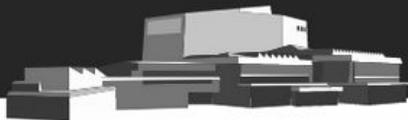




**Staatsbibliothek
zu Berlin**
Preußischer Kulturbesitz



Mensch.Maschine.Kultur

Welche Chancen und Risiken bringt
Künstliche Intelligenz für das kulturelle Erbe?

Clemens Neudecker | Staatsbibliothek zu Berlin - Preußischer Kulturbesitz
HAW Hamburg Ringvorlesung "Bibs & Bits" | 9. Januar 2024

Staatsbibliothek zu Berlin - Preußischer Kulturbesitz

- gehört zur Stiftung Preußischer Kulturbesitz (SPK)
- ist an zwei Standorten in Berlin an 7 Tagen die Woche kostenlos für die Benutzung geöffnet
- sammelt seit 1661 wissenschaftlich relevante Literatur aus allen Sprachen, Zeiten und Ländern
- besitzt Hauptbestand mit ca. 12 Mio. Büchern, jährlicher Zuwachs von etwa 100,000 Titeln
- **Digitalisierte Sammlungen** geben Zugang zu über 210,000 digitalisierten Dokumenten, mit Public Domain Lizenz und via IIF API
- **Stabi Lab** für Experimente, Datensets, Digital Humanities

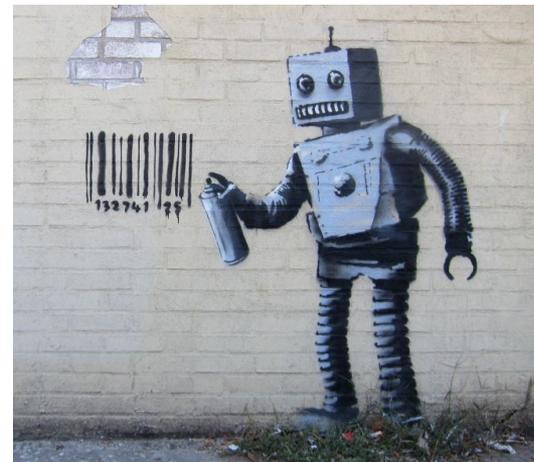


**Staatsbibliothek
zu Berlin**
Preußischer Kulturbesitz



KI (bzw. Machine Learning) in der Stabi Berlin

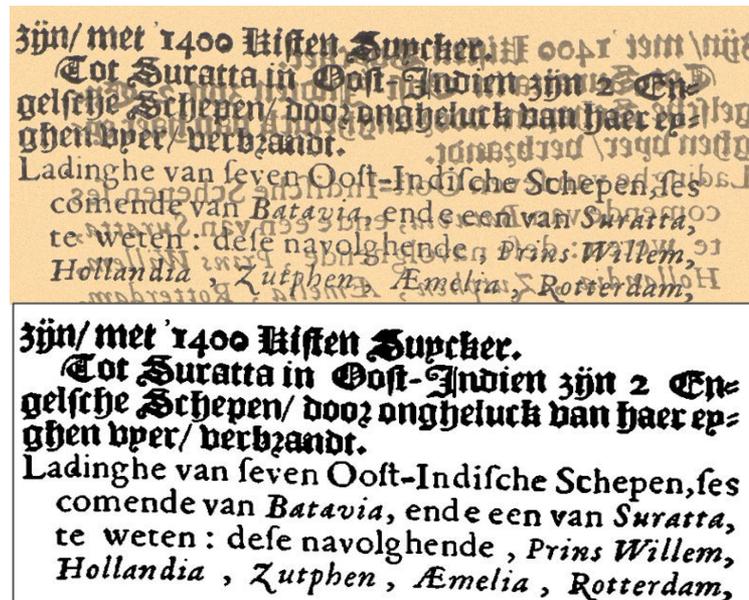
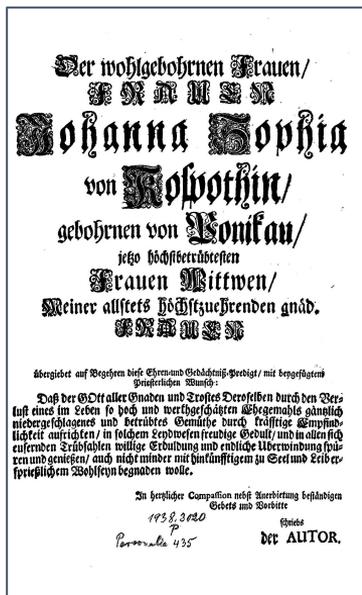
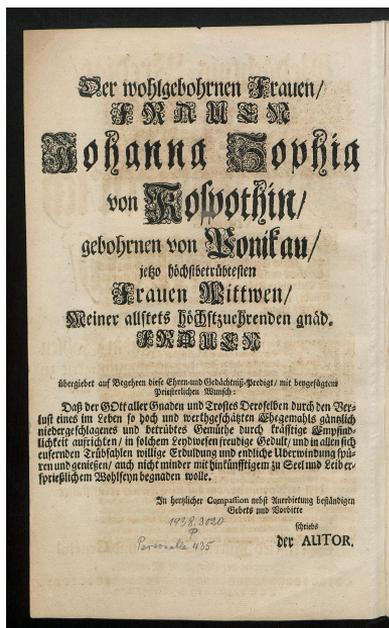
- **Drittmittelprojekte**
 - 2016—2024: **OCR-D** (DFG)
 - 2018—2021: **Qurator** (BMBF)
 - 2022—2025: **Mensch.Maschine.Kultur** (BKM)
- **Team**
 - 6x 100% Machine Learning Engineer
 - 2x 50% Bibliothekswissenschaftler:in
 - 1x 100% Forschungsdatenmanager:in
- **Hardware**
 - GPU Server mit 4x NVIDIA Tesla GPUs V100/A100
- **Ziele**
 - Erforschung und Entwicklung von Open Source Technologien für historische Kulturdaten
 - Durchsuchbarkeit (Texte und Bilder) sowie Erschließung und Strukturierung der Inhalte
 - Bereitstellung als offene und maschinenlesbare Daten (“Collections as Data”)
 - Ethisch, rechtlich, sozial verantwortungsvoller Umgang mit problematischen Inhalten und KI



CC-BY-SA Scott Lynch via [Flickr](#)

Bildoptimierung und Binarisierung

- *A hybrid CNN-Transformer model for Historical Document Image Binarization*, 2023.
<https://doi.org/10.1145/3604951.3605508> | https://github.com/qurator-spk/sbb_binarization



Bildähnlichkeitssuche und Text-Bild-Suche

- *Gauging the Limitations of Natural Language Supervised Text-Image Metrics Learning by Iconclass Visual Concepts, 2023.*

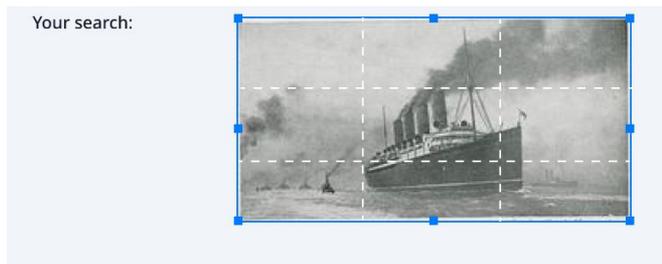
<https://doi.org/10.1145/3604951.3605516> | https://github.com/qurator-spk/sbb_images

Description Search Results

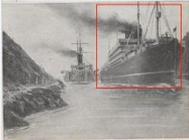
Image Description PPN

Big ship entering port

Enter image description. **English only!**

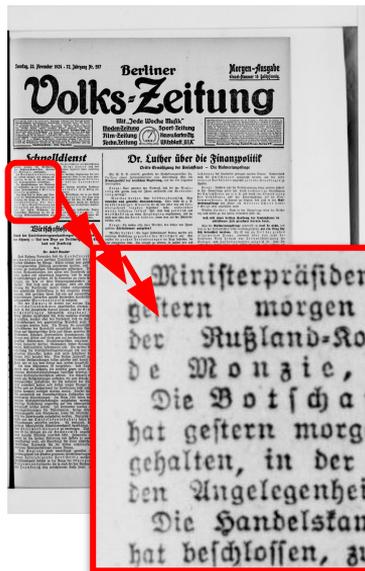


Images matching the description you entered:

<p style="text-align: center;">Search Similar Images</p> <p style="text-align: center;">View in Digitized Collections</p> 	<p style="text-align: center;">Search Similar Images</p> <p style="text-align: center;">View in Digitized Collections</p> 	<p style="text-align: center;">Search Similar Images</p> <p style="text-align: center;">View in Digitized Collections</p> 	<p style="text-align: center;">Search Similar Images</p> <p style="text-align: center;">View in Digitized Collections</p> 	<p style="text-align: center;">Search Similar Images</p> <p style="text-align: center;">View in Digitized Collections</p> 	<p style="text-align: center;">Search Similar Images</p> <p style="text-align: center;">View in Digitized Collections</p> 
---	--	--	--	--	--

Texterkennung (OCR)

- *OCR-D: An end-to-end open source OCR framework for historical printed documents, 2019.*
<https://doi.org/10.1145/3322905.3322917> | <https://github.com/OCR-D/core>



Ministerpräsident Herriot hat
gestern morgen den Vorsitzenden
der Rußland-Kommission, Senator
de Monzie, empfangen.
Die Botschafterkonferenz
hat gestern morgen eine Sitzung ab-
gehalten, in der sie sich mit laufen-
den Angelegenheiten beschäftigte.
Die Handelskammer von Bordeaux
hat beschlossen, zu der neuen franzö-

Ministerpräsident Herriot hat
Die Feierlichkeiten zur Ueberführung
geiern morgen den Vorsitzenden
der Autlanu-Kommission, Senator
de Monzie, empfangen.
Die Botschafterkonferenz
hat gestern morgen eine Sitzung ab-
gehalten, in der sie sich mit laufen-
den Angelegenheiten beschäftigte.
Die Handelskammer von Bordeaux
hat beschlossen, zu der neuen franzö-
sischen Inlandsanleihe 1 Million
Francs zu zeichnen.
Aus Genf sind in Sofia zwei Dele-
gierte der Balkerbundskommission zur
Prüfung der Frage der Massen-
auswanderung der bul-
garischen Bevölkerung aus
Thrazien und Mazedonien
und der letzten Beschwerde der
bulgarischen Regierung an die zu-
ständige Völkerbundskommission ein-
getroffen.

Ministerpräsident Herriot hat
gestern morgen den Vorsitzenden
der Rußland-Kommission, Senator
de Monzie, empfangen.
Die Botschafterkonferenz
hat gestern morgen eine Sitzung ab-
gehalten, in der sie sich mit laufen-
den Angelegenheiten beschäftigte.
Die Handelskammer von Bordeaux
hat beschlossen, zu der neuen franzö-
sischen Inlandsanleihe 1 Million
Francs zu zeichnen.
Aus Genf sind in Sofia zwei Dele-
gierte der Völkerbundskommission zur
Prüfung der Frage der Massen-
auswanderung der bul-
garischen Bevölkerung aus
Thrazien und Mazedonien
und der letzten Beschwerde der
bulgarischen Regierung an die zu-
ständige Völkerbundskommission ein-
getroffen.

OCR Evaluation und Qualitätssicherung

- Verlässliche Ermittlung der OCR Qualität (CER, WER, Reading Order) mit Ground Truth
<https://github.com/qurator-spk/dinglehopper>

Character differences

20

rath mit einer P[**o**]na ficali angelehen worden, und folche durch des Hrn. Graffen von Königsfeld Vorfrpuch, nur aus Gnaden nachgelaffen erhalten. Sonderm man hat auch diefen 4. Wochen lang alle Abend bey der Inquiltin gantz allein gelaffen.

Binnen welcher gantzer Zeit der Schreiber Bredekaw befändig bey Ihme gewefen, und fich in

der am 13 ten Octobr. a. c. in Judicio gegen feinen gewefenen Hrn. introducirer Appellation deffen Beyraths bedienet hat;

§. 33) Dabeneben lifr der Schreiber binnen diefer gantzen Zeit auf freym Fuß geblieben, und

hat nicht nur durch feinen Confulenten, fondem auch, weilen der Inquiltin felbften in Ihrem Gefängnüß

fo viele Freyheit gelaffen worden, daß fie fremden Beluch von Ihren Anverwandten ohngehendend empfangen können, durch andere Perfonen fich mit Ihr über alles, was Er oder fie dereinleiten zu fagen hat-

20

rath mit einer P[**o**]na ficali angelehen worden, und folche durch des Hrn. Graffen von Königsfeld Vorfrpuch, nur aus Gnaden nachgelaffen erhalten. Sonderm man hat auch diefen 4. Wochen lang alle Abend bey der Inquiltin gantz allein gelaffen.

Binnen welcher gantzer Zeit der Schreiber Bredekaw befändig bey Ihme gewefen, und fich in

der am 13 ten Octobr. a. c. in Judicio gegen feinen gewefenen Hrn. introducirer Appellation deffen Beyraths bedienet hat;

-23) Dabeneben lifr der Schreiber binnen diefer gantzen Zeit auf freym Fuß geblieben, und

hat nicht nur durch feinen Confulenten, fondem auch, weilen der Inquiltin felbften in Ihrem Gefängnüß

fo viele Freyheit gelaffen worden, daß fie fremden Beluch von Ihren Anverwandten ohngehendend empfangen können, durch andere Perfonen fich mit Ihr über alles, was Er oder fie dereinleiten zu fagen hat-

Clemens Neudecker, Karolina Zaczynska, Konstantin Baierer, Georg Rehm, Mike Gerber, Julián Moreno Schneider
Methoden und Metriken zur Messung von OCR-Qualität für die Kuratierung von Daten und Metadaten

1 Einleitung

Durch die systematische Digitalisierung der Bestände in Bibliotheken und Archiven hat die Verfügbarkeit von Bilddigitalisaten historischer Dokumente rasant zugenommen. Das hat zunächst konservatorische Gründe: Digitalisierte Dokumente lassen sich praktisch nach Belieben in hoher Qualität vervielfältigen und suchen. Darüber hinaus lässt sich mit einer digitalisierten Sammlung eine wesentlich höhere Reichweite erzielen, als das mit dem Präsenzbestand allein jemals möglich wäre. Mit der zunehmenden Verfügbarkeit digitaler Bibliotheks- und Archivbestände steigen jedoch auch die Ansprüche an deren Präsentation und Nutzbarkeit. Neben der Suche auf Basis bibliografischer Metadaten erwarten Nutzer:innen auch, dass sie die Inhalte von Dokumenten durchsuchen können.

Im wissenschaftlichen Bereich werden mit maschinellen, quantitativen Analysen von Textmaterial große Erwartungen an neue Möglichkeiten für die Forschung verbunden. Neben der Bilddigitalisierung wird daher immer häufiger auch eine Erfassung des Volltextes gefordert. Diese kann entweder manuell durch Transkription oder automatisiert mit Methoden der *Optical Character Recognition* (OCR) geschehen (Engl et al. 2020). Der manuellen Erfassung wird im Allgemeinen eine höhere Qualität der Zeichengenauigkeit zugeschrieben. Im Bereich der Massendigitalisierung fällt die Wahl aus Kostengründen jedoch meist auf automatische OCR-Verfahren.

Die Einrichtung eines massentauglichen und in Ergebnis qualitativ hochwertigen OCR Workflows stellt Bibliotheken und Archive vor hohe technische Herausforderungen, weshalb dieser Arbeitsschritt häufig an dienstleistende Unternehmen ausgelagert wird. Bedingt durch die Richtlinien für die Vergabepraxis und fehlende oder mangelhafte Richtlinien der digitalisierenden Einrichtungen bzw. zwischen Förderinstrumente führt dies jedoch zu einem hohen Grad an Heterogenität der Digitalisierungs- bzw. Textqualitäts sowie des Umfangs der strukturellen und semantischen Anreicherungen. Diese Heterogenität erschwert die Nachnutzung durch die Forschung, die neben einheitlichen

A survey of OCR evaluation tools and metrics

Clemens Neudecker
Konstantin Baierer
Mike Gerber
www.staatsbibliothek-berlin.de
Staatsbibliothek zu Berlin - Preussische Kulturbesitz
Berlin, Germany

Christian Clausner
Apostolos Antonacopoulos
Stefan Pletschacher
www.primaresearch.org
Pattern Recognition and Image Analysis Lab (PRIM)A
University of Salford
Greater Manchester, United Kingdom

ABSTRACT

The millions of pages of historical documents that are digitized in libraries are increasingly used in contexts that have more specific requirements for OCR quality than keyword search. How to comprehensively, efficiently and reliably assess the quality of OCR results against the background of mass digitization, when ground truth can only ever be produced for very small numbers? That to gaps in specifications, results from OCR evaluation tools can return different results, and due to differences in implementations, even commonly used error rates are often not directly comparable. OCR evaluation metrics and sampling methods are also not sufficient when they do not take into account the accuracy of layout analysis, since for advanced use cases like Natural Language Processing or the Digital Humanities, accurate layout analysis and detection of the reading order are crucial. We provide an overview of OCR evaluation metrics and tools, describe two advanced use cases for OCR results, and perform an OCR evaluation experiment with multiple evaluation tools and different metrics for two distant datasets. We analyse the differences and commonalities in light of the presented use cases and suggest areas for future work.

CCS CONCEPTS

• Applied computing → Optical character recognition; Document analysis; Graphics recognition and interpretation; • Information systems → Digital libraries and archives.

KEYWORDS

optical character recognition, evaluation, accuracy, metrics

ACM Reference Format:

Clemens Neudecker, Konstantin Baierer, Mike Gerber, Christian Clausner, Apostolos Antonacopoulos, and Stefan Pletschacher. 2023. A survey of OCR evaluation tools and metrics. In *The 4th International Workshop on Historical Document Imaging and Processing (HDIP '23)*, September 3–6, 2023, Louanoe, Switzerland. ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/364088>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyright for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permission from permissions@acm.org.

© 2023, Association for Computing Machinery.
ACM ISBN 978-1-60959-959-7, 978-1-60959-959-7.
<https://doi.org/10.1145/364088>

1 INTRODUCTION

The efficient, transparent and informative evaluation of the quality of the results of Optical Character Recognition (OCR) is challenging in multiple respects. Established methods require Ground Truth (GT) data to serve as a reference for the desired result quality. Against the background of mass digitization¹, where millions of pages of documents are digitized and OCRed, this is neither feasible nor affordable. Especially in the context of historical documents, the creation of GT requires specialised skills and is a far too time-consuming to perform on a sufficiently large scale.

A further difficulty lies in the fact that standards or established conventions that provide clear and uniform guidelines for the creation of GT for historical documents are only partially available. There remain various un- or under-specified cases that can occur when assessing OCR quality. Examples include ligatures that can be recognised either as individual codepoints or as a combination of codepoints, characters that cannot be represented by a single codepoint, the encoding of special characters² that are not included in the Unicode standard and for which extensions such as METE³ other treatments from the Private Use Area⁴ must be used, and the encoding of punctuation and spaces. The OCR-D Ground Truth Guidelines [5] are an attempt to mediate between the OCR community and the needs of (scholarly) users of OCR results and to establish according specifications and guidelines.

In summary, established procedures and metrics for GT-based quality assessment of OCR results do not provide satisfactory answers when it comes to some of the more detailed questions that arise for historical documents. In addition, the extensive GT-based evaluation of large collections or an OCR in the context of mass digitization is not feasible. The question to which extent OCR confidence values and sample-based statistical evaluations can provide meaningful, reliable and comparable statements needs to be more systematically investigated. Finally, the quality of layout analysis seems to be insufficiently covered by established metrics.

This paper aims to raise and discuss issues of transparency and better direct comparability of OCR evaluation by identifying gaps and ambiguities in current methods and by putting the meaningfulness of OCR evaluation results more into the context of actual use cases for OCR results. The observations and analysis are drawn from Google-estimated in 2010 that there are around 190k unique books published at <http://books.google.com> in 2010. The books of world-wide interest and historical interest and digitized until October 2019: <https://www.kit-ebooks.org/search/3-years-google-books/>.

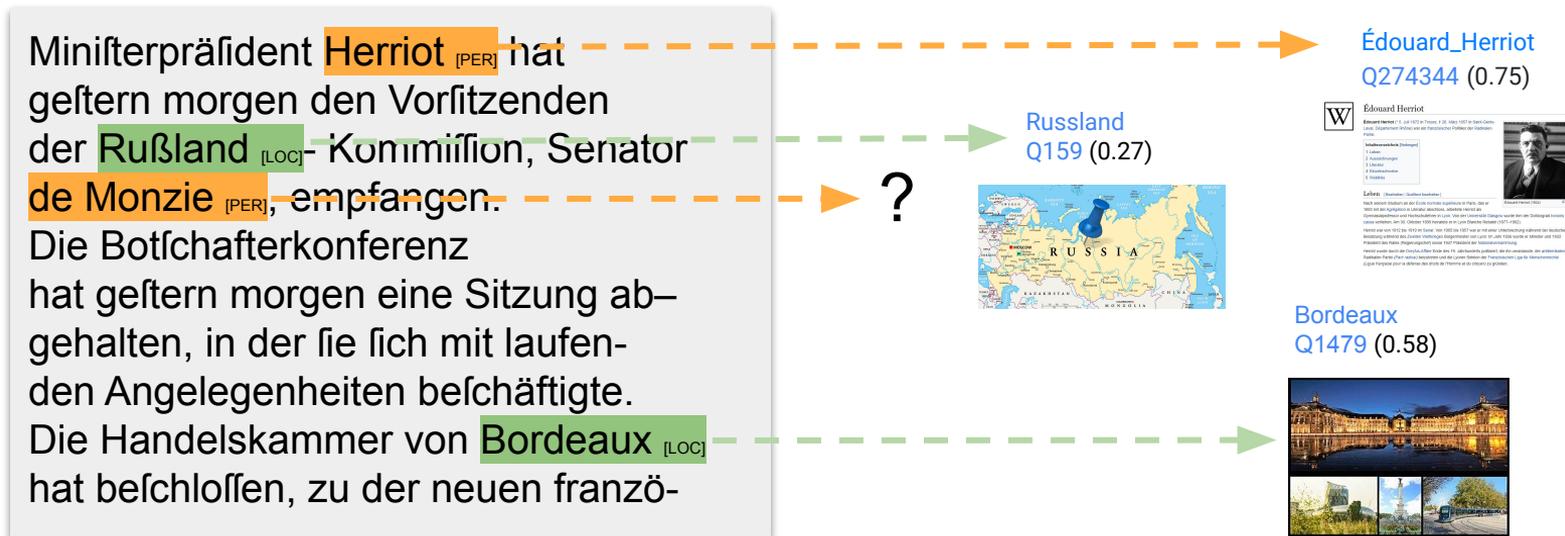
¹<https://www.kit-ebooks.org/search/3-years-google-books/>, accessed 2023-09-01.
²Unicode Standard, Chapter 16: Special Area and Format Characters.
³Unicode Standard, Chapter 16: Special Area and Format Characters.

Named Entity Recognition, Entity Linking

- *BERT for Named Entity Recognition in Contemporary and Historic German, 2019.*

https://konvens.org/proceedings/2019/papers/KONVENS2019_paper_4.pdf |

https://github.com/qurator-spk/sbb_ner | https://github.com/qurator-spk/sbb_ned



Transkription und Annotation

- Annotation von Named Entities, Korrektur und Transkription von Volltexten, Auszeichnung von Bilddaten und Elementen direkt im Browser

<https://github.com/qurator-spk/neat>

neat: neat annotation tool

[User Guide](#) | [Annotation Guidelines](#) | [Issues](#)



[enlarge](#) | [full](#)

<< LOCATION	POSITION	TOKEN	NE-TAG	NE-EMB	ID >>	TEXT >>
9	10	wäre	O	O	-	6. O lit sich nun gefumbler gehan / iu Fuß biß in dië taufent Mann / die hatten Töwens muhte / lie kament hin gehn But- tisholtz / da fandens mengen Engler ftoltz / den fie legten ins Blutte. Es war zu mahl ein harter Streit / das kein theil wolte wei- ch. n/ fie ftunden velt zu beider feit / lefftlich fleng an zaweichen / der English kauff vnd nahm die flucht / alfo handt die Eydgnof- fen/ ihnen felber gemachet lufft.
10	11	,	O	O	-	
11	12	wenn	O	O	-	
12	13	nicht	O	O	-	
13	14	Herr	O	O	-	
14	15	Gambetta	B-PER	O	Q295090	
15	16	als	O	O	-	
16	17	deus	O	O	-	
17	18	ex	O	O	-	
18	19	machina	O	O	-	
19	20	erfchienen	O	O	-	
20	21	wäre	O	O	-	

OK CANCEL

trobert ein fchöne beuñh / an geſchmeidt
 harniſch vnd Roffen / zwey hundert Man
 hands erfchlagen / die warhrit thun ich
 luch fagen / von den freim Eydgnoffen /



© Trustees of British Library Document Supply Centre

© 1999, 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021, 2022, 2023, 2024, 2025, 2026, 2027, 2028, 2029, 2030, 2031, 2032, 2033, 2034, 2035, 2036, 2037, 2038, 2039, 2040, 2041, 2042, 2043, 2044, 2045, 2046, 2047, 2048, 2049, 2050, 2051, 2052, 2053, 2054, 2055, 2056, 2057, 2058, 2059, 2060, 2061, 2062, 2063, 2064, 2065, 2066, 2067, 2068, 2069, 2070, 2071, 2072, 2073, 2074, 2075, 2076, 2077, 2078, 2079, 2080, 2081, 2082, 2083, 2084, 2085, 2086, 2087, 2088, 2089, 2090, 2091, 2092, 2093, 2094, 2095, 2096, 2097, 2098, 2099, 2100, 2101, 2102, 2103, 2104, 2105, 2106, 2107, 2108, 2109, 2110, 2111, 2112, 2113, 2114, 2115, 2116, 2117, 2118, 2119, 2120, 2121, 2122, 2123, 2124, 2125, 2126, 2127, 2128, 2129, 2130, 2131, 2132, 2133, 2134, 2135, 2136, 2137, 2138, 2139, 2140, 2141, 2142, 2143, 2144, 2145, 2146, 2147, 2148, 2149, 2150, 2151, 2152, 2153, 2154, 2155, 2156, 2157, 2158, 2159, 2160, 2161, 2162, 2163, 2164, 2165, 2166, 2167, 2168, 2169, 2170, 2171, 2172, 2173, 2174, 2175, 2176, 2177, 2178, 2179, 2180, 2181, 2182, 2183, 2184, 2185, 2186, 2187, 2188, 2189, 2190, 2191, 2192, 2193, 2194, 2195, 2196, 2197, 2198, 2199, 2200, 2201, 2202, 2203, 2204, 2205, 2206, 2207, 2208, 2209, 2210, 2211, 2212, 2213, 2214, 2215, 2216, 2217, 2218, 2219, 2220, 2221, 2222, 2223, 2224, 2225, 2226, 2227, 2228, 2229, 2230, 2231, 2232, 2233, 2234, 2235, 2236, 2237, 2238, 2239, 2240, 2241, 2242, 2243, 2244, 2245, 2246, 2247, 2248, 2249, 2250, 2251, 2252, 2253, 2254, 2255, 2256, 2257, 2258, 2259, 2260, 2261, 2262, 2263, 2264, 2265, 2266, 2267, 2268, 2269, 2270, 2271, 2272, 2273, 2274, 2275, 2276, 2277, 2278, 2279, 2280, 2281, 2282, 2283, 2284, 2285, 2286, 2287, 2288, 2289, 2290, 2291, 2292, 2293, 2294, 2295, 2296, 2297, 2298, 2299, 2300, 2301, 2302, 2303, 2304, 2305, 2306, 2307, 2308, 2309, 2310, 2311, 2312, 2313, 2314, 2315, 2316, 2317, 2318, 2319, 2320, 2321, 2322, 2323, 2324, 2325, 2326, 2327, 2328, 2329, 2330, 2331, 2332, 2333, 2334, 2335, 2336, 2337, 2338, 2339, 2340, 2341, 2342, 2343, 2344, 2345, 2346, 2347, 2348, 2349, 2350, 2351, 2352, 2353, 2354, 2355, 2356, 2357, 2358, 2359, 2360, 2361, 2362, 2363, 2364, 2365, 2366, 2367, 2368, 2369, 2370, 2371, 2372, 2373, 2374, 2375, 2376, 2377, 2378, 2379, 2380, 2381, 2382, 2383, 2384, 2385, 2386, 2387, 2388, 2389, 2390, 2391, 2392, 2393, 2394, 2395, 2396, 2397, 2398, 2399, 2400, 2401, 2402, 2403, 2404, 2405, 2406, 2407, 2408, 2409, 2410, 2411, 2412, 2413, 2414, 2415, 2416, 2417, 2418, 2419, 2420, 2421, 2422, 2423, 2424, 2425, 2426, 2427, 2428, 2429, 2430, 2431, 2432, 2433, 2434, 2435, 2436, 2437, 2438, 2439, 2440, 2441, 2442, 2443, 2444, 2445, 2446, 2447, 2448, 2449, 2450, 2451, 2452, 2453, 2454, 2455, 2456, 2457, 2458, 2459, 2460, 2461, 2462, 2463, 2464, 2465, 2466, 2467, 2468, 2469, 2470, 2471, 2472, 2473, 2474, 2475, 2476, 2477, 2478, 2479, 2480, 2481, 2482, 2483, 2484, 2485, 2486, 2487, 2488, 2489, 2490, 2491, 2492, 2493, 2494, 2495, 2496, 2497, 2498, 2499, 2500, 2501, 2502, 2503, 2504, 2505, 2506, 2507, 2508, 2509, 2510, 2511, 2512, 2513, 2514, 2515, 2516, 2517, 2518, 2519, 2520, 2521, 2522, 2523, 2524, 2525, 2526, 2527, 2528, 2529, 2530, 2531, 2532, 2533, 2534, 2535, 2536, 2537, 2538, 2539, 2540, 2541, 2542, 2543, 2544, 2545, 2546, 2547, 2548, 2549, 2550, 2551, 2552, 2553, 2554, 2555, 2556, 2557, 2558, 2559, 2560, 2561, 2562, 2563, 2564, 2565, 2566, 2567, 2568, 2569, 2570, 2571, 2572, 2573, 2574, 2575, 2576, 2577, 2578, 2579, 2580, 2581, 2582, 2583, 2584, 2585, 2586, 2587, 2588, 2589, 2590, 2591, 2592, 2593, 2594, 2595, 2596, 2597, 2598, 2599, 2600, 2601, 2602, 2603, 2604, 2605, 2606, 2607, 2608, 2609, 2610, 2611, 2612, 2613, 2614, 2615, 2616, 2617, 2618, 2619, 2620, 2621, 2622, 2623, 2624, 2625, 2626, 2627, 2628, 2629, 2630, 2631, 2632, 2633, 2634, 2635, 2636, 2637, 2638, 2639, 2640, 2641, 2642, 2643, 2644, 2645, 2646, 2647, 2648, 2649, 2650, 2651, 2652, 2653, 2654, 2655, 2656, 2657, 2658, 2659, 2660, 2661, 2662, 2663, 2664, 2665, 2666, 2667, 2668, 2669, 2670, 2671, 2672, 2673, 2674, 2675, 2676, 2677, 2678, 2679, 2680, 2681, 2682, 2683, 2684, 2685, 2686, 2687, 2688, 2689, 2690, 2691, 2692, 2693, 2694, 2695, 2696, 2697, 2698, 2699, 2700, 2701, 2702, 2703, 2704, 2705, 2706, 2707, 2708, 2709, 2710, 2711, 2712, 2713, 2714, 2715, 2716, 2717, 2718, 2719, 2720, 2721, 2722, 2723, 2724, 2725, 2726, 2727, 2728, 2729, 2730, 2731, 2732, 2733, 2734, 2735, 2736, 2737, 2738, 2739, 2740, 2741, 2742, 2743, 2744, 2745, 2746, 2747, 2748, 2749, 2750, 2751, 2752, 2753, 2754, 2755, 2756, 2757, 2758, 2759, 2760, 2761, 2762, 2763, 2764, 2765, 2766, 2767, 2768, 2769, 2770, 2771, 2772, 2773, 2774, 2775, 2776, 2777, 2778, 2779, 2780, 2781, 2782, 2783, 2784, 2785, 2786, 2787, 2788, 2789, 2790, 2791, 2792, 2793, 2794, 2795, 2796, 2797, 2798, 2799, 2800, 2801, 2802, 2803, 2804, 2805, 2806, 2807, 2808, 2809, 2810, 2811, 2812, 2813, 2814, 2815, 2816, 2817, 2818, 2819, 2820, 2821, 2822, 2823, 2824, 2825, 2826, 2827, 2828, 2829, 2830, 2831, 2832, 2833, 2834, 2835, 2836, 2837, 2838, 2839, 2840, 2841, 2842, 2843, 2844, 2845, 2846, 2847, 2848, 2849, 2850, 2851, 2852, 2853, 2854, 2855, 2856, 2857, 2858, 2859, 2860, 2861, 2862, 2863, 2864, 2865, 2866, 2867, 2868, 2869, 2870, 2871, 2872, 2873, 2874, 2875, 2876, 2877, 2878, 2879, 2880, 2881, 2882, 2883, 2884, 2885, 2886, 2887, 2888, 2889, 2890, 2891, 2892, 2893, 2894, 2895, 2896, 2897, 2898, 2899, 2900, 2901, 2902, 2903, 2904, 2905, 2906, 2907, 2908, 2909, 2910, 2911, 2912, 2913, 2914, 2915, 2916, 2917, 2918, 2919, 2920, 2921, 2922, 2923, 2924, 2925, 2926, 2927, 2928, 2929, 2930, 2931, 2932, 2933, 2934, 2935, 2936, 2937, 2938, 2939, 2940, 2941, 2942, 2943, 2944, 2945, 2946, 2947, 2948, 2949, 2950, 2951, 2952, 2953, 2954, 2955, 2956, 2957, 2958, 2959, 2960, 2961, 2962, 2963, 2964, 2965, 2966, 2967, 2968, 2969, 2970, 2971, 2972, 2973, 2974, 2975, 2976, 2977, 2978, 2979, 2980, 2981, 2982, 2983, 2984, 2985, 2986, 2987, 2988, 2989, 2990, 2991, 2992, 2993, 2994, 2995, 2996, 2997, 2998, 2999, 3000, 3001, 3002, 3003, 3004, 3005, 3006, 3007, 3008, 3009, 3010, 3011, 3012, 3013, 3014, 3015, 3016, 3017, 3018, 3019, 3020, 3021, 3022, 3023, 3024, 3025, 3026, 3027, 3028, 3029, 3030, 3031, 3032, 3033, 3034, 3035, 3036, 3037, 3038, 3039, 3040, 3041, 3042, 3043, 3044, 3045, 3046, 3047, 3048, 3049, 3050, 3051, 3052, 3053, 3054, 3055, 3056, 3057, 3058, 3059, 3060, 3061, 3062, 3063, 3064, 3065, 3066, 3067, 3068, 3069, 3070, 3071, 3072, 3073, 3074, 3075, 3076, 3077, 3078, 3079, 3080, 3081, 3082, 3083, 3084, 3085, 3086, 3087, 3088, 3089, 3090, 3091, 3092, 3093, 3094, 3095, 3096, 3097, 3098, 3099, 3100, 3101, 3102, 3103, 3104, 3105, 3106, 3107, 3108, 3109, 3110, 3111, 3112, 3113, 3114, 3115, 3116, 3117, 3118, 3119, 3120, 3121, 3122, 3123, 3124, 3125, 3126, 3127, 3128, 3129, 3130, 3131, 3132, 3133, 3134, 3135, 3136, 3137, 3138, 3139, 3140, 3141, 3142, 3143, 3144, 3145, 3146, 3147, 3148, 3149, 3150, 3151, 3152, 3153, 3154, 3155, 3156, 3157, 3158, 3159, 3160, 3161, 3162, 3163, 3164, 3165, 3166, 3167, 3168, 3169, 3170, 3171, 3172, 3173, 3174, 3175, 3176, 3177, 3178, 3179, 3180, 3181, 3182, 3183, 3184, 3185, 3186, 3187, 3188, 3189, 3190, 3191, 3192, 3193, 3194, 3195, 3196, 3197, 3198, 3199, 3200, 3201, 3202, 3203, 3204, 3205, 3206, 3207, 3208, 3209, 3210, 3211, 3212, 3213, 3214, 3215, 3216, 3217, 3218, 3219, 3220, 3221, 3222, 3223, 3224, 3225, 3226, 3227, 3228, 3229, 3230, 3231, 3232, 3233, 3234, 3235, 3236, 3237, 3238, 3239, 3240, 3241, 3242, 3243, 3244, 3245, 3246, 3247, 3248, 3249, 3250, 3251, 3252, 3253, 3254, 3255, 3256, 3257, 3258, 3259, 3260, 3261, 3262, 3263, 3264, 3265, 3266, 3267, 3268, 3269, 3270, 3271, 3272, 3273, 3274, 3275, 3276, 3277, 3278, 3279, 3280, 3281, 3282, 3283, 3284, 3285, 3286, 3287, 3288, 3289, 3290, 3291, 3292, 3293, 3294, 3295, 3296, 3297, 3298, 3299, 3300, 3301, 3302, 3303, 3304, 3305, 3306, 3307, 3308, 3309, 3310, 3311, 3312, 3313, 3314, 3315, 3316, 3317, 3318, 3319, 3320, 3321, 3322, 3323, 3324, 3325, 3326, 3327, 3328, 3329, 3330, 3331, 3332, 3333, 3334, 3335, 3336, 3337, 3338, 3339, 3340, 3341, 3342, 3343, 3344, 3345, 3346, 3347, 3348, 3349, 3350, 3351, 3352, 3353, 3354, 3355, 3356, 3357, 3358, 3359, 3360, 3361, 3362, 3363, 3364, 3365, 3366, 3367, 3368, 3369, 3370, 3371, 3372, 3373, 3374, 3375, 3376, 3377, 3378, 3379, 3380, 3381, 3382, 3383, 3384, 3385, 3386, 3387, 3388, 3389, 3390, 3391, 3392, 3393, 3394, 3395, 3396, 3397, 3398, 3399, 3400, 3401, 3402, 3403, 3404, 3405, 3406, 3407, 3408, 3409, 3410, 3411, 3412, 3413, 3414, 3415, 3416, 3417, 3418, 3419, 3420, 3421, 3422, 3423, 3424, 3425, 3426, 3427, 3428, 3429, 3430, 3431, 3432, 3433, 3434, 3435, 3436, 3437, 3438, 3439, 3440, 3441, 3442, 3443, 3444, 3445, 3446, 3447, 3448, 3449, 3450, 3451, 3452, 3453, 3454, 3455, 3456, 3457, 3458, 3459, 3460, 3461, 3462, 3463, 3464, 3465, 3466, 3467, 3468, 3469, 3470, 3471, 3472, 3473, 3474, 3475, 3476, 3477, 3478, 3479, 3480, 3481, 3482, 3483, 3484, 3485, 3486, 3487, 3488, 3489, 3490, 3491, 3492, 3493, 3494, 3495, 3496, 3497, 3498, 3499, 3500, 3501, 3502, 3503, 3504, 3505, 3506, 3507, 3508, 3509, 3510, 3511, 3512, 3513, 3514, 3515, 3516, 3517, 3518, 3519, 3520, 3521, 3522, 3523, 3524, 3525, 3526, 3527, 3528, 3529, 3530, 3531, 3532, 3533, 3534, 3535, 3536, 3537, 3538, 3539, 3540, 3541, 3542, 3543, 3544, 3545, 3546, 3547, 3548, 3549, 3550, 3551, 3552, 3553, 3554, 3555, 3556, 3557, 3558, 3559, 3560, 3561, 3562, 3563, 3564, 3565, 3566, 3567, 3568, 3569, 3570, 3571, 3572, 3573, 3574, 3575, 3576, 3577, 3578, 3579, 3580, 3581, 3582, 3583, 3584, 3585, 3586, 3587, 3588, 3589, 3590, 3591, 3592, 3593, 3594, 3595, 3596, 3597, 3598, 3599, 3600, 3601, 3602, 3603, 3604, 3605, 3606, 3607, 3608, 3609, 3610, 3611, 3612, 3613, 3614, 3615, 3616, 3617, 3618, 3619, 3620, 3621, 3622, 3623, 3624, 3625, 3626, 3627, 3628, 3629, 3630, 3631, 3632, 3633, 3634, 3635, 3636, 3637, 3638, 3639, 3640, 3641, 3642, 3643, 3644, 3645, 3646, 3647, 3648, 3649, 3650, 3651, 3652, 3653, 3654, 3655, 3656, 3657, 3658, 3659, 3660, 3661, 3662, 3663, 3664, 3665, 3666, 3667, 3668, 3669, 3670, 3671, 3672, 3673, 3674, 3675, 3676, 3677, 3678, 3679, 3680, 3681, 3682, 3683, 3684, 3685, 3686, 3687, 3688, 3689, 3690, 3691, 3692, 3693, 3694, 3695, 3696, 3697, 3698, 3699, 3700, 3701, 3702, 3703, 3704, 3705, 3706, 3707, 3708, 3709, 3710, 3711, 3712, 3713, 3714, 3715, 3716, 3717, 3718, 3719, 3720, 3721, 3722, 3723, 3724, 3725, 3726, 3727, 3728, 3729, 3730, 3731, 3732, 3733, 3734, 3735, 3736, 3737, 3738, 3739, 3740, 3741, 3742, 3743, 3744, 3745, 3746, 3747, 3748, 3749, 3750, 3751, 3752, 3753, 3754, 3755, 3756, 3757, 3758, 3759, 3760, 3761, 3762, 3763, 3764, 3765, 3766, 3767, 3768, 3769, 3770, 3771, 3772, 3773, 3774, 3775, 3776, 3777, 3778, 3779, 3780, 37

Datenbereitstellung und Kuratierung

- *Datasheets for Digital Cultural Heritage Datasets, 2023.*

<https://doi.org/10.5334/johd.124> | <https://doi.org/10.5281/zenodo.8375034>

The screenshot shows the Europeana Pro website. At the top left is the Europeana Pro logo. The main heading is "Datasheets for digital cultural heritage Working Group". To the right of this heading is a badge that says "Updated on Monday November 6, 2023" and another smaller badge that says "Network Academic Research EuropeanaTech Working group". Below the heading is a paragraph: "This Working Group, set up within the Europeana Research Community and EuropeanaTech Community, works to adapt the concept of datasheets for the cultural heritage sector." Below this is a navigation bar with "Hugging Face" logo and a search bar, and menu items: "Models", "Datasets", "Spaces", "Docs", "Solutions", "Pricing", "Log In", and "Sign Up". Below the navigation bar is a card for "Staatsbibliothek zu Berlin - Preußischer Kulturbesitz" with a "Non-Profit" label and a "Request to join this org" button. Below the card are sections for "AI & ML interests" (listing "Digital Libraries, Digitization, Cultural Heritage") and "Team members" (with a count of 3 and profile icons). At the bottom is an "Organization Card" for "Staatsbibliothek zu Berlin - Preußischer Kulturbesitz" with a description: "Staatsbibliothek zu Berlin - Preußischer Kulturbesitz (Berlin State Library) is one of the largest scientific universal libraries in Germany. The library is digitizing its collections and making them available online: <https://digital.staatsbibliothek-berlin.de/>. In research projects such as [Quator](#) and [Mensch.Maschine.Kultur](#) we produce models and datasets that we want to share here."

Henk Alkemade, Steven Claeysens, Giovanni Colavizza, Nuno Freire, Alba Irollo, Jörg Lehmann, Clemens Neudecker, Giulia Osti, Daniel van Strien

Template: Datasheet for Digital Cultural Heritage Datasets

Version 1 – September 2023

The superscripts added to section headings refer to items in the bibliography at the very end, where every section is thoroughly discussed, explained and further questions can be found. The main structure follows Gebru's (2021) and Pushkarna's templates (2022).

Motivation^{b,f}

[Clearly articulate the reasons for creating the dataset and promote transparency about funding interests. Also provide a brief descriptive overview of the dataset ('at a glance'). For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organisation)? Who funded the creation of the dataset?]

Dataset Description^c

Homepage [Add homepage URL here if available.]

Repository [E.g., if the dataset is hosted on GitHub or has a GitHub homepage, add URL here.]

Paper [If the dataset was introduced by a paper or there was a paper written describing the dataset, add URL here.]

Point of Contact [If known, name and email of at least one person the reader can contact for questions about the dataset.]

Dataset Summary^e

[Briefly summarise the dataset, its intended use and the supported tasks. Give an overview of how and why the dataset was created. The summary should explicitly mention the languages present in the dataset (possibly in broad terms, e.g. translations between several pairs of European languages), and describe the domain, topic, genre covered, keywords, and other relevant metadata.]

Was ist überhaupt KI?

- Der Begriff “Künstliche Intelligenz” wird als problematisch angesehen und sollte besser vermieden werden, da er einen Anthropomorphismus darstellt - als ob es sich hierbei um eine dem Menschen vergleichbare Intelligenz handelt. Dies ist aber eher nicht der Fall.
- Besser wäre entweder allgemein von “maschinellern Lernen” bzw. „deep learning“ zu sprechen oder spezifischer von “stochastischen Vorhersagemodellen” (provokanter „stochastic parrots“, vgl. **Bender et al. 2021**)
- Grundsätzlich gilt: aus möglichst vielen repräsentativen Ausgangsdaten (Beispielen) werden Wahrscheinlichkeitsmodelle trainiert, um diese dann auf weitere Daten anwenden zu können.
- Die Qualität eines Modells (der “KI”) hängt also maßgeblich davon ab, wie umfangreich, qualitativ hochwertig und vielfältig die zum Training verwendeten Ausgangsdaten sind.
- Für die aktuellen KI-Modelle sind jedoch die Antworten auf viele relevante Fragen unbekannt:
 - Welche Daten wurden zum Training verwendet und wo können diese eingesehen werden?
 - Welche Qualitäts- und sonstigen Auswahlkriterien kamen dabei zur Anwendung?
 - Welche Personen waren ggf. an der Annotation von Daten beteiligt?
 - Welche Kontaktmöglichkeiten gibt es um Probleme und Fehler zu melden?

Herausforderungen bei Verwendung von KI

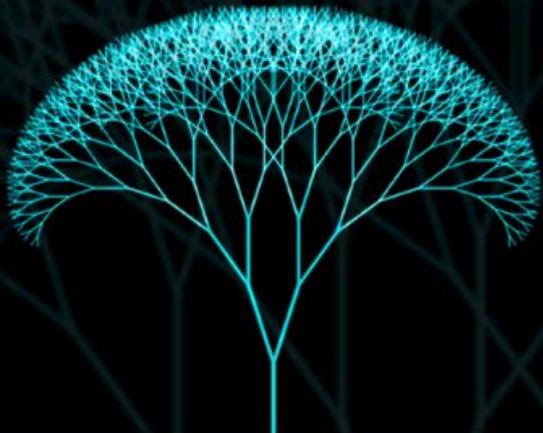
- Aufgrund von geltendem Urheberrecht werden vor allem historische Werke digitalisiert – aber diese enthalten historische Begriffe und Rechtschreibung, für die eine KI erst angepasst bzw. trainiert werden muss
- Digitalisate und Metadaten liegen nicht in der geeigneten Form und in den Formaten vor, wie sie für das Training von KI benötigt werden und müssen daher zunächst evaluiert, konvertiert und/oder (von Expert:innen) transkribiert und annotiert werden
- Kulturelle Kontexte (z.B. Zeit, Ort) müssen bei der Verwendung von Kulturerbe und KI Berücksichtigung finden
- Ethisch, sozial oder rechtlich problematische Inhalte (wie z.B. Kolonialismus, Nationalsozialismus oder die Unterrepräsentation von marginalisierten Gruppen) in den Daten müssen identifiziert und entsprechend vorsichtig behandelt und kontextualisiert werden
- Spezielle Hardware (GPUs) und Expertise (Einwerbung von Personal mit entsprechenden Kompetenzen) wird benötigt

Potentiale von Kulturerbeeinrichtungen und KI

- Der Einsatz von KI bietet viele Möglichkeiten zur effizienten Erschließung, Analyse und Anreicherung von Digitalisaten sowie für neue Services für Nutzende und die Wissenschaft
- Durch die fortschreitende Massendigitalisierung verfügen Kulturerbeeinrichtungen über große und stetig wachsende Mengen an (zumeist offenen) Daten für Training und Verbesserung von KI
- In den jeweiligen Fachbereichen in Kulturerbeeinrichtungen gibt es große Expertise zu den Sammlungen und Inhalten, von denen eine KI lernen bzw. profitieren kann
- Grundsätzlich besteht ein (vor allem im Vergleich zu großen Tech-Unternehmen) hohes Qualitätsbewusstsein und Sensibilität bei der Erstellung, Pflege und Nutzung von Daten
- Als öffentliche Einrichtungen und Dienstleister für Forschung und Wissenschaft werden Transparenz, Datenschutz und Verantwortung im Umgang mit Daten und KI ernst genommen – auch über sehr lange Zeiträume

“In closing, we advocate for a turn in the culture towards **carefully collected** datasets, **rooted in their original contexts**, distributed only in ways that **respect the intellectual property and privacy rights** of data creators and data subjects, and constructed **in conversation with the relevant scientific and scholarly fields** required to create datasets that faithfully model tasks and tasks which target **relevant and realistic capabilities**. Such datasets will undoubtedly be **more expensive to create, in time, money and effort**, and therefore smaller than today’s most celebrated benchmarks. This, in turn, will encourage work on approaches to machine learning (and to artificial intelligence beyond machine learning) that **go beyond the current paradigm of techniques idolizing scale**. Should this come to pass, we predict that machine learning as a field will be better positioned to **understand how its technology impacts people** and to design solutions that work with **fidelity and equity** in their deployment contexts.”

A. Paullada, I. D. Raji, E. M. Bender, E. Denton, A. Hanna. Data and its (dis)contents: A survey of dataset development and use in machine learning research, 2021. <https://doi.org/10.1016/j.patter.2021.100336>



AI4LAM

Artificial Intelligence for Libraries, Archives & Museums



Cultural AI
a lab for
culturally
valued AI

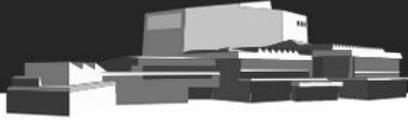
[Mission](#) [News](#) [Topics](#) [Projects](#) [Team](#) [Results](#) [Join](#)

Cultural AI is the study, design and development of socio-technological AI systems that are implicitly or explicitly aware of the subtle and subjective complexity of human culture.

A photograph showing two men. The man on the left is wearing a highly decorative, colorful costume with a large, voluminous pink feathered headdress and large, white, rectangular sunglasses. He has a surprised or expressive facial expression. The man on the right is wearing a dark, collared jacket and is looking towards the man in the costume. The background appears to be the interior of a vehicle or a confined space with overhead lights.

Culture for AI

AI for Culture



**Staatsbibliothek
zu Berlin**
Preußischer Kulturbesitz



Danke für die Aufmerksamkeit!

Fragen?

Clemens Neudecker | Staatsbibliothek zu Berlin - Preußischer Kulturbesitz

HAW Hamburg Ringvorlesung "Bibs & Bits" | 9. Januar 2024