Something is rotten in the Dewey Decimal system. https://www.careharder.com/blog/systemic-injustice-in-the-dewey-decimal-system

Custers, B., Calders, T., Schermer, B., Zarsky, T. (Eds.) (2013). Discrimination and Privacy in the Information Society. Data Mining and Profiling in Large Databases. Springer

Dustin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. Reuters. Oct 11, 2018. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. ACM Transactions on Information Systems 14, 3 (July 1996), 330–347. https://doi.org/10.1145/230538.230561

Literatur (2)

Joseph, C. (2021). Move Over, Melvil! Momentum Grows to Eliminate Bias and Racism in the 145-year-old Dewey Decimal System. School Library Journal, Aug 18, 2021. https://www.slj.com/?detailStory=move-over-melvil-momentum-grows-to-eliminate-bias-and-racism-in-the-145-year-old-dewey-decimal-system

Nunes, Z.C. (2018). Remembering the Howard University Librarian Who Decolonized the Way Books Were Catalogued. Smithsonian Magazine, November 26, 2018. https://www.smithsonianmag.com/history/remembering-howard-university-librarian-who-decolonized-way-books-were-catalogued-180970890/

O'Hara, M. (2018). Bad Dewey. Goldsmiths University of London library blog. https://sites.gold.ac.uk/library-blog/bad-dewey/

Olson, H. A., Ward, D. B. (1998). Charting a journey across knowledge domains: Feminism in the Dewey Decimal Classification. In W. Mustafa el-Hadi, J. Maniez, A.S. Pollitt (Eds.), Structure and Relations in Knowledge Organization: Proceedings of the Fifth International ISKO Conference, 238–244. Würzburg: ERGON Verlag. https://www.ergon-verlag.de/isko_ko/downloads/aikovol06199833.pdf

Presner, T. (2016) “The Ethics of the Algorithm: Close and Distant Listening to the Shoah Foundation Visual History Archive.” Probing the Ethics of Holocaust Culture. Eds. Fogu, Kansteiner, and Presner. Kindle Edition. Cambridge: Harvard University Press. 175-202. cited from http://www.toddpresner.com/wp-content/uploads/2014/12/Presner_Ethics_Dec2014.pdf

Schwartz, M. (2018). Classifying books, classifying people. Medium. May 3, 2018. https://medium.com/the-bytegeist-blog/classifying-books-classifying-people-302430282a3d

Das Zitat im Original

Computer kamen gerade rechtzeitig,
um soziale Erfindungen unnötig erscheinen zu lassen.

What the coming of the computer did, "just in time," was to make it unnecessary to create social inventions, to change the system in any way. So in that sense, the computer has acted as fundamentally a conservative force, a force which kept power or even solidified power where is already existed.

(ben-Aaron, 1985, interview with J. Weizenbaum)

Abbildungen und Screenshots

https://algorithmwatch.org/de/automatisierte-diskriminierung-facebook-verwendet-grobe-stereotypen-um-die-anzeigenschaltung-zu-optimieren/

https://www.nytimes.com/2021/06/30/technology/artificial-intelligence-bias.html

https://www.dnb.de/DE/Professionell/ProjekteKooperationen/Projekte/KI/KI.html

https://wiki.dnb.de/display/FNMVE/Netzwerk+maschinelle+Verfahren+in+der+Erschliessung

https://www.journals.uchicago.edu/doi/abs/10.1093/bjps/29.4.394?journalCode=bjps

https://bellard.org/textsynth

https://ddc.typepad.com/025431/2019/11/why-are-the-200s-so-heavily-focused-on-christianity-an-explainer.html

https://bookriot.com/library-of-congress-subject-heading-change/

https://blog.cpanel.com/feature-requests-and-the-user-story/

https://journal.code4lib.org/articles/3022

https://www.unwomen.org/en/news/stories/2013/10/women-should-ads

https://blog.oclc.org/next/the-dewey-decimal-classification-needs-you/

https://ddc.typepad.com/025431/2019/01/local-option-for-classifying-works-about-indigenous-peoples.html

https://sfi.usc.edu/news/2017/09/17961-nearly-1000-english-transcripts-added-visual-history-archive

https://www.facctconference.org

https://pair-code.github.io/what-if-tool/ai-fairness.html

Weitere: s. Logos und URLs auf den Folien