Libraries,
Collections as Data,
and AI
Thomas Padilla

collections as data?

PREFACE.

AS this treatiſe is calculated for the improvement of the riſing generation of *Females* in America, the Lady of faſhion and fortune will not be diſpleaſed, if many hints are ſuggeſted for the more general and univerſal knowledge of thoſe females in this country, who by the loſs of their parents, or other unfortunate circumſtances, are reduced to the neceſſity of going into families in the line of domeſtics, or taking refuge with their friends or relations, and doing thoſe things which are really eſſential to the perfecting them as good wives, and

```
<recipe class1="soups" ethnicgroup="creole" region="south" occasion="other"><p><purpose align="center"
size="smaller" placement="heading">SOUPE MAIGRE, WITHOUT MEAT, FOR LENT</purpose>

Melt half a pound of <ingredient>butter</ingredient> in a <implement>stewpan,</implement> put in six <
ingredient>onions</ingredient> sliced; add two heads of <ingredient>celery</ingredient> cut small, one-
half a head of <ingredient>white cabbage,</ingredient> and a bunch of chopped <ingredient>parsley;</
ingredient> let them boil twenty minutes, then stir in three rolled <ingredient>crackers;</ingredient>
pour in two quarts of boiling <ingredient>milk,</ingredient> or <ingredient>milk</ingredient> and <
ingredient>water;</ingredient> let this boil up gently for half an hour, and just before serving stir in
two well-beaten <ingredient>eggs.</ingredient></p></recipe>
```

Feeding America: The Historic American Cookbook Dataset

View of encoded cookbook text data

```
<recipe class1="soups" ethnicgroup="creole" region="south" occasion="other"><p><purpose align="center"
size="smaller" placement="heading">SOUPE MAIGRE, WITHOUT MEAT, FOR LENT</purpose>

Melt half a pound of <ingredient>butter</ingredient> in a <implement>stewpan,</implement> put in six <
ingredient>onions</ingredient> sliced; add two heads of <ingredient>celery</ingredient> cut small, one-
half a head of <ingredient>white cabbage,</ingredient> and a bunch of chopped <ingredient>parsley;</
ingredient> let them boil twenty minutes, then stir in three rolled <ingredient>crackers;</ingredient>
pour in two quarts of boiling <ingredient>milk,</ingredient> or <ingredient>milk</ingredient> and <
ingredient>water;</ingredient> let this boil up gently for half an hour, and just before serving stir in
two well-beaten <ingredient>eggs.</ingredient></p></recipe>
```

Download Data

The Feeding America: The Historic American Cookbook Dataset

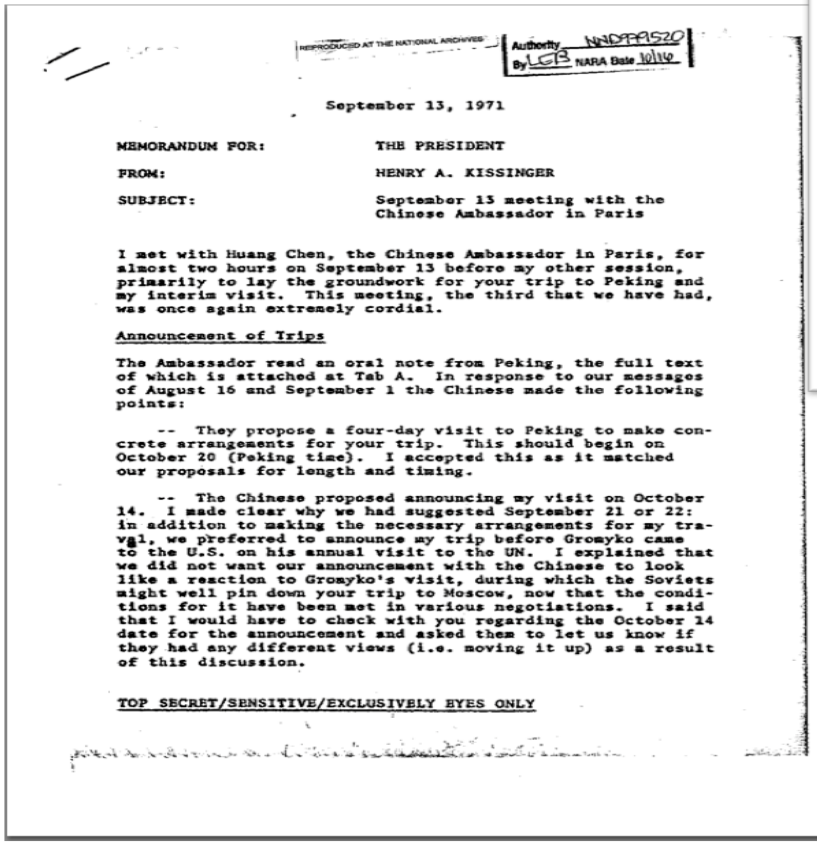
Description

The Feeding America: The Historic American Cookbook dataset contains transcribed and encoded text from 76 influential American cookbooks held by MSU Libraries Special Collections. Features encoded within the text include but are not limited to recipes, types of recipes, cooking implements, and ingredients. The 76 texts were chosen among more than 7000 cookbooks that MSU Libraries holds as representative of periods and themes in American cookbook history spanning the late 18th to early 20th century.

Preferred Citation

Feeding America: The Historic American Cookbook Dataset. East Lansing: Michigan State University Libraries Special Collections. https://www.lib.msu.edu/feedingamericadata/

Example of a 'Memcon' (Meeting Memo)



REPRODUCED AT THE NATIONAL ARCHIVES

Authority NND979520
By LGB NARA Date 10/16

September 13, 1971

MEMORANDUM FOR: THE PRESIDENT

FROM: HENRY A. KISSINGER

SUBJECT: September 13 meeting with the Chinese Ambassador in Paris

I met with Huang Chen, the Chinese Ambassador in Paris, for almost two hours on September 13 before my other session, primarily to lay the groundwork for your trip to Peking and my interim visit. This meeting, the third that we have had, was once again extremely cordial.

Announcement of Trips

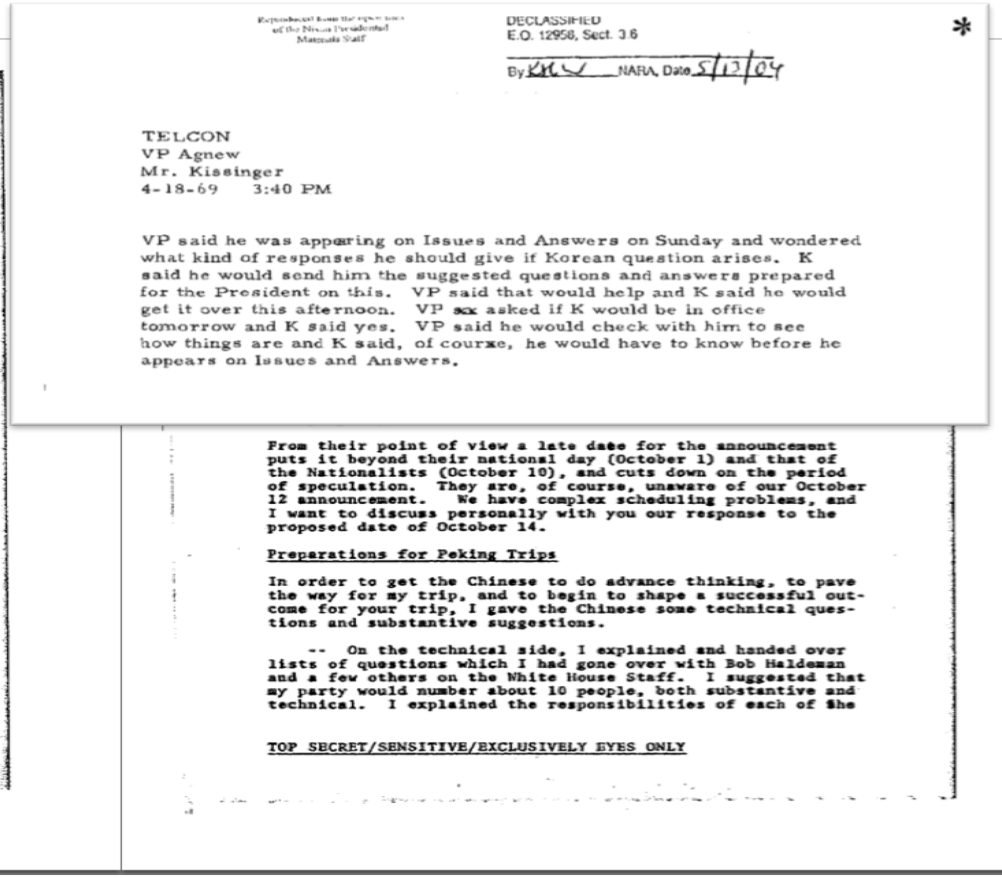
The Ambassador read an oral note from Peking, the full text of which is attached at Tab A. In response to our messages of August 16 and September 1 the Chinese made the following points:

-- They propose a four-day visit to Peking to make concrete arrangements for your trip. This should begin on October 20 (Peking time). I accepted this as it matched our proposals for length and timing.

-- The Chinese proposed announcing my visit on October 14. I made clear why we had suggested September 21 or 22: in addition to making the necessary arrangements for my travel, we preferred to announce my trip before Gromyko came to the U.S. on his annual visit to the UN. I explained that we did not want our announcement with the Chinese to look like a reaction to Gromyko's visit, during which the Soviets might well pin down your trip to Moscow, now that the conditions for it have been met in various negotiations. I said that I would have to check with you regarding the October 14 date for the announcement and asked them to let us know if they had any different views (i.e. moving it up) as a result of this discussion.

TOP SECRET/SENSITIVE/EXCLUSIVELY EYES ONLY

Example of a 'Telcon' (Telephone Conversation Transcript)



Reproduced from the open files of the Nixon Presidential Materials Staff

DECLASSIFIED
E.O. 12958, Sect. 3.6

By KHW NARA, Date 5/13/04

TELCON
VP Agnew
Mr. Kissinger
4-18-69 3:40 PM

VP said he was appearing on Issues and Answers on Sunday and wondered what kind of responses he should give if Korean question arises. K said he would send him the suggested questions and answers prepared for the President on this. VP said that would help and K said he would get it over this afternoon. VP ~~xx~~ asked if K would be in office tomorrow and K said yes. VP said he would check with him to see how things are and K said, of course, he would have to know before he appears on Issues and Answers.

From their point of view a late date for the announcement puts it beyond their national day (October 1) and that of the Nationalists (October 10), and cuts down on the period of speculation. They are, of course, unaware of our October 12 announcement. We have complex scheduling problems, and I want to discuss personally with you our response to the proposed date of October 14.

Preparations for Peking Trips

In order to get the Chinese to do advance thinking, to pave the way for my trip, and to begin to shape a successful outcome for your trip, I gave the Chinese some technical questions and substantive suggestions.

-- On the technical side, I explained and handed over lists of questions which I had gone over with Bob Haldeman and a few others on the White House Staff. I suggested that my party would number about 10 people, both substantive and technical. I explained the responsibilities of each of the

TOP SECRET/SENSITIVE/EXCLUSIVELY EYES ONLY

Source: Digital National Security Archive (DNSA) http://nsarchive.chadwyck.com

Micki Kaufman, Quantifying Kissinger

Relevance ▼ Sort

Narrow results by

⊟ DNSA record type

Document (48)

⊟ Date

1969 - 1974 (years)

Update

Enter a specific date range

⊞ Document type

⊞ Subject

⊞ Classification

Select 1-20 | 0 Selected items | Brief view | Detailed view

1 Memcon of Meeting with Matt Meselson and Information as Requested [Includes Attachment] Preview

United States. National Security Council. Staff. Secret, Information Memorandum. December 17, 1970: 5 pp.

DNSA collection: Kissinger Transcripts

...Kissinger and Matthew Meselson debate use of herbicides and tear gas in Vietnam;

Abstract/Details | Full text - PDF (317 KB)

2 MemCon of Your July 4 Talk with Thieu [Attachment Not Included] Preview

United States. National Security Council. Staff. Top Secret, Action Memorandum. August 4, 1971: 3 pp.

DNSA collection: U.S. Policy in the Vietnam War, 1969-1975

...commitments made to Nguyen Van Thieu during meeting with Henry Kissinger.

Abstract/Details | Full text - PDF (159 KB)

3 MEMCON of Your July 4 Talk with Thieu [Includes Attachment] Preview

United States. National Security Council. Staff. Top Secret, Action Memorandum. August 4, 1971: 21 pp.

DNSA collection: Kissinger Transcripts

...Kissinger to Nguyen Van Thieu related to Paris Peace Talks, retention of South

Abstract/Details | Full text - PDF (2 MB)

4 Meeting between the President, Ambassador Chow, and Henry A. Kissinger [U.S. Relations with China and Taiwan; Includes Action Memorandum Entitled "Chinese Representation at the United Nations"] Preview

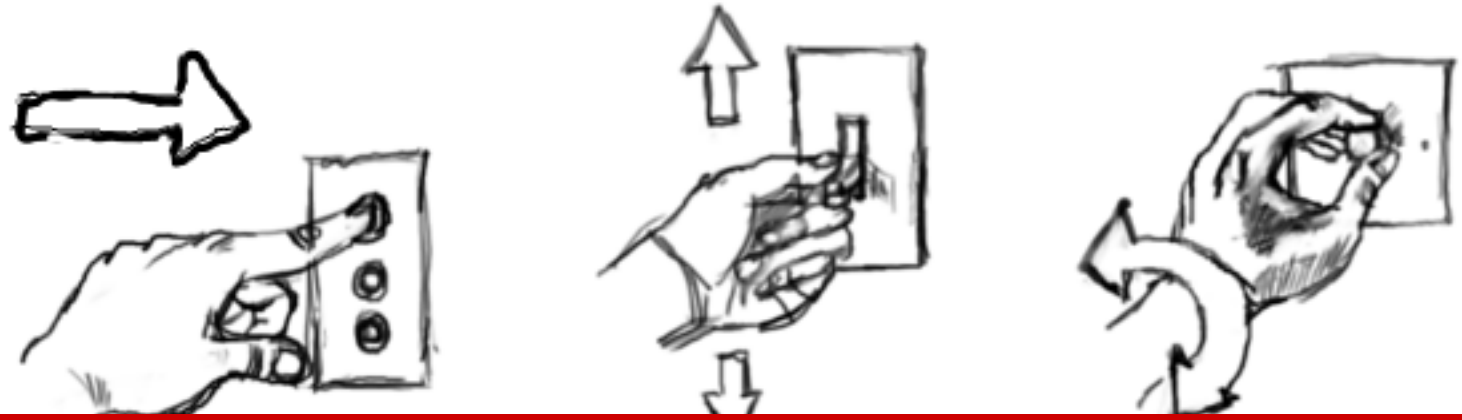
United States. Assistant to the President for National Security Affairs. Top Secret, Memorandum of Conversation. April 12, 1971: 7 pp.

DNSA collection: Kissinger Transcripts

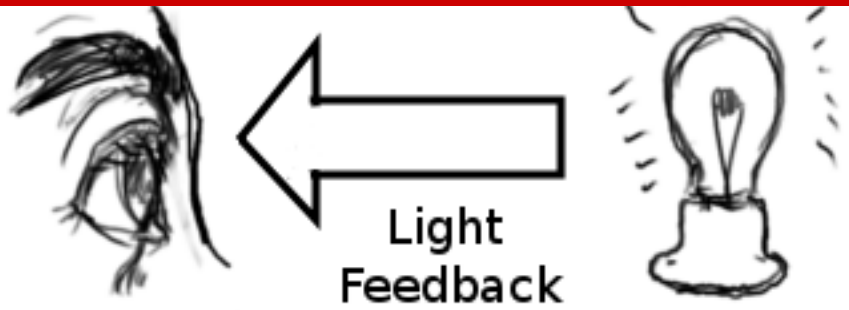
Abstract/Details | Full text - PDF (430 KB)

collections as data

… ordered information
… stored digitally
… amenable to computation



design for qualities of collections that define possible use



INSTITUTE of Museum and Library SERVICES

Always Already Computational: Collections as Data
2016-2018

thomas padilla, unlv
laurie allen, university of pennsylvania
stewart varner, university of pennsylvania
hannah frost, stanford university
sarah potvin, texas a&m university
elizabeth russey roke, emory university

collectionsasdata.github.io

Collections as Data: Part to Whole
2019-2023

thomas padilla, internet archive
hannah scates kettler, iowa state university
stewart varner, university of pennsylvania
yasmeen shorish, james madison university

collectionsasdata.github.io/part2whole

LSU

FIU

CARNEGIE
MUSEUM OF
ART

THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL

VE
RI
TAS
HARVARD

UNIVERSITY OF
RICHMOND

UNIVERSITY of WISCONSIN
UWM
1849 1885 1956
MILWAUKEE

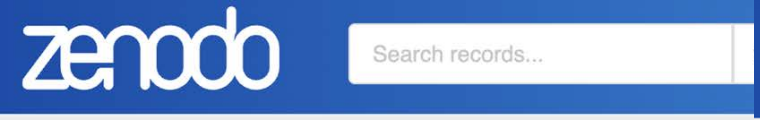
CLAREMONT
McKENNA
COLLEGE

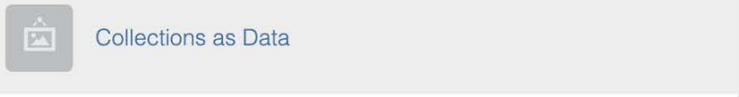
NORTHWESTERN UNIVERSITY
QUAECUMQUE SUNT VERA
1851

UNIVERSITY OF
PITTSBURGH
VERITAS 1787 VIRTUS

UNIVERSITY OF DENVER
COLORADO SEMINARY
PRO SCIENTIA ET RELIGIONE
FOUNDED 1864

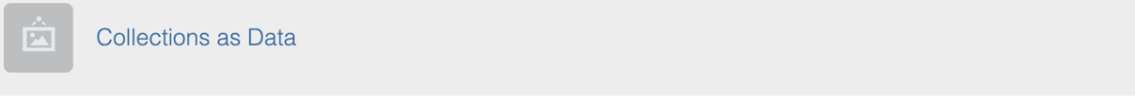
May - June 2023
Draft Collections as Data: State of the Field and Future Directions + Santa Barbara Statement on Collections as Data
July 2023
Seek public feedback on white paper + statement
August 2023
Translate white paper + statement into additional languages
Publish Collections as Data: State of the Field and Future Directions and revised Santa Barbara Statement on Collections as Data





Published December 20, 2023 | Version v1

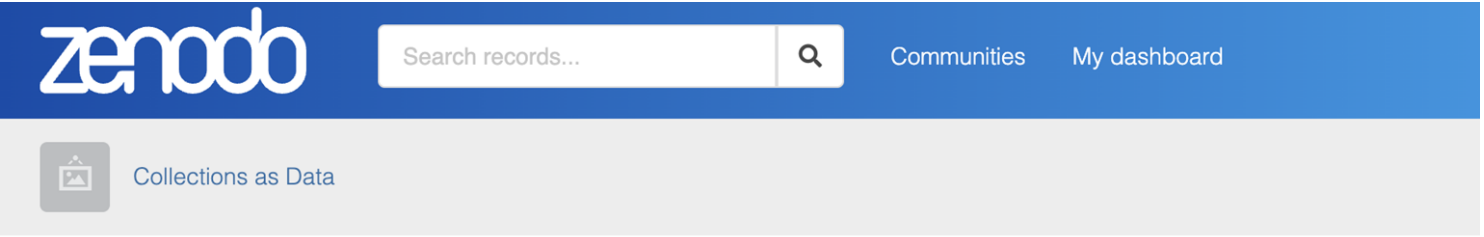
Declaraciones de posturas -> Collec Field and Future Directions (ES)



Published September 13, 2023 | Version v2

Vancouver Statement on Collections as Data (ES)

Padilla, Thomas[1]; Scates Kettler, Hannah[2]; Varner, Stewart[3]; Shorish, Yasmeen[4]



Published November 27, 2023 | Version 1

Report Open

Collections as Data: Part to Whole Final Report (ES)

Padilla, Thomas[1]; Scates Kettler, Hannah[2]; Shorish, Yasmeen[3]

Show affiliations

Collections-as-data (2017) se ha producido
acionalmente. Las instituciones, tanto gran
ilitación de acceso y el respaldo del uso c
lo, es necesaria una declaración actualizad
de datos en el contexto de un entorno de c

C&RL College & Research Libraries

Transfiguring the Library as Digital Research Infrastructure: Making KBLab at the National Library of Sweden

*Love Börjeson, Chris Haffenden, Martin Malmsten, Fredrik Klingwall, Emma Rende, Robin Kurtz, Faton Rekathati, Hillevi Hägglöf and Justyna Sikora**

Collections as data: Acceso computacional a colecciones digitales

Gustavo Candela
Universidad de Alicante
https://orcid.org/0000-0001-6122-0777

How do we responsibly support computational engagement with memory organization collections as data?

Libraries & AI

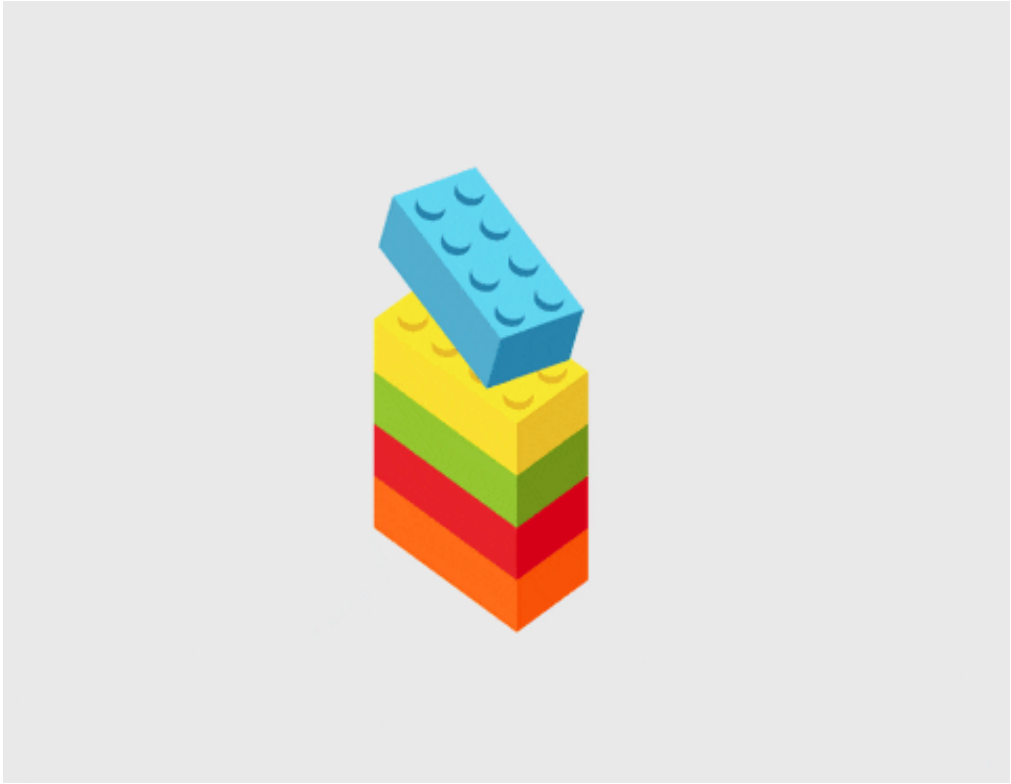
1. **reusable** - ✨ open ✨, not 🙄 open 🙄
2. **accountable -** development and use subject to specific community needs
3. **sustainable -** adoption guided by stewardship mindset

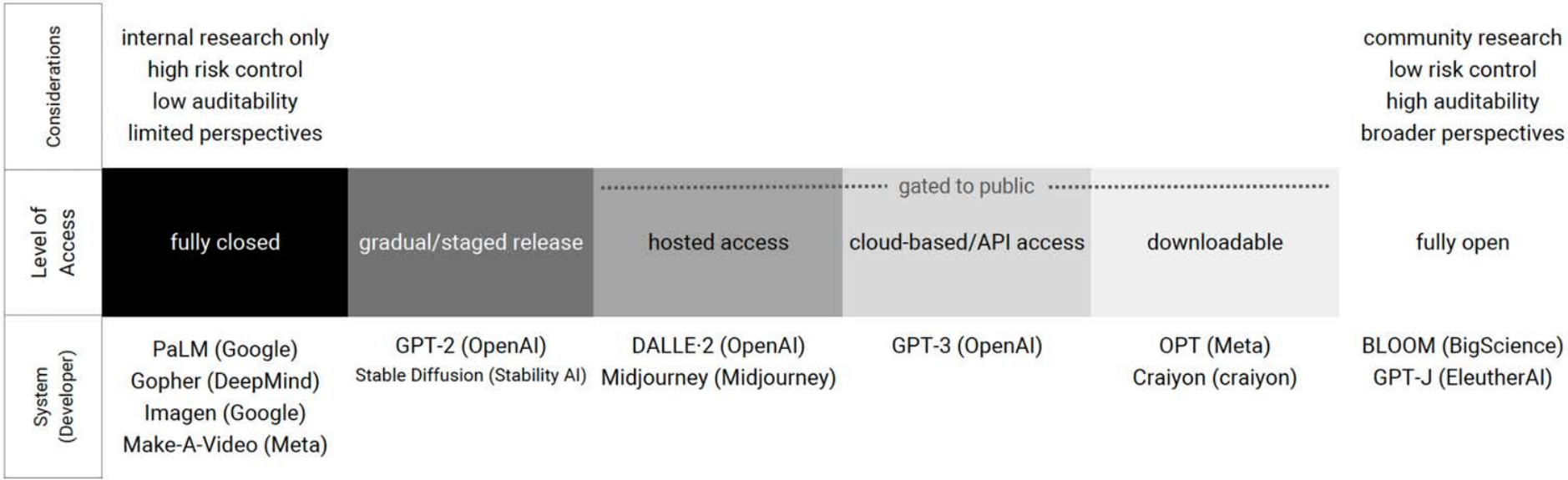
Libraries & AI

1. **reusable** - ✨ open ✨, not 🙄 open 🙄
2. **accountable -** development and use subject to specific community needs
3. **sustainable -** adoption guided by stewardship mindset

open AI =

Maximize our agency through ability to comprehensively assess, freely build upon, reuse, and share





Irene Solaiman, **The Gradient of Generative AI Release: Methods and Considerations**

Large language model

Llama 2: open source, free for research and commercial use

We're unlocking the power of these large language models. Our latest version of Llama – Llama 2 – is now accessible to individuals, creators, researchers, and businesses so they can experiment, innovate, and scale their ideas responsibly.

Download the model

Product Solutions Open Source Pricing

Search or jump to... /

facebookresearch / llama Public

Notifications Fork 8.6k

Code Issues 269 Pull requests 43 Actions Projects Security Insights

Files

main

Go to file

- .github
- llama
- .gitignore
- CODE_OF_CONDUCT.md
- CONTRIBUTING.md
- LICENSE
- MODEL_CARD.md

llama / LICENSE

SulimanSagindykov Update LICENSE 7c0a08e · 7 mont

Code Blame 126 lines (102 loc) · 6.86 KB

```
1  LLAMA 2 COMMUNITY LICENSE AGREEMENT
2  Llama 2 Version Release Date: July 18, 2023
3
4  "Agreement" means the terms and conditions for use, reproduction, distribution and
5  modification of the Llama Materials set forth herein.
6
7  "Documentation" means the specifications, manuals and documentation
8  accompanying Llama 2 distributed by Meta at ai.meta.com/resources/models-and-
9  libraries/llama-downloads/.
```

1. License Rights and Redistribution.

a. Grant of Rights. You are granted a non-exclusive, worldwide, non-transferable and royalty-free limited license under Meta's intellectual property or other rights owned by Meta embodied in the Llama Materials to use, reproduce, distribute, copy, create derivative works of, and make modifications to the Llama Materials.

2. Additional Commercial Terms. If, on the Llama 2 version release date, the monthly active users of the products or services made available by or for Licensee, or Licensee's affiliates, is greater than 700 million monthly active users in the preceding calendar month, you must request a license from Meta, which Meta may grant to you in its sole discretion, and you are not authorized to exercise any of the rights under this Agreement unless or until Meta otherwise expressly grants you such rights.

https://ai.meta.com/llama/use-policy), which is hereby incorporated by reference into this Agreement.

v. You will not use the Llama Materials or any output or results of the Llama Materials to improve any other large language model (excluding Llama 2 or derivative works thereof).

“**Open source software** often becomes an industry standard,” Zuckerberg told investors on an earnings call on Feb. 1. “When companies **standardize on** building with **our stack**, that then becomes easier to **integrate** new **innovations** into **our products**.”

Billy Perrigo, Yann LeCun On How An Open Source Approach Could Shape AI

Open Language Model: OLMo

A State-of-the-Art, Truly Open LLM and Framework

Each model comes with the following:

- Full training data used for these models, including code that produces the training data, from AI2's Dolma, and WIMBD for analyzing pretraining data.
- Full model weights, training code, training logs, training metrics in the form of Weights & Biases logs, and inference code.
- 500+ checkpoints per model, from every 1000 steps during the training process, available as revisions on HuggingFace.
- Evaluation code under the umbrella of AI2's Catwalk and Paloma.
- Fine-tuning code and adapted models (coming soon with Open Instruct)
- All code, weights, and intermediate checkpoints are released under the Apache 2.0 License.

Libraries & AI

1. **reusable** - ✨ open ✨, not 🙄 open 🙄
2. **accountable -** development and use subject to specific community needs
3. **sustainable -** adoption guided by stewardship mindset

Administration Priorities

OCTOBER 30, 2023

FACT SHEET: President Biden Iss Executive Order on Safe, Secure, a Trustworthy Artificial Intelligen

BRIEFING ROOM › STATEMENTS AND RELEASES

English

内閣府 Cabinet Office

内閣府の政策 組織・制度 広報・報道 活動・白

内閣府ホーム > 内閣府の政策 > 科学技術・イノベーション > AI戦略

AI戦略

・AIに関する暫定的な論点整理(本文)(PDF形式488KB) (要旨)(PDF形式:217KB) (要旨(仮訳))(PDF形式:270KB)

・AI戦略2022:(本文)(PDF形式455KB) (別紙)(PDF形式:282KB) (概要)(PDF形式:455KB)

GOV.UK

Home > Business and industry > Science and innovation > Artificial intelligence > AI Safet

AI Safety Institute

Department for Science, Innovation & Technology

Policy paper

Introducing the AI Safety Inst

Updated 17 January 2024

REUTERS® World Business Markets Sustainability Legal Breakingviews Technology Investig

Disrupted

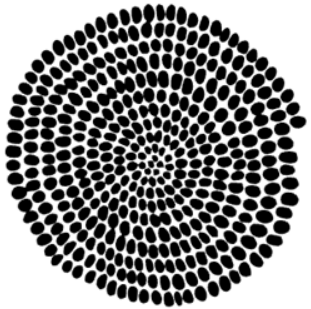
Europe within reach of landmark AI rules after nod from EU countries

By Foo Yun Chee

February 2, 2024 2:31 PM PST · Updated 9 days ago

Distributed AI Research Institute

Learn more



Community

Reducing the distance between researchers and community collaborators

Our research centers lived experiences from communities which are not typically represented in AI research. We strongly believe in a bottom-up approach, supporting and prioritizing research and researchers that emerge from the communities that we center in our work and their specific needs.

Interrogating Power

Creating spaces for communities to question elites

Understanding why people and communities “on the margins” are broadly excluded from AI development, and how this affects all of us, requires interrogating power. Part of our work is to expand the field of inquiry to study decision makers and tech elites that have remained significantly unquestioned in AI research.

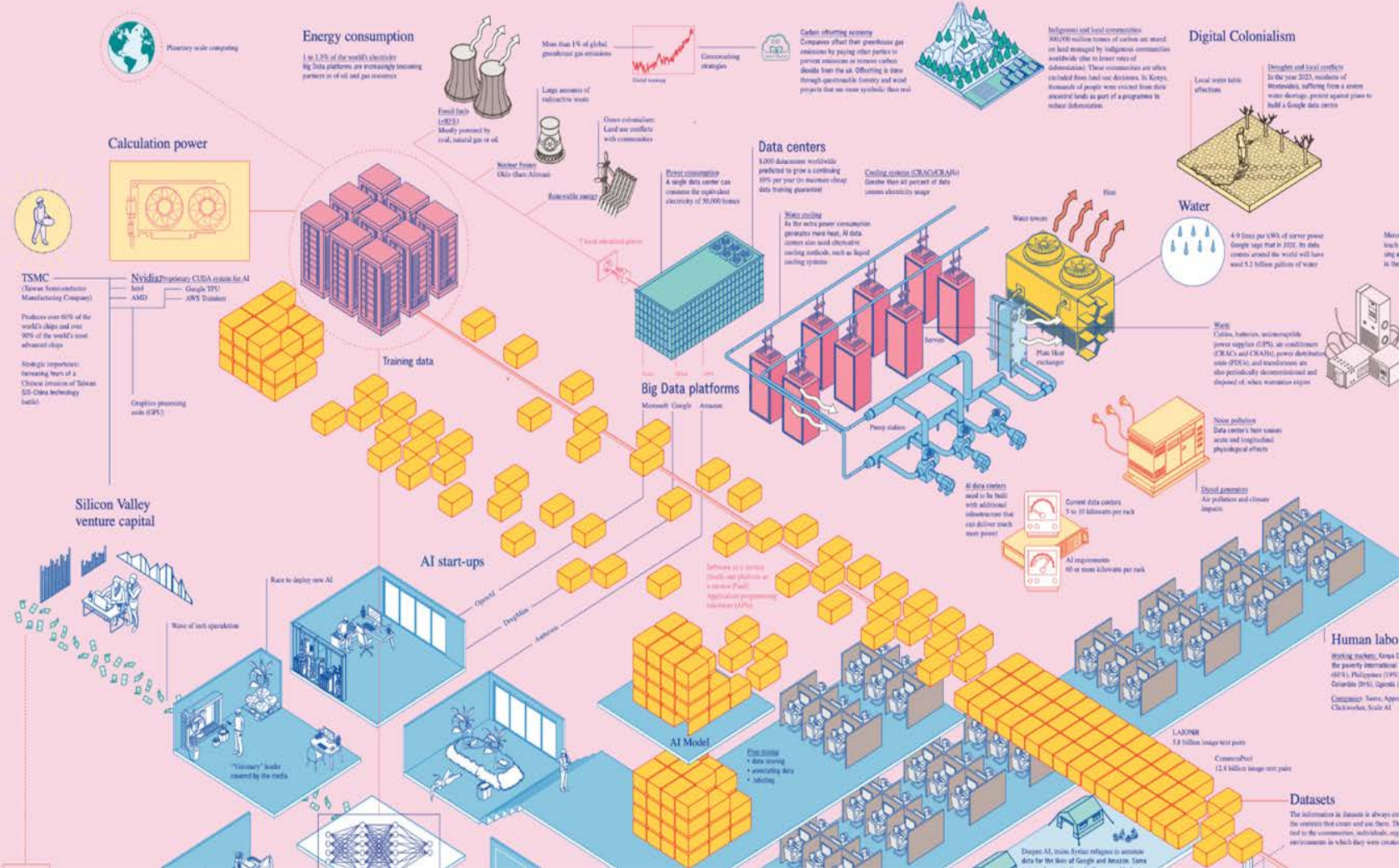
Libraries & AI

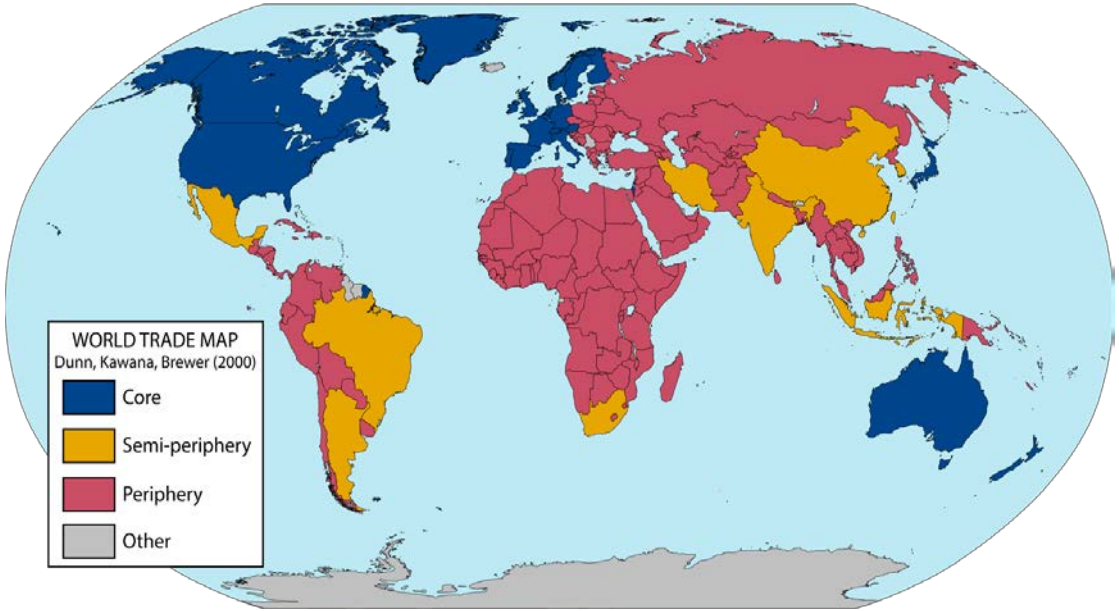
1. **reusable** - ✨ open ✨, not 🙄 open 🙄
2. **accountable -** development and use subject to specific community needs
3. **sustainable -** adoption guided by stewardship mindset

Sustainability strategy depends on awareness of interdependence, threats, and opportunities.

Suggested methods:

1. **Exploded view -** Cartography of generative AI
2. **Systems view -** World Systems Theory (Wallerstein)
3. **Replacement view** - AI is Automation (Bender)


Planetary scale computing
Energy consumption
1 to 1.5% of the world's electricity
Big Data platforms are increasingly becoming partners in oil and gas resources
Fossil fuels (~80%)
Mostly powered by coal, natural gas or oil
More than 1% of global greenhouse gas emissions
Greenwashing strategies
Large amounts of radioactive waste
Nuclear fusion
Okio (San Alfonso)
Renewable energy
Green colonialism: Land use conflicts with communities
Carbon offsetting economy
Companies offset their greenhouse gas emissions by paying other parties to prevent emissions or remove carbon dioxide from the air. Offsetting is done through questionable forestry and wind projects that are more symbolic than real
Indigenous and local communities
300,000 million tonnes of carbon are stored on land managed by indigenous communities worldwide (due to lower rates of deforestation). These communities are often excluded from land-use decisions. In Kenya, thousands of people were evicted from their ancestral lands as part of a programme to reduce deforestation.
Digital Colonialism
Local water table affections
Droughts and local conflicts
In the year 2025, residents of Montevideo, suffering from a severe water shortage, protest against plans to build a Google data centre
Calculation power
TSMC
(Taiwan Semiconductor Manufacturing Company)
Produces over 60% of the world's chips and over 90% of the world's most advanced chips
Strategic importance: Increasing fears of a Chinese invasion of Taiwan (US-China technology battle)
Nvidia Proprietary CUDA system for AI
Intel
AMD
Google TPU
AWS Trainium
Graphics processing units (GPU)
Training data
Power consumption
A single data center can consume the equivalent electricity of 50,000 homes
Data centers
8,000 datacenters worldwide predicted to grow a continuing 10% per year (to maximum cheap data training guarantee
Cooling systems (CRAC/CRAH)
Greater than 40 percent of data centers electricity usage
Water cooling
As the extra power consumption generates more heat, AI data centers also need alternative cooling methods, such as liquid cooling systems
Heat
Water towers
Water
4-9 times per kWh of server power Google says that in 2032, its data centres around the world will have used 5.2 billion gallons of water
Servers
Plate Heat exchanger
Pump station
Waste
Cables, batteries, uninterruptible power supplies (UPS), air conditioners (CRACs and CRAHs), power distribution units (PDUs), and transformers are also periodically decommissioned and disposed of, when warranties expire
Noise pollution
Data center's hum causes acute and long(rednu) physiological effects
Diesel generators
Air pollution and climate impacts
AI data centers
need to be built with additional infrastructure that can deliver much more power
Current data centers
5 to 10 kilowatts per rack
AI requirements
60 or more kilowatts per rack
Big Data platforms
Microsoft Google Amazon
Silicon Valley venture capital
AI start-ups
OpenAI
DeepMind
Anthropic
Race to deploy new AI
Wave of tech speculation
"Visionary" leader revered by the media
AI Model
Fine-Tuning
• data training
• annotating data
• labeling
Human labo
Working markets: Kenya (the poverty international (68%), Philippines (19% Colombia (8%), Uganda (
Companies: Sama, Appe Clickworker, Scale AI
LAION5B
5.8 billion image-text pairs
CommonPool
12.8 billion image-text pairs
Datasets
Deepen AI, trains Syrian refugees to annotate data for the likes of Google and Amazon. Sama



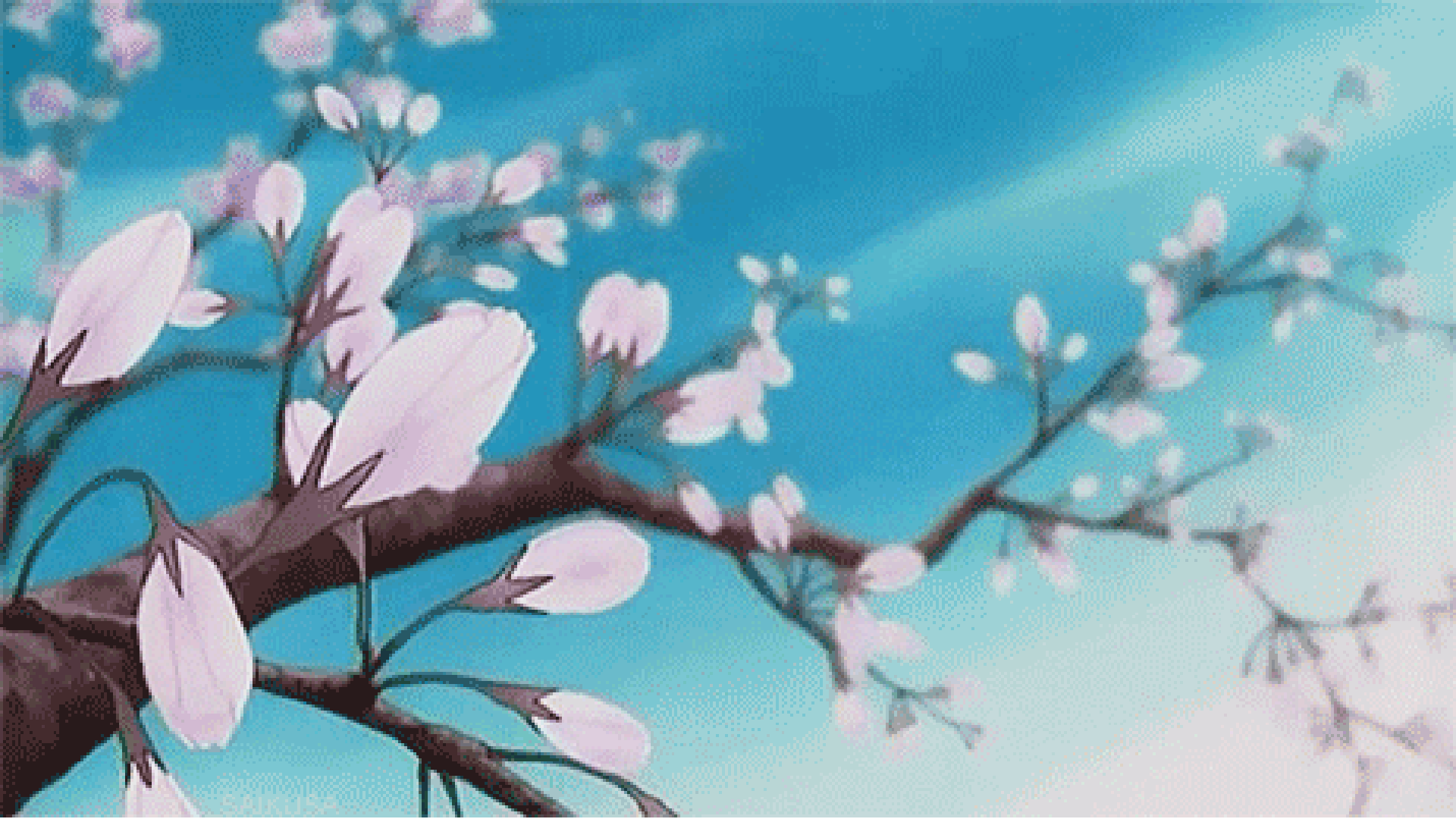
- Maintained distribution of capital concretizes system roles
- Exclusivity - e.g., compute
- Purgatory - e.g., raw materials, training data, content moderation

I think that discussions of this technology become much clearer when we replace the term AI with the word “automation”.

Then we can ask:

- What is being automated?
- Who’s automating it and why?
- Who benefits from that automation?
- How well does the automation work in its use case that we’re considering?
- Who’s being harmed?
- Who has accountability for the functioning of the automated system?
- What existing regulations already apply to the activities where the automation is being used?

Emily Bender, Opening remarks on “AI in the Workplace: New Crisis or Longstanding



call for collaboration!

ARCTIC OCEAN
Greenland
(Denmark)
Svalbard
(Norway)
ARCTIC OCEAN
United States
Canada
Iceland
Norway
Sweden
Finland
Russia
Estonia
Latvia
Lithuania
Ger-many
Poland
Belar.
Ukraine
Mol.
Rom.
Bul.
Geor.
Az.
Armenia
Kazakhstan
Mongolia
Uzbek.
Turkm.
Kyrgyzstan
Tajikistan
China
N.K.
S.K.
Portugal
Spain
France
Italy
Turkey
Cyprus
Syr.
Iraq
Iran
Afghan.
Pakistan
Nepal
Bhutan
India
Laos
Vietnam
Taiwan
Bur.
Bangl.
Thai.
Cambodia
Philippines
Morocco
Tunisia
Greece
Lebanon
Israel
Jordan
Kuwait
Qatar
Saudi Arabia
U.A.E.
Yemen
Oman
Djibouti
Western Sahara
Algeria
Libya
Egypt
Mexico
Caribbean
Senegal
Maurit.
Mali
Niger
Chad
Eritrea
Sudan
The Gambia
B. Faso
Guinea-Bissau
Guinea
Sierra Leone
Liberia
d'Ivoire
Nigeria
Cam.
C.A.R.
South Sudan
Ethiopia
Ghana
Togo
Benin
Equatorial Guinea
Gabon
Dem. Rep. of Congo
Congo
Ken.
Somalia
Uganda
Rwanda
Burundi
Tanz.
Angola
Zambia
Madagascar
Sri Lanka
Malaysia
Papua New Guinea
Indonesia
Solomon Islands
Timor-Leste
New Caledonia
Fiji
New Zealand
PACIFIC OCEAN
Central America
Guyana
Suriname
French Guiana
Venez.
Colombia
Ecuador
Peru
Bol.
Chile
Argentina
Uruguay
ATLANTIC OCEAN
ATLANTIC OCEAN
INDIAN OCEAN
OCEAN
SOUTHERN OCEAN
© 2020 Geology.com

United Nations | Department of Economic and Social Affairs
Sustainable Development

Home | SDG Knowledge | Intergovernmental Processes | HLPF | SIDS | SDG Actions | Engage | News | About

THE 17 GOALS

169 Targets

3937 Events

1348 Publications

7845 Actions

Interested in collections as data or AI and libraries work?

Seeking partners for international convenings and workshops to address challenges and opportunities together.

Please do reach out!

Emily Bender, University of Washington Seattle
Chris Bourg, MIT
Laurie Bridges, Oregon State University
F. Stuart Chapin, University of Alaska Fairbanks
Kate Crawford, USC Annenberg
Timnit Gebru, Distributed Artificial Intelligence Institute (DAIR)
Hanna Hajishirzi, Allen Institute for AI, University of Washington Seattle
Yacine Jernite, Hugging Face
Bergis Jules, Shift Collective
Mia Lund, Open Source Initiative
A Maxwell
Margaret Mitchell, Hugging Face
Rachael Samberg, UC Berkeley
Irene Solaiman, Hugging Face
Anna Tsing, UC Santa Cruz
Immanuel Wallerstein, Yale University
Sarah West, AI Now Institute
David Gray Widder, Carnegie Mellon University
Meredith Whittaker, Signal