BigScience 🌸 Data Governance Group

July 2021 Initial Progress and Future Directions

With: Stella Biderman, Jesse Dodge, Gérard Dupont, Yacine Jernite, Maraim Masoud, Meg Mitchell, Huu Nguyen, Keith Porcaro, Anna Rogers, Zeerak Waseem

Exploratory Phase

Who, What, and Why?

Ethics-informed ML begins with ethics-informed data collection/access.

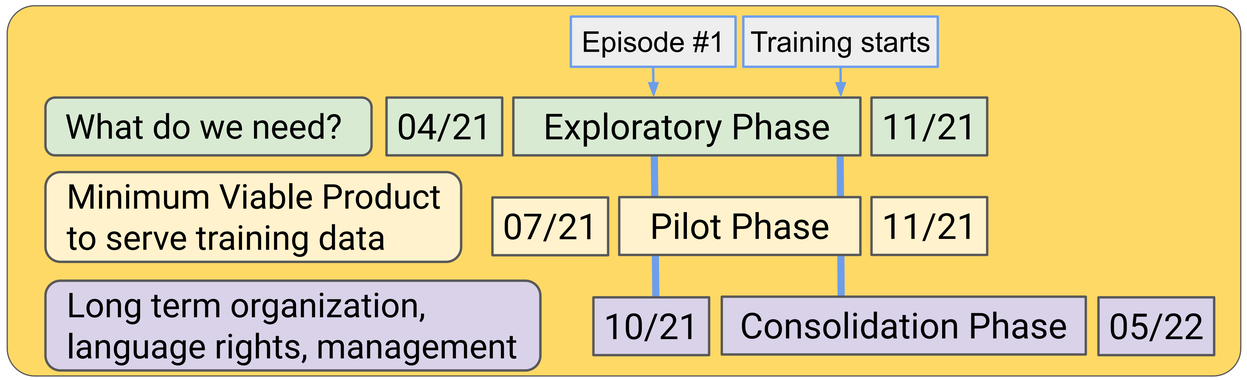
Requires governance structure that can operationalize ethical values.

Stakeholders: researchers, authors, data subjects & custodians

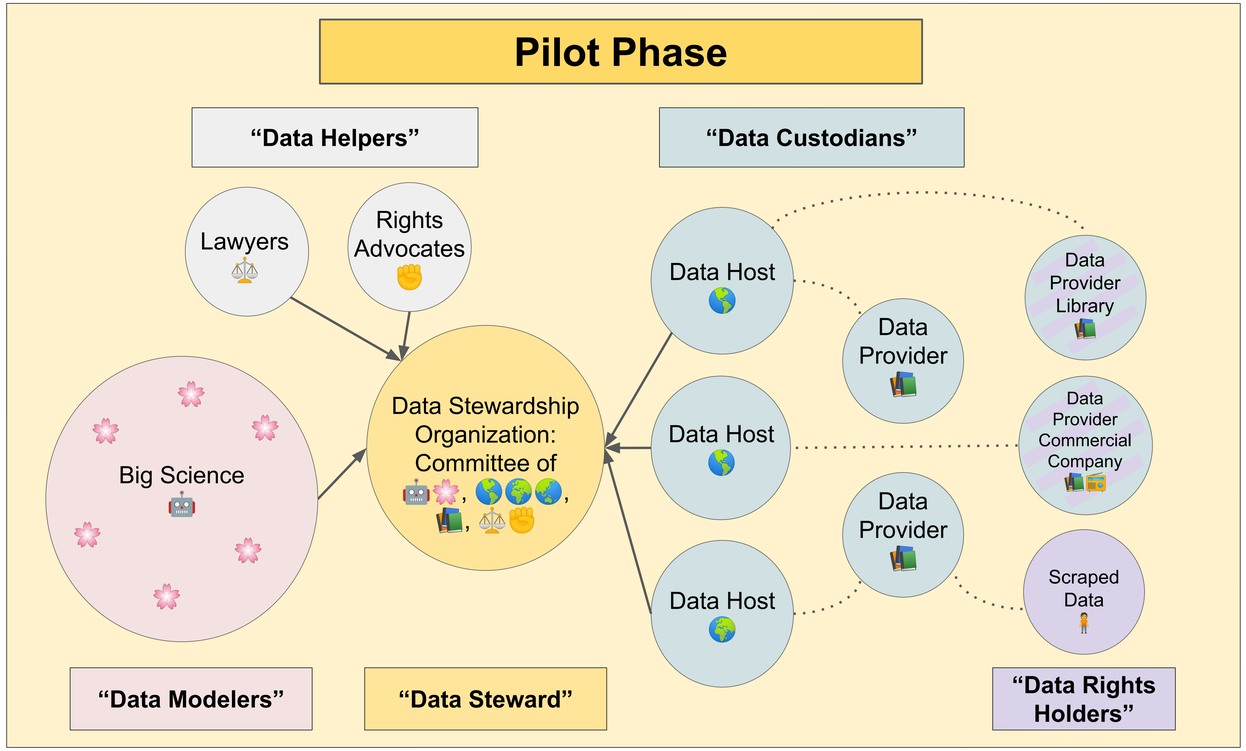
Values	Summary
Licensing/ Attribution	Right to legal controls
Just rewards	Right to benefit from uses of one's data
Privacy	Right to control one's personal data and information
Inclusion/ Representation	Right to equally participate in culture and language tech
Autonomy, *incl.* Consent Contestation	Right to meaningfully participate in application of rights
Beneficence	Above rights subject to *"do no harm"* first

From Values to Data Stewardship

Data Stewardship Org.: DSO	- Connect parties - Establish standards that operationalize values - Provide support to enact these standards
Data Providers, Data Hosts	- DP provide language data & conditions for use - DH gather, index, serve data based on DSO+DP



Africa 🌍	LATAM 🌎	APAC 🌏	EU 🌍	NA 🌎



Relationship to Other Data Working Groups

Data Sourcing
Provides sources of language data to meet DSO values and requirements

Data Tooling
Provides data preprocessing, PII removal, automatic documentation, serving, logging, indexing

Consolidation Phase

Agreements Needed Between

Data users & Data Hosts

DSO & Data Hosts

Data Hosts & Data Providers

Topics of legal and social scholarship (w/ advocates)

Human Rights, Cultural Rights

IP and text data mining & modeling

Privacy in data and models

Cross border cooperation & disputes

Legal playbook for stakeholders

Open Hard Questions

How do we address contestation, including handling trained models?

How do we prove legal compliance?

How do we prevent dissemination of the data beyond approved uses?

How do we keep trusted data secure?

Post-Workshop Governance

Data after Big Science

DSO successor