# Data Governance in the Age of Large-Scale Data-Driven Language Technology

YACINE JERNITE, Hugging Face, United States
HUU NGUYEN, Ontocord, United States
STELLA BIDERMAN, EleutherAI, United States
ANNA ROGERS, University of Copenhagen, Denmark
MARAIM MASOUD, Independent, Ireland
VALENTIN DANCHEV, University of Essex, United Kingdom
SAMSON TAN*, AWS AI Research & Education, United States
ALEXANDRA SASHA LUCCIONI, Hugging Face, Canada
NISHANT SUBRAMANI, Allen Institute for Artificial Intelligence, United States
GÉRARD DUPONT, Independent, France
JESSE DODGE, Allen Institute for Artificial Intelligence, United States
KYLE LO, Allen Institute for Artificial Intelligence, United States
ZEERAK TALAT, Simon Fraser University, Canada
DRAGOMIR RADEV, Yale University, United States
ISAAC JOHNSON, Wikimedia, United States
SOMAIEH NIKPOOR, CAIDP, Canada
JÖRG FROHBERG, apergo.ai, Germany
AARON GOKASLAN, Cornell University, United States
PETER HENDERSON, Stanford University, United States
RISHI BOMMASANI, Stanford University, United States
MARGARET MITCHELL, Hugging Face, United States

The recent emergence and adoption of Machine Learning technology, and specifically of Large Language Models, has drawn attention to the need for systematic and transparent management of language data. This work proposes an approach to global language data governance that attempts to organize data management amongst stakeholders, values, and rights. Our proposal is informed by prior work on distributed governance that accounts for human values and grounded by an international research collaboration that brings together researchers and practitioners from 60 countries. The framework we present is a multi-party international governance structure focused on language data, and incorporating technical and organizational tools needed to support its work.

CCS Concepts: • **Social and professional topics** → **Information system economics**; **Digital rights management**;

Additional Key Words and Phrases: datasets, technology governance, data rights, language data

*Work done prior to joining AWS.

Permission to make digital or hard copies of part or all 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. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
© 2022 Copyright held by the owner/author(s).
Manuscript submitted to ACM

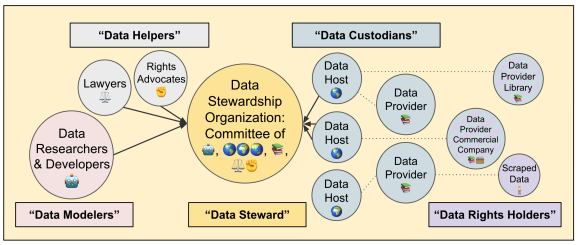


Fig. 1. Overview of the Data Stewardship Organization and Actors

**ACM Reference Format:**
Yacine Jernite, Huu Nguyen, Stella Biderman, Anna Rogers, Maraim Masoud, Valentin Danchev, Samson Tan, Alexandra Sasha Luccioni, Nishant Subramani, Gérard Dupont, Jesse Dodge, Kyle Lo, Zeerak Talat, Dragomir Radev, Isaac Johnson, Somaieh Nikpoor, Jörg Frohberg, Aaron Gokaslan, Peter Henderson, Rishi Bommasani, and Margaret Mitchell. 2022. Data Governance in the Age of Large-Scale Data-Driven Language Technology. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 34 pages. https://doi.org/10.1145/3531146.3534637

## 1 Introduction

New families of algorithms relying on *deep learning* have made it possible to extract ever more complex language statistics from growing numbers of text and speech records to drastically improve the performance and applicability of data-driven *Natural Language Processing* (NLP) systems. As a result, language technologies have become an integral part of daily lived experience in a greater variety of areas both online (Internet search engines, content recommendation and moderation in social media) and offline (automatic translation and speech transcription in official documents and interactions) to the point of becoming near ubiquitous. Consequently, through being so deeply embedded into modern human life, the governance of these new forms of infrastructure —or the lack thereof —has come to exert power over individuals' and communities' lives and access to technology.

These practical applications of language technology are increasingly reliant on approaches based on trained Large Language Models (LLMs) [23, 41, 104, 111], whereby models are first exposed to as large and varied a collection of language data as possible with the aim of extracting *"general"* properties of a language of interest. This first step then makes it easier to fine-tune models that learn to perform a range of *"specific"* NLP tasks more efficiently in that same language setting. As such, the language corpora used to train LLMs need to meet significantly different requirements than the more purpose-specific *datasets* that have hitherto supported major advances in data-driven NLP. Indeed, while concerns of *"generality"* are not new to the field of Machine Learning (ML), this two-stage approach of (pre-)training followed by further training and fine-tuning for a task has given them a new scope; where the properties identified by the model are expected to hold across a much greater variety of tasks, domains, and settings as long as they are in the same "language(s)" as the text it was pre-trained on.

However, whereas recent advances in modeling and hardware have increased the **data training capacity** of LLMs, increasing from Wikipedia-scale corpora to close to three orders of magnitude more, [1] devising methods for carefully identifying, obtaining, and managing a **sufficiently large and diverse collection** of language data to take full advantage of this increased capacity has remained an elusive endeavor. Indeed, in order to support such ambitions of generality, this collection would need to include language data from a great diversity of carefully curated sources

[1]The *Chinchilla* model of Hoffmann et al. [71] was trained on over 1.4 trillion tokens compared to the earlier BERT's 3.3B words corpus.

to minimize harms in downstream applications [12, 120], with international rights holders spanning multiple jurisdictions, and extend to multiple languages beyond the common English (further discussed by Blasi et al. [21]). This requires a more intentional approach to collecting and working with data, but designing a **data governance structure** to appropriately handle such varied data sources while respecting the **rights and interests of their stakeholders** presents a unique challenge that is only partly met by existing language data management approaches.

To better address these needs, we propose a new model for data governance in the form of a Data Stewardship Organization (DSO, see Figure 1 diagram) working in conjunction with related stakeholders and rights holders. The DSO primarily aims to foster the agency of data subjects and rights holders with respect to the uses of their data as the amount and diversity of contexts for this data grows. It is designed to enable multiple stakeholders to collaborate on the decisions that go into building and managing a collection of language resources, so as to meet goals of responsible data governance at a scale and diversity that may support this new generation of data-driven language technology. While our work is grounded by the goal of training a multilingual LLM, we also note that many of the constraints and impacts of the design choices proposed in this work hold across a greater variety of uses of human-centric research and development data. We endeavor to also consider these related applications when relevant.

### 1.1 Research Context and Paper Outline

The research presented in this paper is conducted in the context of a year-long, distributed, collaborative workshop on Large Language Models. [2] The workshop brings together over 1000 participants from 60 countries and is organized into smaller working groups focused on key aspects of the topic, including model architecture and training procedure, evaluation of performance and social biases, multilinguality, and data sourcing and governance. Part of the value in this project is making connections between different fields of knowledge that are not normally very connected, and do so in a practical case study that pursues ethical, legal and technical goals all together.

The following sections present the findings of the data governance group in our effort to build a governance structure to manage and preserve the training data used in the project while promoting agency of all stakeholders and contending with multiple legal contexts. Section 2 overviews prior work on the theory and mechanisms of distributed governance, outlining the special role of *values* and defining the object of our governance effort. Section 3 then examines the social, legal, and technical context for using language data, and Section 4 reviews current approaches to data management in ML/NLP and in Wikimedia, a distributed collaborative project whose goals and requirements have significant overlap with ours. Finally, Section 5 describes our proposed governance structure, describing its various actors and outlining a framework for their interactions.

## 2 Distributed Governance: Values and Definitions

Our proposed organization aims to promote *better* data governance in the context of data-driven language technology research and development. To support this project, we start by reviewing literature on the processes and mechanisms of distributed governance (Section 2.1), and in particular on the *values* that underpin them (2.2). We then position our governance proposal with respect to these processes by defining both its object, namely human-centric data used in NLP (2.3), and its relationship to other aspects of data management (2.4).

[2] https://bigscience.huggingface.co/

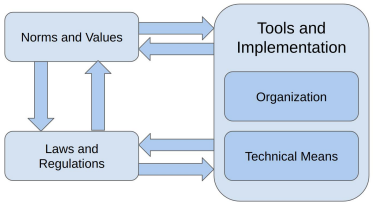


Fig. 2. Collaborative governance mechanisms rely on interacting pillars.

Fig. 3. Machine Learning Data Triad

### 2.1 Approaching Collaborative Governance: Theories and Mechanisms

Governance is a nebulous concept, defined by the Commission on Global Governance [32] as *"the sum of the many ways individuals and institutions, public and private, manage their common affairs"*. Topics such as *technology governance* have received increasing attention in the last few decades as the digital transformation of the late 20th and early 21st century has increased the speed at which technological innovation changes people's lives around the world [139], leading to extensive analyses of the processes, dynamics, and particular challenges of global governance.

One such challenge has proven to be the impossibility of governing any individual subject in isolation in a fully integrated world, a phenomenon studied under the name of **regime complexes** [99]. Keohane and Victor [77] study the case of the regime complex for climate change, whose global governance happens at the intersection of *e.g.*, UN and local legal regimes and bilateral agreements and spans topics such as trade regulation, technology, or geoengineering. Consequently, governance efforts need to account for **fragmentation** when organizations in inter-connected areas make choices that have bearing on each other; by examining these connections and positioning any decision within a dense network of issues and entities [143]. *Data governance*, especially of language data, is similarly integrated in a multitude of related areas, of which Section 3 will discuss the social, legal, and technical dimensions.

Having a broad classification of the mechanisms that underpin governance can help us better navigate this network. In addition to the **laws & regulations** in their various regimes, previous work has focused on the specific role of **tools & implementation** (such as indicators [83] or ICT tools [103]) and on the importance of **norms & values** [80] in governance efforts. In general, we can map mechanisms reviewed in governance literature to one or to the intersection of two of these pillars (Figure 2). For example, Ada Lovelace Institute et al. [4]'s aspects of *algorithmic accountability* include mechanisms such as *prohibitions and moratoria* (regulations), *principles and guidelines* (norms), or *independent oversight bodies* (organizational tools). A similar analysis may be applied to works studying governance's aim to **identify and resolve tensions** between actors. Emerson et al. [46]'s proposed framework considers *principled engagement* and *shared motivation* between all the participants in a governance structure (their shared values) as the basis for resolving tensions. The approach of Feiock [52] addresses dilemmas stemming from different externalities, such as different local regulations of the object of governance. Wareham et al. [137] examine the case of governance of software platforms (specifically the tools they rely on) through the lens of striking a balance between a system's stability and ability to evolve. In order to position our own governance efforts with respect to all these processes, we review its values in Section 2.2, relations to technical tools in 2.4 and to regulations in 3.2.

Finally, previous work has also pointed out how the very mechanisms used to resolve tensions can **shift or entrench power imbalances**, and advised to pay special attention to this phenomenon. Barnett and Duvall [10] and Purdy [108] examine how authority, resources, and discursive legitimacy can lead to exclusion within collaborative governance

| Value | Description |
|---|---|
| Inclusion, Representation, & Non-Discrimination | Equal access to cultural resources and ability to interact with language infrastructures and technology without prejudice |
| Autonomy, Consent & Contestation | Right of individuals and communities to meaningfully control the inclusion of their language data in public resources |
| Privacy | Right of individuals to control who may have access to their personal identifying information (PII) |
| Just Rewards | Right to share in the financial and social benefits stemming from uses of an individual's or communities' language and data |
| Licensing & Attribution | Right to legal controls over one's data and the product of one's work |
| Local Knowledge | Local expressions of values and their context take precedence when making and implementing local decisions |
| Participation | Above values and definitions evolve based on actors' needs and feedback |
| Beneficence | Above values subject to a general "do no harm" approach |

Table 1. Set of values proposed to guide our data governance effort.

efforts. In particular, Mohamed et al. [95] call attention to the "first-mover advantage" phenomenon in setting standards in the contact of AI governance: values that protect and are of interest to the people who write the standards will necessarily be prioritized over values whose upholding is more urgently needed by other parties. We endeavor to be cognizant of these risks in our own governance proposal, both in the expression of its driving values (Section 2.2) and of its structure and processes (Section 5).

### 2.2 Values of Governance, Governance of Values

Section 2.1 identifies *norms* and *values* as main pillars of governance, which are implicitly or explicitly defined by the organizations contributing the the governance structure [80]. These shape design choices and trade-offs, and a static set of values, or ones expressed exclusively by the originators of the project, can lead to exclusion [10] and reinforce disparities [95]. In this context, taking time to examine the values driving our own project, the framework that is used to contextualize them, and the way they themselves are governed is particularly important.

Birhane et al. [18] review recent literature in ML to identify values that are typically put forward to motivate work in that field. They note that most of these focus on endogenous notions of technical performance and novelty, and leave out considerations of broader context and impact that are necessary to shaping a governance effort. Inspired by their approach, we reviewed the working documents of the LLM workshop grounding this paper (see Section 1.1)[3], and found that notions of **inclusivity** and **non-discrimination** regularly appeared. Many of the participants' comments were also informed by the recent European drive towards more data and algorithmic regulation, including their focus on **respecting privacy** and promoting the **agency of data and algorithm subjects**[4]. Additionally, and in reaction to recent practices of indiscriminate use of crawled web text, participants expressed a concern for **respecting rights** of the text creators (e.g., copyright or intellectual property laws). Finally, participants, especially ones with ties to Africa[5]

[3]Records organized by working groups are publicly available at https://drive.google.com/drive/folders/1db2hYZuRs2VjoIrVaVtZJ5FLE2iS7z3p. Appendix A describes the interactions that led to the initial set of values in more details.
[4]https://eur-lex.europa.eu/eli/reg/2016/679/oj
[5]https://www.masakhane.io/

and South-East Asia, pointed out the potential for exploitative data practices in fully open ML data and research [92], and stressed the need for **equitable distribution of the benefits** stemming from data use and work.

Among these values, the stated goal of inclusivity merits further examination. Our project aims to govern global language data, which as we shall see in Section 3.1 shows significant variation across cultural and social contexts. Meanwhile, the participants of our research project remain embodied in their own subjectivities [68], which necessarily represent but a small portion of these contexts. As such, devising a governance structure based solely on values expressed by our participants runs the risk of prioritizing their interests and excluding visions that may be more relevant to other language users [65]. Additionally grounding the definition of our proposed values in *Human Rights frameworks* constitutes an appealing starting point to addressing these limitations given their global reach, varied realizations (both historically and geographically), and general recognition as an accepted foundation of good governance [55]. Indeed, we find that documents such as the UDHR [8], ICCPR [6], or ICESCR [7] echo the proposed values of non-discrimination, privacy and just rewards respectively, and help ground them in an external system [106].

At the same time, while the principle of Human Rights does have a universal scope, the staggering number and diversity of human rights documents written both at the UN [6] and regional level brings to light the inadequacy of focusing on a limited set of human rights document as absolute grounding when outlining values that apply *equitably* to a global and contemporary setting, as it arises from significant differences in their philosophical foundations across the globe [122]. Scholarship at the intersection of decoloniality and human rights in particular has called out the need to question the universality of how we conceptualize the "human" in Human Rights [51, 89]. Furthermore, the focus of human rights discourse has historically been on the relationship between the individual and the state [117], whereas the language data we propose to govern is created and managed at various meeting points between individuals, communities, corporations, and state organizations [134]. Thus, acknowledging both the need for respecting local expertise and conceptualization of human rights [5] and for more *relational and community-oriented* notions of justice [17, 98], we complement our initial set of values as outlined in Table 1 to explicitly include the prioritization of **local knowledge** when realizing shared aspirations in the context of participants' own expression of values, and conversely **participation** in shaping these over-arching governance values to better account for their specific needs.

### 2.3 Object of Governance: Distributed Language Data

We apply the norms and values described above to a governance structure for language data, paying special attention to its use in ML datasets. In the ML world, data refers to any digital representation of acts, facts or other information in forms that include text, images, video, or audio recordings [2, 100, 134], which may be collated or formed into more complex information [13]. It is often described via analogy as food and fuel for ML systems, water and oil as an AI resource, and increasingly as records of human activity [128]. As such, it is a fundamental catalyst for the creation of artificial intelligence systems [13], used for training, developing and/or testing AI systems in the form of datasets, an organized collection of data for a defined task [100].

The proposed data governance structure introduced in this paper focuses on **digital language data**, which includes text from news and academic articles, reports, white papers, blogs, social media posts, radio shows, and digitized books. All such data is created by a person, group of people, or an organization who may hold the rights to that data. All of these dimensions of the language data and datasets it's organized as have bearing on the governance choices (see Table 2 for examples of categories). Of particular sensitivity in data governance is **human-centric data**, data that

[6] https://www.ohchr.org/en/professionalinterest/pages/coreinstruments.aspx

| Category | Examples |
|---|---|
| *Domain* | news, medical, legal |
| *Genre* | literature, social media, articles |
| *Legal status* | public use, non-commercial |
| *Origin* | person, organization |
| *Source* | book, social media platform, radio |
| *Modality* | text, audio, video |
| *Goals* | curated corpus, benchmark, convenience sample |

Table 2. Dimensions of digital language datasets.

| Stakeholders | Examples |
|---|---|
| *Data subjects* | people(s) being talked to/about |
| *Data creators* | journalists, social media users |
| *Data aggregators* | social media platforms |
| *Dataset creators* | researchers, organizations |
| *Dataset distributors* | researcher, university, dataset hub |
| *Dataset users* | model developers |
| *Those affected* | users/subjects of ML systems |

Table 3. Stakeholders of human-centric data.

additionally refers to or represents the ideas of a person. Some kinds of textual data such as weather reports are less likely to harm an individual in case of lax governance or misuse, but human-centric data brings with it concerns around how an individual is represented, how that representation may affect them, the represented individual's consent, and other fundamental and legal rights [1]. The specific risks to be considered depend on the individual's relationship to the data, from creators who may have commercial rights on the product of their work to users and (passive) subjects of the technology developed based on the data. Table 3 lists these various stakeholders, and we explore their different needs and interactions with roles in our proposed structure in Section 5.

### 2.4 Focus of a Data Governance Organization

The life cycle of the data and datasets we aim to govern spans many different stages, including: data creation, selection, curation, documentation, dissemination, hosting and serving, conservation, tracking, versioning, and deletion [74, 100]. Management of each of these different stages will impact our ability to support the values outlined in Table 1, which will also depend on the characteristics of the data along the dimensions illustrated in Table 2 and on which stakeholders are most directly involved (Table 3). In order to better define the specific scope of our governance proposal in all of the many combinations of these parameters, we need to be able to differentiate what falls within its purview from what may better be addressed by other aspects of data management at these various stages.

Specifically, our approach to data governance separates work done with the data, such as selection and curation, from work done around data access, control, and exchanges between different data actors. Our focus may be seen as *people-centric*, narrowing in on the people represented and the users of the data, rather than on its analytics. Thus, we make a distinction between **Data Governance**, **Data Sourcing**, and **Data Tooling** as outlined in Figure 3. These three directions are complementary: Data Governance provides an overall structure wherein Sourcing and Tooling can come into play. The governance work provides norms, frameworks, and communication mechanisms in order to *e.g.* help operationalize definitions of contestation in different legal contexts to allow for the development of *locally relevant* supporting tools, or to formalize relations between actors in different roles and parts of the world. We illustrate this categorization further on three concrete aspects of data management next.

**Data Governance supports Data Sourcing.** While governance focuses on data stakeholders and interacting norms, values, laws, etc., the governance structure operates over the datasets provided through data sourcing efforts, which accumulate, categorize, organize, and document data for datasets. Governance provides a framework for helping sourcing actors formalize rules on how the data they propose may be used and processing requirements. These frameworks are designed to enable values such as representation by e.g. removing barriers to entry and lowering risk of participation in technology to enable actors with local knowledge to identify and fill gaps in available language resources

within the global organization [90]; however the diversity of the sources represented in the governance structure will, ultimately, depend on the quality of the sourcing efforts.

**Handling Personal Information and Privacy.** Table 1 includes values of privacy and consent. To uphold these we need to understand what constitutes and be able to identify instances of Personal Identifiable Information (PII; a term used in the U.S.) and Personal Data (a term used in the U.K., E.U., and some other jurisdictions). All three directions illustrated in Figure 3 come into play for this aspect: (a) governance helps guide the focus on the relevant aspects of personal information depending on the data types (Table 2), with local legal context shaping policies for what to do with that information (e.g. whether it is indexed, obfuscated, accessible); (b) tooling implements these definitions into software that can look for instances of personal information at scale in large amounts of text data; (c) sourcing makes decisions on what data to prioritize based on the identified privacy risks and impact on various stakeholders.

**Contestation and Removal of Data.** We also want our proposed structure to promote contestation rights and control over one's data, in particular by allowing parties who have personal or commercial rights on data included in the organization to request its removal. This aspect also exemplifies the interaction between governance and tooling responsibilities. The former defines actionable guidelines and processes for identifying what consitutes a *legitimate removal request* depending on the local norms and regulations of the requester and data custodian. The latter needs to ensure that the data instance can be easily found and deleted in datasets. In particular, deletion can only be meaningfully enacted if the governance structure ensures *non-dissemination of the data*; that is, if data modelers and researchers can use it without making and broadcasting their own copies. This needs to rely on a combination of technical tools and, when they aren't mature enough, signed agreements or licenses defining the parameters of their data access. We describe a framework for these agreements in Section 5.

## 3 Language Data: Social, Legal, and Technical Context

Section 2 outlines the general mechanism of governance (2.1), the values supporting our effort (section 2.2), the kind of human-centric data we focus on (section 2.3), and its relationship to other aspects of data manangement (2.4). Our next step is to investigate the interplay between the object of our governance effort and its broader context: the social context of language data (section 3.1), the relevant legal principles and frameworks (section 3.2), and the culture around language data use in ML and NLP (section 3.3).

### 3.1 Social Context: Social Variation and Language Discrimination

The governance values outlined in Table 1 include inclusion & non-discrimination. Section 2.1 also cautions against the risk that governance mechanisms might entrench inequalities and power disparities if those are not explicitly taken into account. In order to apply these values and avoid those risks in the context of language data, we need to consider its social dimension. Here, we review sociolinguistics literature to identify social variables that can engender discrimination in inter-personal interactions and meetings with technology.

Most named human languages are collections of language varieties with differences that stem from demographic factors such as education, geography, race, and socio-economic class [82]. However, there is a common misunderstanding that there are well-defined boundaries between languages, each with only a single grammar, lexicon, and orthography, and this has resulted in the stigmatization of the language varieties not associated with status and power [58, 62, 94, 114], negatively impacting speakers's access to social infrastructures (*e.g.* schools [36, 69], courtrooms [118]).

This misunderstanding permeates in modern NLP practices. For instance, texts which display sociolinguistic variation, e.g., social media text, are often labelled as "noisy", while text from prestige variants are deemed "clean". Such *politics of dirt* [44] reveal attitudes that stigmatize minority language variants [58, 62, 114] (as well as demeaning the people that speak them) whilst obscuring values and information signaled through dialectal use of language [115]. "Clean" text additionally has been misrepresented as being "unbiased" against any community—a notion that has been strongly contested [22, 73, 130]. Unsurprisingly, gendered and racial disparities have been documented in a number of language technologies [37, 79, 141], and processes of creating resources and technologies may further entrench such disparities [25, 38, 144]. For more detail see [53].

While social and linguistic discrimination do not originate or end in language technologies, such technologies do engage in society as sociotechnical systems that are imbued with values [18], and it is therefore important to consider their role in discrimination, and the ways in which values of non-discrimination can be implemented when governing data. To this effect, we should be cognizant of the existing linguistic discrimination present in our societies [118] and be careful not to inadvertently replicate them [132, 133]. In the context of the various interactions of data governance, sourcing, and tooling mentioned in Section 2.4, this requires prioritizing the needs of currently under-represented language communities in their sourcing efforts, promoting notions of data quality that do not confound noise with sociolinguistic variation [131], and explicitly including and giving a say in the various governance choices to speakers of all language variants. The role of these data advocates is further outlined in Section 5.

### 3.2 Legal Context: Rights and Regulations

Figure 2 presents laws and regulations as one of the pillars of governance. In particular, the notion of **protected rights** can help us understand how the guiding values presented in Table 1 are understood and regulated in various legal contexts. The global landscape of relevant laws is vast, but in this section we provide a brief overview of how the values of just rewards, attribution, and contestation are related to the **property rights**, consent and privacy - to **privacy rights**, and non-discrimination - to **user rights**.

First, we examine **property rights** for language data creators. In the U.S., property rights are often thought of as a "bundle of sticks" [66]. That is, property rights are composed of different types of rights: the right to profit from the property (i.e., receive just rewards), to require proper crediting for re-use (i.e., attribution), etc. For example, an artist is entitled to fully profit from their work, or to remove it from circulation at any time. This bundle of rights comes with some common limitations. Copyright, trademarks, and patents can expire and "fair use" exemptions to copyright exist to allow certain uses of copyrighted data deemed socially beneficial, such as keeping content from disappearing [84]. In the context of ML, there is an ongoing debate about whether and when using copyrighted data for training models constitutes "fair use" [85]. The U.S. Copyright Office recently issued an exemption to liability for removing digital rights management software for the purposes of text and data mining for non-commercial research.[7] Japan and Europe have passed similar legislation making it easier to use data for text and data mining for research purposes[8] [9]; this tension between social benefits from allowing re-use of data and the social harms to data creators has led some in the U.S. to call this a "fair use crisis" [127]. The role of a governance structure will be to help data creators, hosts, and modelers navigate these tensions by providing locally relevant frameworks for contestation and use case restrictions.

[7] https://public-inspection.federalregister.gov/2021-23311.pdf
[8] https://eare.eu/japan-amends-tdm-exception-copyright/
[9] https://www.europarl.europa.eu/news/en/press-room/20190321IPR32110/european-parliament-approves-new-copyright-rules-for-the-internet

Second, we examine **privacy rights** of data creators. The view of privacy protection based on data as property [75, 119, 129] has been criticized as placing a substantial burden on the free flow of information, while potentially not improving privacy protections [78]. An alternative approach to privacy is restricting the processing of "personal data", as it is done in the E.U.'s General Data Protection Regulation (GDPR) [49]. This approach hinges on defining what is "personal", and how that interacts with "publicly available". For digital language data, a big issue is that most longer texts are unique and difficult to anonymize [31, 81, 96], and by themselves could identify people: e.g. a simple search could identify the author of a tweet, who willingly made the authorship information public. Since that author may not be aware of the visibility of their language data and conceptions of the social benefits (or lack thereof) of various research practices [54], privacy legislation such as GDPR may require that data be used with *(revocable) consent*, and for the *specific purposes* that are clearly explained to the data subjects, who should also have the right to delete or rectify existing records (so as to enable e.g. factual corrections, updates to the previously accurate records, or the 'right to be forgotten'). Given that trained ML models might be queried for specific information about individuals [26], a governance model would have to consider not only whether and how to remove specific instances from its datasets, but also how to minimize the risk of memorization when sharing the data for model training and development.

Third, we examine **user rights**. Depending on the jurisdiction, there may be an orthogonal set of laws that aims to ensure the rights of the *users of models created from the data*. A number of prior works, particularly from a U.S.-centric perspective, have connected ML to legal frameworks for human rights, especially anti-discrimination [47, 48, 64, 70, 72, 140, 142, 146]. These works often focus on the difficulties of constraining algorithmic discrimination in many contexts, proposing alternative legal frameworks that would allow for more regulatory enforcement of algorithmic bias. The evolving nature of human rights law, civil rights law, and ML may place more constraints on data curators to ensure that downstream models are more fair – and respect rights like equal protection, anti-discrimination, or constraints on arbitrary enforcement. For example, New York City now regulates automated employment algorithms and would require yearly bias audits.[10] However, the effectiveness of these relatively new laws has yet to be tested, and in the past governments themselves have tried to leverage ML systems in potential violations of human rights. [11] [12] We refer the reader to cited works for more in-depth analysis of these issues, including: accessibility rights, [13] a right to explanation, [14] and a right to a certain level of performance. [15] Data governance supports these user rights by allowing marginalized populations better control over how they are represented in the data used to train ML systems in an effort to lessen algorithmic discrimination, and by supporting auditability of these systems to promote accountability [4].

### 3.3 Machine Learning Context: Challenges and Incentives

One of the major challenges in creating a data governance structure for ML datasets lies in the limited amount of research on this subject within the ML community. Very recent research — most published within the last year — has

[10] Administrative Code of the City of New York, Title 20, Section 1, Chapter 5, Subchapter 25.
[11] https://notechforice.com/wp-content/uploads/2021/10/Deadly.Digital.Border.Wall_.pdf
[12] https://www.reuters.com/world/china/china-uses-ai-software-improve-its-surveillance-capabilities-2022-04-08/
[13] For example, the Americans with Disabilities Act of 1990 (ADA) 42 U.S.C. §§12101-12213, in the U.S. enabled the National Association of the Deaf to argue that automated captions in some cases were of such unacceptable quality that they did not satisfy the accessibility rights of deaf data users. *See, e.g., National Ass'n of the Deaf v. Harvard University*, 377 F. Supp. 3d 49 (D. Mass. 2019); *National Ass'n of the Deaf v. Netflix, Inc.*, 869 F. Supp. 2d 196 (D. Mass. 2012).
[14] GDPR [49] does not explicitly guarantee it, but it it does require the data processor to provide 'meaningful information about the logic involved' in fully automated decisions, which could be interpreted that way [126]. In May 2021 a Dutch court upheld this principle for the first time: a ridesharing company was obliged to "communicate the main assessment criteria and their role in the automated decision [to the drivers], so that they can understand the criteria on the basis of which the decisions were taken and they are able to check the correctness and lawfulness of the data processing" [101].
[15] The current proposal for the EU AI Act [50] distinguishes between application areas on the basis of risk they pose, and would institute external "conformity assessments" for the more risky applications.

begun to analyze dataset values [18, 40], question assumptions around dataset use [100], unpack what is represented in ML datasets [43, 88, 124] and establish basics of how an organized dataset lifecycle might proceed [74]. These just begin to scratch the surface of what well-defined data systems may look like in ML.

We see several reasons for the limited attention to data governance in ML to date. First, the mainstream ML research focuses predominantly on improvements to the model architecture, training procedure, and (hyper)parameters [100, 123]. For LLMs in particular, the data used to train them are one further step removed from the task-specific models built from them, so the link between data and ML progress is even more abstracted [87, 116]. Second, research addressing dataset choices, creation, and curation, is systematically "under-valued and de-glamorised" [3, 123][16]. Even works that do include significant curation efforts for the sake of improving models [57, 113] focus on definitions of quality that prioritize technical performance over the agency of data and algorithm subjects, which can result in widespread data that proliferates misogyny, pornography without consent, and malignant stereotypes [19].

One approach put forward in recent years to foster more accountability of these data practices has been documentation standards for data and models in natural language processing [11, 59] and ML in general [93]. There has also been an increased focus on analyzing other dimensions of data quality and stewardship [102, 107, 121, 123], with several noteworthy initiatives aiming to document both existing [9, 20, 30, 43], and newly developed [16, 60, 136] resources. These efforts have gone hand-in-hand with efforts centered around values of *transparency and replicability* in scientific work, through the introduction of standards and conference checklists [42, 105]. [17] The two directions have come together in the last year to extend the approach beyond simply reproducibility, with newer checklists for "Responsible NLP" [15, 27, 121] asserting the importance of respecting values including non-discrimination (fairness), consent, or privacy in the development and use of datasets and encouraging intentional handling of data tools (see Section 2.4). Given the importance of conferences in the field, we may hope that these paper checklists will have a significant role to play in spreading norms and best practices of data curation and documentation. Still, within this context, comparatively little attention is paid to the later stages of the data life cycle (see Section 2.3), or to data management models that intentionally include data subjects. We review common approaches to hosting and distributing ML data in Section 4.

## 4 Efforts and Challenges in ML Data Governance

Despite its unquestionable importance in contributing towards higher-quality LLMs and stakeholder agency, explicit data *governance* remains a relatively new field of practice in the ML and NLP communities. In this Section, we first survey existing data management efforts in AI, then provide a short description of the data governance practices in the Wikimedia project (Section 4), an example of a governance framework with goals and priorities similar to ours.

**Centralized Dataset Management.** Perhaps the most common method for managing NLP datasets is for the developers themselves to host the data upon release on platforms such as GitHub and personal websites. Commonly used larger organizations include Microsoft Research Open Data and Allen Institute for AI Datasets, as well as consortia such as the Linguistic Data Consortium (LDC), European Language Resources Association (ELRA) and CLARIN [76], which aim to centralize and standardize access to textual resources for members of the community. An advantage of such centralized repositories is that members can access a wide range of datasets that persist unchanged over time. For example, any researcher who downloads the popular OntoNotes will have the same version as other researchers, enabling reproducibility and fair comparisons. There are also downsides (such as membership cost or time lags), but critically for this work, there is no place for multiple stakeholders and rights-holders to align on priorities, giving the

[16] For a direct example of how the ML community treats work on datasets and values, see reviews for [3] here
[17] mostly geared towards code and experiment tracking, but also covering training and evaluation data

governing organization full say over how the data should be shared and used. Data subjects and providers generally do not have visibility into the data decisions, nor recourse to address how they are represented, and centralized decisions regarding content do not necessarily account for knowledge local to where the data instance is sourced.

**Public Dataset Repositories.** In recent years public repositories of datasets, like the UCI ML Repository [45] and the Hugging Face Dataset Repository [86], have become popular. These repositories resemble centralized dataset management, but rely primarily on user contributions, both to source datasets and to govern them. For example, dataset submitters must independently determine whether or not they have appropriate legal grounds to use the data, something they frequently lack the resources or expertise to do. In practice, compliance is hard to enforce, and while datasets are increasingly accompanied by datasheets [59] and similar documentation, navigating the legal structures involved is not always straightforward. Public repositories do present a unique opportunity to help harmonize emerging standards around documentation [91] mentioned in Section 3.3, but they are structurally unable to support the oversight and management that are essential to our purposes. Our values of autonomy, consent, and contestation are difficult if not practically impossible for public dataset repositories, due to the full reliance on self-governance by dataset submitters (but see the Wikimedia model for related mechanisms for content curation, Section 4).

**Open Data Initiatives.** In NLP, open data initiatives involve collecting, processing, and sharing data that is public, but inaccessible or difficult to use [57]. Some prominent open data initiatives have developed in response to the practice at many companies of training ML models on unreleased data, including OpenWebText [63], which seeks to replicate the dataset that GPT-2 [112] was trained on; BookCorpus2 [57] and Smashwords21 [9], which seek to replicate the formerly public BookCorpus dataset [145]; and LAION-400M [125] which seeks to replicate the WebImageText dataset that CLIP and DALL-E were trained with [110]. Another form of data replication effort seeks to provide public access to previously privately held data. C4, the dataset that the T5 language model was trained on [113], went unreleased for almost two years until it was replicated and shared by the Allen AI institute, enabling other scholars to study it and use it for training their own models. Open data initiatives meet many of our desiderata for data governance, but possess some key omissions. Critically, the goals of reproducible research that underlie the public recording of datasets are inherently in tension with the need to update datasets to accommodate requests to remove personal information [121], and unredacted copies may circulate for years [35].

**Example: Distributed Data Governance in the Wikimedia Project.** The Wikimedia projects offer a wealth of experience in highly collaborative and largely self-regulated data curation [56], similar to the goals in the proposed governance structure. The core stakeholders map to the proposed governance structure in Figure 1 as follows: the many contributors to the knowledge that is gathered on Wikimedia projects (data rights holders), editors (data custodians), the Wikimedia Foundation (data stewards and helpers), and the researchers, digital platforms, and many additional end-users of Wikimedia content (data modelers). The Wikimedia projects face many of the same tensions that would face governance of global digital language data, such as diverse needs and goals of editors [109], the need to navigate varying local laws such as "freedom of panorama" [138] when determining whether an image can be hosted [39], and how they are situated within existing power imbalances [135].

For example, to create some consistency for editors and end-users of Wikimedia data, the data is governed in part through content licenses. Content licenses vary in attribution requirements between projects, but restrict contributors' rights on how their work is used. This can be in conflict with cultural values, e.g., in the case of indigenous communities that are generally underrepresented on Wikipedia but have concerns about how their knowledge might be exploited if shared [28, 29]. To ensure that the content adheres to the chosen licenses (and other regulations [34]), editors have

| | |
|---|---|
| **Data Rights-holders** | Decide whether to share their data. |
| **Data Providers** | Make data available to others. |
| **Data Hosts** | Gather and hold data aligned to constraints defined within the governance structure. |
| **Data Modelers** | Specify dataset values and requirements. |
| **Data Stewardship Org.** | Discussion space for all actors involved. |
| **Data Helpers** | Ensure decisions respect rights and regulations |

Table 4. Actors within the proposed Data Governance structure

| | |
|---|---|
| **Data Modelers + Data Hosts** | Data dissemination<br>Specific use case restrictions |
| **Data Hosts + Data Providers** | Data dissemination<br>Conditions for serving data<br>Rights with respect to derived products |

Table 5. Binding agreements needed in the Data Governance Structure.

written policies (norms as in Section 2.2) that are constantly evolving and being contested themselves [14, 24]. Similar to the proposed DSO, the success of the Wikimedia editor community is facilitated by a large ecosystem of tools [61, 97] (as in Section 2.4) such as APIs, dumps, database replicas, and various cloud environments that can be used by tool developers to provide local access to this data [33]. The ability for the community to build the tools required for data governance has been crucial to their success at scale [67].

## 5 A New Data Governance Structure

Let us now review the needs we have identified for a governance structure in Sections 2, 3, and 4. We want an organization driven by a set of guiding values outlined in Table 1, and notably the inclusion & representation of all categories of stakeholders identified in Table 3, in a fashion that fosters equitable access across social, cultural and geographical contexts (Section 3.1). In so doing, the governance structure needs to account for the complexity and diversity of corresponding legal contexts (3.2). We reviewed some issues and promising directions around the culture of data use in ML (3.3) and current approaches to data management in the field (4), and found coordination across stakeholders following the desiderata detailed above to be a particular challenge.

The need to collect, share, access, and define norms, management, policies, guidelines, and values around the use of data suggests a structure with multiple categories of distributed actors prioritizing different aspects and communicating with one another for alignment on end goals, legal issues, values, and interoperability. To that end, we propose a data governance structure with six main actors, whose roles and relationships are summarized in Figure 1. The actors additionally interact with Data Sourcing and Data Tooling, as discussed in Section 2.4. In this Section, we start by describing the specific roles of the **data governance entities** involved in this structure. We then review the relationships between these entities through two lenses: the **journey of the data** through the structure from its initial creators to the data modelers and the **role of the DSO** in formalizing frameworks and aggregating feedback and expressions of the various stakeholders' needs, especially with the aim of fostering the values in Table 1.

**Data Governance Entities** Table 4 summarizes the roles around which we organize the governance structure; specific entities may take **one or more** of these roles at various times (e.g., a data modeler may also make their own dataset and entrust it back to a data host as data provider). Figure 1 maps some of these roles to traditional categorization of data governance, including data steward and data custodian. We review each of these roles next.

Our effort toward defining data governance roles starts with asking *where the data is found*, and *whose rights* need to be accounted for. **Data Rights-holders** are varied: they can be individuals, organizations or companies. An individual who wrote on social media, for example, might have legally protected privacy rights on their language data used in a dataset (Section 3.2), and organizations such as radio stations, newspapers, or content platforms have property rights on the data they create or host. In general, the Data Rights-holders correspond to the *data subjects* and *data creators* categories of stakeholders in Table 3 and represent the focus of the values of contestation, consent, privacy, attribution,

and just rewards described in Table 1. In particular, Rights-holders can inform how specific items of their language data may be used, in accordance with legal protections and values.

Data is brought into our proposed governance structure by **Data Providers**. Companies that host or create language data can act as Data Providers, as can research organizations that create datasets from public or private data or archival institutions that work on preserving online or offline content (*e.g.*, the Internet Archive). The Data Providers can be identical to or separate from the Data Rights-holders, and can either fully specify what the data they bring into the governance structure may be used for, or specify it to the extent permitted by the original rights-holders.

Data is served by **Data Hosts** who gather and hold data from the Data Providers so as to meet the goals of the governance project and comply with legal requirements. This data is in turn made available to **Data Modelers**. Data Hosts maintain their own, possibly post-processed version of the language data offered by data providers, and can decide which data they want to host (*i.e.* they may decline to host some of the data offered by a data provider). Depending on the jurisdiction of the Data Host and Data Provider, the Hosts may need a specific legal basis for holding certain kind of data or being eligible for some of the research exceptions outlined in Section 3.2, which may go from being a non-profit organization to having some form of public interest mission. As outlined in Section 2.4, Data Sourcing happens at the intersection between Data Providers and Data Hosts; the diversity and representativity of the data available to the governance organization will depend on the ability of the hosts to establish relationships and support the need of the greater variety of data providers. Notably, this is easier when there is a degree of proximity between the hosts and providers so they have similar social and legal contexts —which motivates the need to have data hosts around the world to foster linguistic and cultural diversity of the available language data. This proximity is also necessary to enacting meaningful contestation rights at the Data Host level, as it will allow the requester (Data Rights-holders) and the enacter (Data Host) to share similar understandings of the notion and rely on a similar legal framework (as the extraterritorial applicability of data protection laws like the GDPR around the world is still an open question).

The last category of distributed actors of our proposed data governance organization are the **Data Modelers**, who can request access to data held by the Data Hosts to use according to the requirements set forth by the Data Rights-holders, Data Providers, and Data Hosts. The Data Modelers have their own data needs, including visibility of the data available across the data hosts, and ease of processing (e.g. through a unified format for all data sources). Researchers may also need some degree of replicability for experiments run using the data held with the governance organization (see Section 3.3), which needs to be understood in the context of the contestation rights within the organization.

Finally, a **Data Stewardship Organization** provides a discussion space for all the above-mentioned actors and connections involved, communicating between Hosts, Providers, Data Advocates, and Data Rights-holders. The DSO brings together representatives of all of the other roles, and is supported by **Data Helpers**, including lawyers and legal scholars representing all regions where the governance organization operates and advocacy groups focused on representing the interests of populations affected by data use and technology [18]. The role of the DSO in establishing and formalizing relationships between the actors is further outlined below, and it also serves as a central repository of technical tools, relevant documentations, and as a facilitator of interoperability. Given this role, both the input of the Lawyers and of the Data Advocates is requested on new choices to account for both relevant regulations and their impact on the values listed in Table 1.

**Journey of the Data.** Before eventually being used for research or development of NLP system, the data follows a journey from its original creators, to the Data Providers who introduce it to the governance structure, to the Data

[18] Data 4 Black Lives, Our Data Bodies are two such organizations in the US.

Hosts that aggregate various sources, to the Data Modelers. Each of these transfers and nodes in the path defines its own agreements between parties. In particular, these agreements are structured around the notion of **contractual flow-down**: each subsequent actor on this path is responsible for communicating the requirements and restrictions formulated by its predecessors in addition to its own.

In the exchanges **between Data Hosts and Data Providers**, Providers make data available, and if they have full rights on the data, they may specify use conditions. Some selection criteria might include research only, or use by organizations that meet certain criteria (such as non-profit status, or value statements). These restrictions may be *explicitly* set down in a license agreement signed by the Host and the Provider (Table 5). As an additional incentive for Providers to share their data, this license may require the Host to give the Provider access to any by-product of their data, such as analyses or processed versions (*e.g.* with PII removed; see Section 2.4). If the Provider is proposing data that is curated from external sources, especially text that is regulated by general data protection laws, there is also an *implicit* relationship **between Data Host and Rights-holders**. In particular, the Data Host will be bound to honor contestation requests when an individual finds that their private information, or data that they have a commercial right to, is included in data shared with the Host without their explicit consent. The Host would then be required to remove the particular data items from their dataset. This aspect requires Hosts to share data via access restrictions or binding agreements, as opposed to allowing copies to be freely downloaded and proliferated.

Finally, the exchange **between Data Hosts and Data Modelers** is bound by another set of licensing agreements, which need to reflect the restrictions flowing down from the Data Providers and other Rights-holders, any additional constraints expressed by the Data Host, and a non-dissemination clause. The Modeler may be required to obtain a fresh version of any dataset reflecting the most recent version of the data host after a fixed amount of time. The latter two are essential to ensure that data that has been deleted from a Host to answer a contestation request, or whose license with the Data Provider has expired, does not remain available.

**Role of the DSO.** While the DSO itself is not a direct party in any of the agreements outlined above, its role is to facilitate interactions between all entities involved and assure interoperability between actors. For example, Data Providers might have reason to propose their own license, especially to support values that are misrepresented in legal frameworks relevant to them [19]. In such cases, respecting our value of inclusivity and the local knowledge of the Providers on how best to represent their community's interests means allowing them to use their own licensing in their interactions with Data Hosts rather then requiring them to use one designed by the DSO. Conversely, some of the Data Providers, especially ones with fewer proper resources, might not have the legal expertise to develop their own data sharing licenses – then, a DSO standard data license would foster inclusion. We provide our proposal for such an agreement between the Data Hosts and Data Providers in Appendix B.

This tension reflects the governance trade-off between harmonization and independence mentioned in Section 2.1. One additional complexity of allowing Hosts and Providers to use custom licenses arises when a Host aggregates data from several Providers to share with a Data Modeler. Without any categorization of the various Provider licenses, the Host would have to either develop a new custom license for each aggregation of data sources, or leave it to the Modeler to understand the interplay of the various constraints. We address these issues through a dual approach. First, the DSO provides a license template for use in exchanges between the Data Providers and Data Hosts. Second, the DSO *maintains a taxonomy of licenses* designed to support rules for aggregating use case restrictions from Providers for the agreement between the Hosts and Modelers, to be updated when a new license appears that is not easily categorized.[20]

[19] Maori Data Sovereignty License
[20] CLARIN (Section 4) uses a restricted version of such a taxonomy for non-commercial data only [76]

This approach exemplifies the general role of the Data Stewardship Organization at various points of exchange in the governance organization, in its two roles as a fallback mechanism or default option for actors that do not have the resources to develop their own processes, and as an enabler of interoperability between processes when they do: whether the process in question corresponds to the license agreement between the Data Host and Data Modelers, the framework for identifying a legitimate contestation request between a Data Host and Rights-holders, or the technical tools used for identifying personal information or managing access to data.

## 6 Discussion

We have introduced an approach to data governance, grounded in an ongoing year-long case study that coordinates data internationally to train a Large Language Model. Critical aspects of the governance structure include protocols for achieving different values, working with established norms, and contending with the different laws applicable across datasets. This requires coordinating multiple stake-holders and rights-holders. Our approach is modular, where different parties focus on different aspects of the dataset processing and sharing, interconnecting data providers, data hosts, and data developers. This is coordinated by a Data Stewardship Organization that develops appropriate management plans, access restrictions, and legal scholarship. A complementary Data Tooling efforts help to provide resources common to the legal and ethical needs of the participating institutions, We have found that one of the most difficult hurdles is developing legal agreements for Providers, Hosts, and Modelers that respect the laws and copyrights set forth in the data, as well as the laws of the institutions' regions. We tackle this problem through the lens of stated governance *values*, which inform the kinds of agreements that are necessary.

## References


[1] 2018. REGULATION (EU) 2018/1725 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32018R1725&from=EN

[2] 2020. Data Governance Act. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52020PC0767&from=EN

[3] 2021. "The Values Encoded in Machine Learning Research" Reviews. https://openreview.net/forum?id=oioB7Te7Bo

[4] The Ada Lovelace Institute, The AI Now Institute, and The Open Government Partnership. 2021. Algorithmic accountability for the public sector. http://www.opengovpartnership.org/documents/

[5] Rebecca Adami. 2014. Human rights for more than one voice: rethinking political space beyond the global/local divide. *Ethics & Global Politics* 7 (2014), 163 – 180.

[6] UN General Assembly. 1966. International covenant on civil and political rights. *United Nations, Treaty Series* 999 (1966), 171.

[7] UN General Assembly. 1966. International covenant on economic, social and cultural rights. *United Nations, Treaty Series* 993, 3 (1966), 2009–2057.

[8] United Nations. General Assembly. 1949. *Universal declaration of human rights*. Vol. 3381. Department of State, United States of America.

[9] Jack Bandy and Nicholas Vincent. 2021. Addressing "Documentation Debt" in Machine Learning Research: A Retrospective Datasheet for BookCorpus. *arXiv preprint arXiv:2105.05241* (2021). https://arxiv.org/abs/2105.05241

[10] Michael Barnett and Raymond Duvall. 2004. *Power in global governance*. Vol. 98. Cambridge University Press.

[11] Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics* 6 (2018), 587–604. https://doi.org/10.1162/tacl_a_00041

[12] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021*, Madeleine Clare Elish, William Isaac, and Richard S. Zemel (Eds.). ACM, 610–623. https://doi.org/10.1145/3442188.3445922

[13] Misha Benjamin, Paul Gagnon, Negar Rostamzadeh, Chris Pal, Yoshua Bengio, and Alex Shee. 2019. Towards Standardization of Data Licenses: The Montreal Data License. *CoRR* abs/1903.12262 (2019). arXiv:1903.12262 http://arxiv.org/abs/1903.12262

[14] Amber Berson, Monika Sengul-Jones, and Melissa Tamani. 2021. Unreliable Guidelines: Reliable sources and marginalized communities in French, English, and Spanish Wikipedias. https://artandfeminism.org/wp-content/uploads/2021/06/Unrealiable-Guidelines_Final.pdf

[15] Alina Beygelzimer, Yann Dauphin, Percy Liang, and Jennifer Wortman Vaughan. 2021. Introducing the NeurIPS 2021 Paper Checklist. https://neuripsconf.medium.com/introducing-the-neurips-2021-paper-checklist-3220d6df500b

[16] Stella Biderman, Kieran Bicheno, and Leo Gao. 2022. Datasheet for the Pile. *arXiv preprint arXiv:2201.07311* (2022).

[17] Abeba Birhane. 2021. Algorithmic injustice: a relational ethics approach. *Patterns* 2 (2021).

[18] Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2021. The values encoded in machine learning research. *arXiv preprint arXiv:2106.15590* (2021).

[19] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. 2021. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *CoRR* abs/2110.01963 (2021). arXiv:2110.01963 https://arxiv.org/abs/2110.01963

[20] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. 2021. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *ArXiv* abs/2110.01963 (2021).

[21] Damián E. Blasi, Antonios Anastasopoulos, and Graham Neubig. 2021. Systematic Inequalities in Language Technology Performance across the World's Langu ages. *CoRR* abs/2110.06733 (2021). arXiv:2110.06733 https://arxiv.org/abs/2110.06733

[22] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. arXiv:1607.06520 [cs.CL]

[23] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165* (2020).

[24] Brian Butler, Elisabeth Joyce, and Jacqueline Pike. 2008. Don't look now, but we've created a bureaucracy: the nature and roles of policies and rules in wikipedia. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1101–1110.

[25] Yang Trista Cao and Hal Daumé III. 2021. Toward Gender-Inclusive Coreference Resolution: An Analysis of Gender and Bias Throughout the Machine Learning Lifecycle. *Comput. Linguistics* 47, 3 (2021), 615–661. https://doi.org/10.1162/coli_a_00413

[26] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2020. Extracting training data from large language models. *arXiv preprint arXiv:2012.07805* (2020).

[27] Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz. 2021. Responsible NLP Research Checklist. http://aclrollingreview.org/responsibleNLPresearch/

[28] Stephanie Russo Carroll, Ibrahim Garba, Oscar L Figueroa-Rodríguez, Jarita Holbrook, Raymond Lovett, Simeon Materechera, Mark Parsons, Kay Raseroka, Desi Rodriguez-Lonebear, Robyn Rowe, et al. 2020. The CARE Principles for Indigenous Data Governance. *Data Science Journal* (2020).

[29] Nathalie Casemajor, Christian Coocoo, and Karine Gentelet. 2019. Openness, inclusion and self-affirmation: Indigenous knowledge in open knowledge projects. *Journal of Peer Production* 13 (2019).

[30] Isaac Caswell, Julia Kreutzer, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Javier Ortiz Suárez, Iroro Orife, Kelechi Ogueji, Rubungo Andre Niyongabo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2021. Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *arXiv:2103.12028 [cs]* (April 2021). arXiv:2103.12028 [cs] http://arxiv.org/abs/2103.12028

[31] Carole E Chaski. 2013. Best practices and admissibility of forensic author identification. *Journal of law and policy* 21, 2 (2013), 332–376.

[32] The Commission on Global Governance. 1995. *Our global neighbourhood: the report of the Commission on Global Governance*. Oxford University Press.

[33] Wikitech contributors. 2021. Help:Cloud Services Introduction — Wikitech,. https://wikitech.wikimedia.org/w/index.php?title=Help:Cloud_Services_Introduction&oldid=1929345 [Online; accessed 18-January-2022].

[34] Wikipedia contributors. 2021. Wikipedia:Five pillars. https://en.wikipedia.org/w/index.php?title=Wikipedia:Five_pillars&oldid=1060800392 [Online; accessed 18-January-2022].

[35] Frances Corry, Hamsini Sridharan, Alexandra Luccioni, Mike Ananny, Jason Schultz, and Kate Crawford. 2021. The Problem of Zombie Datasets: A Framework For Deprecating Datasets. *ArXiv* abs/2111.04424 (2021).

[36] Cristiana Cremona and Elizabeth Bates. 1977. The development of attitudes toward dialect in Italian children. *Journal of Psycholinguistic Research* 6, 3 (1977), 223–232.

[37] Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2019. Plug and play language models: A simple approach to controlled text generation. *arXiv preprint arXiv:1912.02164* (2019).

[38] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial Bias in Hate Speech and Abusive Language Detection Datasets. In *Proceedings of the Third Workshop on Abusive Language Online*. Association for Computational Linguistics, 25–35. https://doi.org/10.18653/v1/W19-3504

[39] Mélanie Dulong De Rosnay and Pierre-Carl Langlais. 2017. Public artworks and the freedom of panorama controversy: a case of Wikimedia influence. *Internet Policy Review* 6, 1 (2017).

[40] Emily Denton, Alex Hanna, Razvan Amironesei, Andrew Smart, and Hilary Nicole. 2021. On the genealogy of machine learning datasets: A critical history of ImageNet. *Big Data & Society* 8, 2 (2021), 20539517211035955.

[41] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 4171–4186. https://doi.org/10.18653/v1/n19-1423

[42] Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. 2019. Show your work: Improved reporting of experimental results. *arXiv pre-print* 1909.03004 (2019), 1–21. https://arxiv.org/abs/1909.03004.

[43] Jesse Dodge, Maarten Sap, Ana Marasovic, William Agnew, Gabriel Ilharco, Dirk Groeneveld, and Matt Gardner. 2021. Documenting the English Colossal Clean Crawled Corpus. *arXiv:2104.08758 [cs]* (April 2021). arXiv:2104.08758 [cs] http://arxiv.org/abs/2104.08758

[44] Mary Douglas. 1978. *Purity and danger: an analysis of the concepts of pollution and taboo*. Routledge.

[45] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

[46] Kirk Emerson, Tina Nabatchi, and Stephen Balogh. 2012. An integrative framework for collaborative governance. *Journal of public administration research and theory* 22, 1 (2012), 1–29.

[47] David Freeman Engstrom and Daniel E Ho. 2020. Algorithmic accountability in the administrative state. *Yale J. on Reg.* 37 (2020), 800.

[48] David Freeman Engstrom, Daniel E Ho, and Cristina Isabel Ceballos. 2021. Disparate Limbo: How Administrative Law Erased Antidiscrimination. *Yale Law Journal, Forthcoming* (2021).

[49] European Commission. 2018. 2018 reform of EU data protection rules. https://ec.europa.eu/commission/sites/beta-political/files/data-protection-factsheet-changes_en.pdf

[50] European Commission. 2021. *Proposal for a Regulation Laying down Harmonised Rules on Artificial Intelligence*. Technical Report 2021/0106 (COD). European Comission, Brussels. https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_1&format=PDF

[51] Frantz Fanon. 1952. Black Skin, White Masks.

[52] Richard C Feiock. 2013. The institutional collective action framework. *Policy Studies Journal* 41, 3 (2013), 397–425.

[53] Anjalie Field, Su Lin Blodgett, Zeerak Waseem, and Yulia Tsvetkov. 2021. A Survey of Race, Racism, and Anti-Racism in NLP. arXiv:2106.11410 [cs.CL]

[54] Casey Fiesler and Nicholas Proferes. 2018. "Participant" Perceptions of Twitter Research Ethics. *Social Media + Society* 4 (2018).

[55] Jessica Fjeld, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar. 2020. Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. *Berkman Klein Center Research Publication* (2020).

[56] Andrea Forte, Vanesa Larco, and Amy Bruckman. 2009. Decentralization in Wikipedia governance. *Journal of Management Information Systems* 26, 1 (2009), 49–72.

[57] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2021. The Pile: An 800GB Dataset of Diverse Text for Language Modeling. *CoRR* abs/2101.00027 (2021). arXiv:2101.00027 https://arxiv.org/abs/2101.00027

[58] Cristina García-Bermejo Gallego. 2015. A Sociolinguistic Approach to Vernacular Varieties: Stigmas and Prejudices in the Case of the West Country Dialect. (2015).

[59] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2020. Datasheets for Datasets. *arXiv:1803.09010 [cs]* (March 2020). arXiv:1803.09010 [cs] http://arxiv.org/abs/1803.09010

[60] Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Aremu Anuoluwapo, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna Clinciu, Dipanjan Das, Kaustubh D Dhole, et al. 2021. The gem benchmark: Natural language generation, its evaluation and metrics. *arXiv preprint arXiv:2102.01672* (2021).

[61] R Stuart Geiger and David Ribes. 2010. The work of sustaining order in Wikipedia: The banning of a vandal. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*. 117–126.

[62] Robert LeRoy Giron. 1982. Chicano Spanish: Cross Hispanic Language Attitudes toward Specific Lexical Items. (1982).

[63] Aaron Gokaslan and Vanya Cohen. 2019. OpenWebText Corpus. http://Skylion007.github.io/OpenWebTextCorpus.

[64] Jamie Grace. 2019. Machine Learning Technologies and Human Rights in Criminal Justice Contexts. *Available at SSRN 3487454* (2019).

[65] Daniel Greene, Anna Lauren Hoffmann, and Luke Stark. 2019. Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning. In *HICSS*.

[66] Thomas C Grey. 2014. 2 The Disintegration of Property. In *Formalism and Pragmatism in American Law*. Brill, 30–45.

[67] Aaron Halfaker and R Stuart Geiger. 2020. Ores: Lowering barriers with participatory machine learning in wikipedia. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–37.

[68] Donna Harawy. 1988. Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. *Feminist Studies* 14 (1988).

[69] Ruqaiya Hasan. 2009. *Semantic Variation: Meaning in society and in sociolinguistics*. Vol. 2. Citeseer.

[70] Daniel E Ho and Alice Xiang. 2020. Affirmative Algorithms: The Legal Grounds for Fairness as Awareness. *U. Chi. L. Rev. Online* (2020), 134.

[71] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and L. Sifre. 2022. Training Compute-Optimal Large Language Models. *ArXiv* abs/2203.15556 (2022).

[72] Aziz Z Huq. 2019. Constitutional Rights in the Machine-Learning State. *Cornell L. Rev.* 105 (2019), 1875.

[73] Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social biases in NLP models as barriers for persons with disabilities. *arXiv preprint arXiv:2005.00813* (2020).

[74] Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. 2021. Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 560–575. https://doi.org/10.1145/3442188.3445918

[75] Paulius Jurcys, Christopher Donewald, Mark Fenwick, Markus Lampinen, and Andrius Smaliukas. 2020. Ownership of User-Held Data: Why Property Law Is the Right Approach. *Harvard Journal of Law and Technology Digest [2021]* (2020).

[76] A. Kelli, Krister Lindén, Kadri Vider, P. Labropoulou, E. Ketzan, Pawel Kamocki, and P. Stranák. 2018. Implementation of an Open Science Policy in the context of management of CLARIN language resources. In *CLARIN Annual Conference 2018*.

[77] Robert O Keohane and David G Victor. 2011. The regime complex for climate change. *Perspectives on politics* 9, 1 (2011), 7–23.

[78] Cameron F Kerry and John B Morris. 2019. Why data ownership is the wrong approach to protecting privacy. *Brookings. https://www. brookings. edu/blog/techtank/2019/06/26/why-data-ownership-is-the-wrong-approach-toprotecting-privacy* (2019).

[79] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences* 117, 14 (2020), 7684–7689.

[80] Jan Peter Kooiman and Svein Jentoft. 2009. META-GOVERNANCE: VALUES, NORMS AND PRINCIPLES, AND THE MAKING OF HARD CHOICES. *Public Administration* 87 (2009), 818–836.

[81] Srijan Kumar, Justin Cheng, Jure Leskovec, and VS Subrahmanian. 2017. An army of me: Sockpuppets in online discussion communities. In *Proceedings of the 26th International Conference on World Wide Web*. 857–866.

[82] William Labov. 1994. *Principles of Linguistic Change, Volume 1: Internal Factors*. John Wiley & Sons.

[83] Markku Lehtonen, Léa Sébastien, and Thomas Bauler. 2016. The multiple roles of sustainability indicators in informational governance: Between intended use and unanticipated influence. *Current Opinion in Environmental Sustainability* 18 (2016), 1–9.

[84] Mark A Lemley. 2020. Disappearing Content. *Available at SSRN 3715133* (2020).

[85] Mark A Lemley and Bryan Casey. 2020. Fair Learning. *Tex. L. Rev.* 99 (2020), 743.

[86] Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. 2021. Datasets: A Community Library for Natural Language Processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 175–184. arXiv:2109.02846 [cs.CL] https://aclanthology.org/2021.emnlp-demo.21

[87] Thomas Liao, Rohan Taori, Inioluwa Deborah Raji, and Ludwig Schmidt. 2021. Are We Learning Yet? A Meta Review of Evaluation Failures Across Machine Learning. In *Datasets and Benchmarks Proceedings at the 35th Conference on Neural Information Processing Systems (NeurIPS 2021)*, J. Vanschoren and S. Yeung (Eds.). Neural Information Processing Systems, San Diego, CA, 1–20. https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/757b505cfd34c64c85ca5b5690ee5293-Abstract-round2.html.

[88] Alexandra Sasha Luccioni and Joseph D Viviano. 2021. What's in the Box? An Analysis of Undesirable Content in the Common Crawl Corpus. *arXiv preprint arXiv:2105.02732* (2021).

[89] Nelson Maldonado-Torres. 2017. On the coloniality of human rights. *Revista Crítica de Ciências Sociais* (2017), 117–136.

[90] Angelina McMillan-Major, Zaid Alyafeai, Stella Biderman, Kimbo Chen, Francesco De Toni, Gerard Dupont, Hady Elsahar, Chris Emezue, Alham Fikri Aji, Suzana Ilic, Nurulaqilla Khamis, Colin Leong, Maraim Masoud, Aitor Soroa, Pedro Ortiz Suarez, Zeerak Talat, Daniel van Strien, and Yacine Jernite. 2022. Documenting Geographically and Contextually Diverse Data Sources: The BigScience Catalogue of Language Data and Resources. *In Submission.* (2022).

[91] Angelina McMillan-Major, Salomey Osei, Juan Diego Rodriguez, Pawan Sasanka Ammanamanchi, Sebastian Gehrmann, and Yacine Jernite. 2021. Reusable Templates and Guides For Documenting Datasets and Models for Natural Language Processing and Generation: A Case Study of the HuggingFace and GEM Data and Model Cards. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*. Association for Computational Linguistics, Online, 121–135. https://doi.org/10.18653/v1/2021.gem-1.11

[92] Budiman Minasny, Dian Fiantis, Budi Mulyanto, Yiyi Sulaeman, and Wirastuti Widyatmanti. 2020. Global soil science research collaboration in the 21st century: Time to end helicopter research. *Geoderma* 373 (2020), 114299.

[93] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 220–229. https://doi.org/10.1145/3287560.3287596

[94] Claudia Mitchell-Kernan. 1971. Language Behavior in a Black Urban Community. *Monograph of the Language-Behavior Research Laboratory, No. 2* (1971).

[95] Shakir Mohamed, Marie-Thérèse Png, and William S. Isaac. 2020. Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence. *ArXiv* abs/2007.04068 (2020).

[96] Maximilian Mozes and Bennett Kleinberg. 2021. No Intruder, no Validity: Evaluation Criteria for Privacy-Preserving Text Anonymization. *arXiv preprint arXiv:2103.09263* (2021).

[97] Claudia Müller-Birn, Leonhard Dobusch, and James D Herbsleb. 2013. Work-to-rule: the emergence of algorithmic governance in Wikipedia. In *Proceedings of the 6th International Conference on Communities and Technologies*. 80–89.

[98] Iroro Orife, Julia Kreutzer, Blessing Sibanda, Daniel Whitenack, Kathleen Siminyu, Laura Martinus, Jamiil Toure Ali, Jade Z. Abbott, Vukosi Marivate, Salomon Kabongo, Musie Meressa, Espoir Murhabazi, Orevaoghene Ahia, Elan Van Biljon, Arshath Ramkilowan, Adewale Akinfaderin, Alp Öktem, Wole Akin, Ghollah Kioko, Kevin Degila, Herman Kamper, Bonaventure Dossou, Chris Emezue, Kelechi Ogueji, and Abdallah Bashir. 2020. Masakhane - Machine Translation For Africa. *CoRR* abs/2003.11529 (2020). arXiv:2003.11529 https://arxiv.org/abs/2003.11529

[99] Amandine Orsini, Jean-Frédéric Morin, and Oran Young. 2013. Regime complexes: A buzz, a boom, or a boost for global governance. *Global governance* 19 (2013), 27.

[100] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. 2021. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns* 2, 11 (2021), 100336. https://doi.org/10.1016/j.patter.2021.100336

[101] Steve Peers. 2021. EU Law Analysis: The Ola & Uber Judgments: For the First Time a Court Recognises a GDPR Right to an Explanation for Algorithmic Decision-Making. http://eulawanalysis.blogspot.com/2021/04/the-ola-uber-judgments-for-first-time.html

[102] Kenny Peng, Arunesh Mathur, and A. Narayanan. 2021. Mitigating dataset harms requires stewardship: Lessons from 1000 papers. *ArXiv* abs/2108.02922 (2021).

[103] Ângela Guimarães Pereira, Jean Daniel Rinaudo, Paul Jeffrey, J E M Blasques, Serafin Corral Quintana, Nathalie Courtois, Silvio Funtowicz, and V. Petit. 2003. ICT Tools to Support Public Participation in Water Resources Governance & Planning: Experiences from the Design and Testing of a Multi-Media Platform. *Journal of Environmental Assessment Policy and Management* 05 (2003), 395–420.

[104] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, Marilyn A. Walker, Heng Ji, and Amanda Stent (Eds.). Association for Computational Linguistics, 2227–2237. https://doi.org/10.18653/v1/n18-1202

[105] Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché Buc, Emily Fox, and Hugo Larochelle. 2021. Improving reproducibility in machine learning research: a report from the NeurIPS 2019 reproducibility program. *Journal of Machine Learning Research* 22 (2021).

[106] Vinod Prabhakaran, Iason Gabriel, Timnit Gebru, and Margaret Mitchell. 2021. A Human Rights Approach to Responsible AI. (2021).

[107] Vinay Uday Prabhu and Abeba Birhane. 2020. Large image datasets: A pyrrhic win for computer vision? *arXiv preprint arXiv:2006.16923* (2020).

[108] Jill M Purdy. 2012. A framework for assessing power in collaborative governance processes. *Public administration review* 72, 3 (2012), 409–417.

[109] Anna C. Rader. 2020. *Why Do People Edit?* Technical Report. The Wikimedia Foundation. https://upload.wikimedia.org/wikipedia/commons/8/82/WDPE_Literature_Review_Anna_Rader.pdf

[110] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*.

[111] Alec Radford and Karthik Narasimhan. 2018. Improving Language Understanding by Generative Pre-Training.

[112] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[113] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683* (2019).

[114] Jacquelyn Rahman. 2008. Middle-class African Americans: Reactions and attitudes toward African American English. *American Speech* 83, 2 (2008), 141–176.

[115] Jacquelyn Rahman. 2012. The N Word: Its History and Use in the African American Community. *Journal of English Linguistics* 40, 2 (Jun 2012), 137–171. https://doi.org/10.1177/0075424211414807

[116] Inioluwa Deborah Raji, Emily M. Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. 2021. AI and the Everything in the Whole Wide World Benchmark. *arXiv pre-print* 2111.15366 (2021), 1–20. https://arxiv.org/abs/2111.15366.

[117] Steven Ratner. 2001. Corporations and Human Rights: A Theory of Legal Responsibility. *Yale Law Journal* 111 (2001), 1.

[118] John Rickford and Sharese King. 2016. Language and linguistics on trial: Hearing Rachel Jeantel (and other vernacular speakers) in the courtroom and beyond. *Language* 92 (12 2016), 948–988. https://doi.org/10.1353/lan.2016.0078

[119] Jeffrey Ritter and Anna Mayer. 2017. Regulating data as property: a new construct for moving forward. *Duke L. & Tech. Rev.* 16 (2017), 220.

[120] Anna Rogers. 2021. Changing the World by Changing the Data. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 2182–2194. https://aclanthology.org/2021.acl-long.170

[121] Anna Rogers, Tim Baldwin, and Kobi Leins. 2021. Just What do You Think You're Doing, Dave?' A Checklist for Responsible Data Use in NLP. arXiv:2109.06598 [cs.CL]

[122] Cornelia Roux and Petro Du Preez. 2013. Human rights literacy: A quest for meaning. *Retrieved* 20 (2013), 2017.

[123] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.

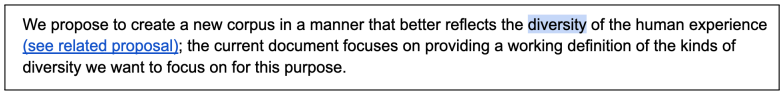


Fig. 4. Snippet from initial Data Governance planning doc, with *diversity* value highlighted.

[124] Morgan Klaus Scheuerman, Alex Hanna, and Emily Denton. 2021. Do datasets have politics? Disciplinary values in computer vision dataset development. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–37.

[125] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs. *arXiv preprint arXiv:2111.02114* (2021).

[126] Andrew D Selbst and Julia Powles. 2017. Meaningful Information and the Right to Explanation. *International Data Privacy Law* 7, 4 (Nov. 2017), 233–242. https://doi.org/10.1093/idpl/ipx022

[127] Benjamin LW Sobel. 2017. Artificial Intelligence's Fair Use Crisis. *Colum. JL & Arts* 41 (2017), 45.

[128] Luke Stark and Anna Lauren Hoffmann. 2019. Data Is the New What? Popular Metaphors & Professional Ethics in Emerging Data Culture. *Journal of Cultural Analytics* 4, 1 (1 5 2019). https://doi.org/10.22148/16.036

[129] Ivan Stepanov. 2020. Introducing a property right over data in the EU: the data producer's right–an evaluation. *International Review of Law, Computers & Technology* 34, 1 (2020), 65–86.

[130] Zeerak Talat, Smarika Lulz, Joachim Bingel, and Isabelle Augenstein. 2021. Disembodied Machine Learning: On the Illusion of Objectivity in NLP. *arXiv preprint arXiv:2101.11974* (Jan 2021). http://arxiv.org/abs/2101.11974

[131] Samson Tan. 2022. *Linguistically-Inclusive Natural Language Processing*. Ph. D. Dissertation. National University of Singapore.

[132] Samson Tan, Shafiq Joty, Min-Yen Kan, and Richard Socher. 2020. It's Morphin' Time! Combating Linguistic Discrimination with Inflectional Perturbations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 2920–2935. https://doi.org/10.18653/v1/2020.acl-main.263

[133] Nanna Thylstrup and Zeerak Talat. 2020. Detecting 'Dirt' and 'Toxicity': Rethinking Content Moderation as Pollution Behaviour. *SSRN Electronic Journal* (2020). https://doi.org/10.2139/ssrn.3709719

[134] Salomé Viljoen. 2020. Democratic data: A relational theory for data governance. *Available at SSRN 3727562* (2020). https://www.yalelawjournal.org/pdf/131.2_Viljoen_1n12myx5.pdf

[135] AG Vrana, A Sengupta, S Bouterse, J Reagle, and J Koerner. 2020. Toward a Wikipedia For and From Us All.

[136] Boxin Wang, Chejian Xu, Shuohang Wang, Zhe Gan, Yu Cheng, Jianfeng Gao, Ahmed Hassan Awadallah, and Bo Li. 2021. Adversarial glue: A multi-task benchmark for robustness evaluation of language models. *arXiv preprint arXiv:2111.02840* (2021).

[137] Jonathan Wareham, Paul B Fox, and Josep Lluís Cano Giner. 2014. Technology ecosystem governance. *Organization science* 25, 4 (2014), 1195–1215.

[138] Wikimedia Commons contributors. 2021. Mission. https://commons.wikimedia.org/wiki/Commons:Freedom_of_panorama

[139] David E Winickoff and Sebastian M Pfotenhauer. 2018. Technology governance and the innovation process. *Science OECD, editor. Technology and innovation outlook: Adapting to technological and societal disruption. Paris: OECD Publishing* (2018), 221–240.

[140] Alice Xiang. 2021. Reconciling legal and technical approaches to algorithmic bias. *Tennessee Law Review* 88, 3 (2021).

[141] Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. 2021. Bot-Adversarial Dialogue for Safe Conversational Agents. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 2950–2968. https://doi.org/10.18653/v1/2021.naacl-main.235

[142] Peter K Yu. 2018. Intellectual Property and Human Rights 2.0. *U. Rich. L. Rev.* 53 (2018), 1375.

[143] Fariborz Zelli and Harro Van Asselt. 2013. Introduction: The institutional fragmentation of global environmental governance: Causes, consequences, and responses. *Global environmental politics* 13, 3 (2013), 1–13.

[144] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, 2979–2989. https://doi.org/10.18653/v1/D17-1323

[145] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*. 19–27.

[146] Frederik J Zuiderveen Borgesius. 2020. Strengthening legal protection against discrimination by algorithms and artificial intelligence. *The International Journal of Human Rights* 24, 10 (2020), 1572–1593.

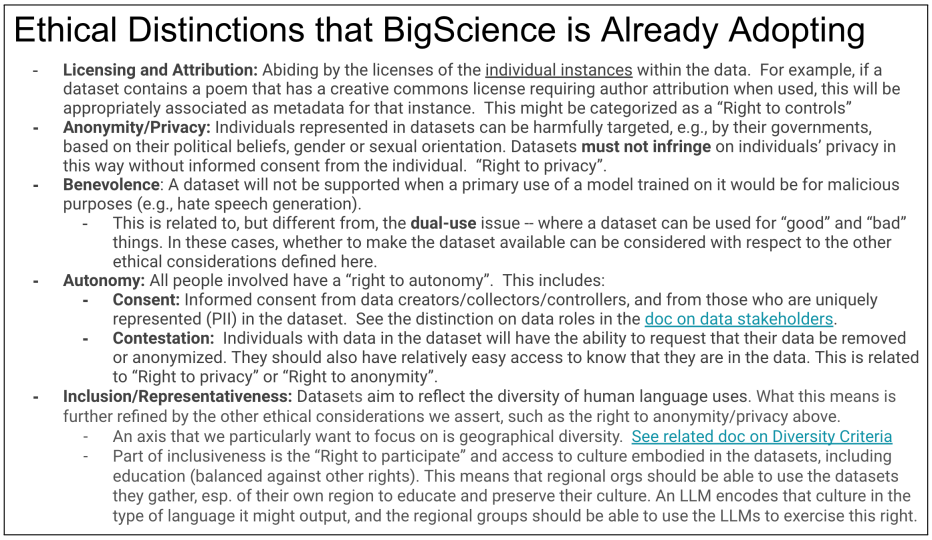


Fig. 5. Screenshot of the earliest draft of values and definitions discussed live.

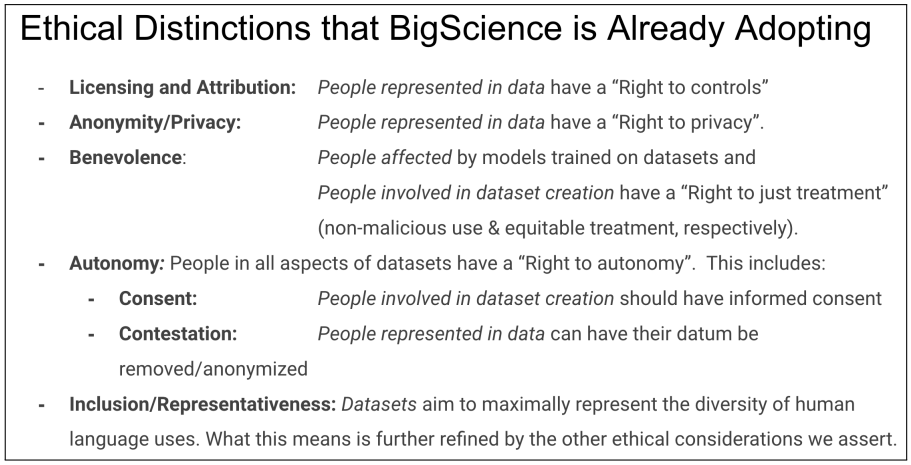


Fig. 6. Screenshot of an early draft of values and definitions discussed live.

## A Crafting Values in Data Governance

### A.1 Overview of Approach

An initial seed set of values for the data governance project were first implicitly expressed by the project planners, in our initial planning documents. These documents were a product of roughly a handful of people: The group co-leads (3 people) who primarily authored them, and a set of people who provided comments and additions after the planning documents were shared more broadly within the BigScience effort.

Discussions and debates within the larger working group (around 10 people) at our regular meetings refined and expanded on these values in light of what everyone in the group wanted to prioritize in the project. For example,

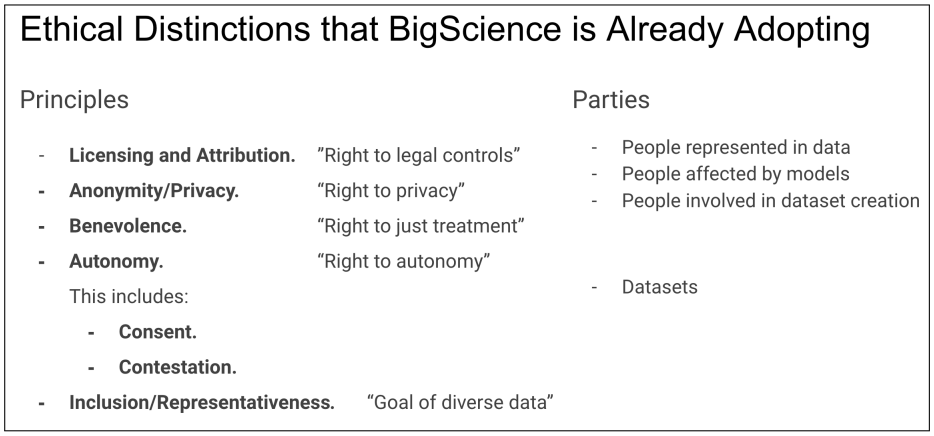


Fig. 7. Screenshot of a revised draft of values and definitions discussed live.

notes on needing to take care to make sure the data wasn't inappropriately reductive towards some populations was tentatively labelled as *inclusion*, and with the working group, this eventually evolved into a value of *representativeness*, defined as capturing the full diversity of human language use.

Similarly, the working group together decided on the best terms to use for the different value proposals. Our goal was to align on shared values to help prioritize different aspects of the work and to have some guidance to inform the decisions and potential disagreements we'd have as a working group moving forward. We recognized this was especially important as more people became involved, and so sought to have a basic set of values in place within the first 2 months of the project. Notably, prioritizing different aspects of *inclusion* was a strongly shared goal across participants.

### A.2 Steps

To create the initial set of values, we first reflected on the fact that no one would be operating as a "blank slate" in this working group; that we all had our own values, and our own goals and motivations in working on the project. As such, we focused on identifying what values we were *already bringing to the table*. This was an exercise of making the implicit explicit, and required annotating the initial planning documents alongside larger working group discussions.

First, we organized all documents and notes for the initial creation of BigScience and the working group in chronological order. Then, the co-lead went through each, highlighting specifically mentioned values – such as geographical *diversity* – as well as annotating implicit values expressed in the text by the various authors and commenters. For the latter approach, the terms used for the annotations served as placeholders for further discussion within the larger working group.

Then, the working group discussed the highlighted values and value annotations in light of their surrounding text, what the implicit ideas behind the text were, what we all felt we should be doing relevant to the value, and what we were all understanding and not understanding. Throughout these discussions, we crafted definitions live for what these value terms meant. Once the definitions were in place and generally agreed upon, we discussed the specific terms used as value labels, and in some cases changed them, or broke up definitions into different components to identify more than one value.

Screenshots representative of how these discussions evolved, in chronological order, are depicted in Figures 4, 5, 6, and 7

### A.3 Growing

Over the course of the project, the size of the working group grew. From an initial set of around 10 people, we became a group of 50+ (some more involved than others), with some individuals taking on different roles as needs arose (for example, legal scholars and others interested worked on crafting a Data Host-Provider Agreement). All participants were introduced to the grounding values and the overall plan for the governance structure as they joined; indeed, presentations on this content arguably brought more people into the group.

## B DSO Standard Host-Provider Agreement

As outlined in Section 5, one of the ways the DSO fulfills its purpose is by providing templates for licenses and leagal agreements between parties. The following license can be used to formalize a legal agreement between a **Data Host** and **Data Provider** in a way that supports our proposed governance values.

# DATA PROVIDER-HOST AGREEMENT

*v0.1*

## 1. PREAMBLE

BigScience is an open research collaboration involving over 1000 participants from 60 countries, focusing its collaborative research efforts in the study and development of natural language processing systems (hereinafter NLP).

The project is motivated by recent evolutions in the field brought about by the growing capabilities, popularity, size and cost of Large Language Model-based methods. The computational resources and data needed to develop LLMs are affordable by a handful of institutions, who often conduct this research behind closed doors despite its significant impact on society.

Thanks to the support of a large compute grant on the French Jean Zay public super-computer, the participants of BigScience can instead collaborate across a range of academic institutions and organizations to create an openly accessible Large Language Model (LLM), available for the general public. This can be used to fuel research, governance, regulation, and future technology.

In particular, the choice and governance of the Data used to develop these technologies are of paramount importance. Previous work has mainly relied on text obtained from snapshots of the Internet, due to the large amount of Data and availability. Unfortunately, this convenience choice raises multiple ethical and legal issues and leads the technology to amplify harmful biases in its deployed applications.

BigScience takes an alternative approach of identifying Data sources for a training corpus. Namely, our participants built an annotated catalog of high-quality language resources to cover the diversity of languages and social contexts that should make up such a training corpus. There are two essential parties in charge of making this data available, under the auspices of BigScience: First, the Data Providers, any institution willing to license datasets of interest purely for research purposes on a royalty free basis; and Second, the Data Host, institutions willing to contribute their technical capabilities in order to host the data provided, enabling society to access it. These are the champions of data sharing and openness in research. This License governs the use of Data as informed by the BigScience Ethical Charter. BigScience has set forth its Ethical Charter representing the values of its community. Although the BigScience community does not aim to impose its values on potential users of the Data, it is determined to take tangible steps towards protecting the community from inappropriate uses of the work being developed by BigScience.

Consequently, the main objective of this Data Provider Agreement (the Agreement) is to serve as the core instrument enabling and governing the sharing of data between the interested parties, for the benefit of open research. Both parties strive to serve this goal by entering into this Agreement.

## 2. DEFINITIONS

**“Agreement**” means this Agreement including all its Exhibits.

“**Confidential Information**” means information that one Party discloses to the other Party under this Agreement and that is marked as confidential or would normally be considered confidential.

“**Data**” means machine-readable informational content (individually or as a whole i.e., collection of Datasets) made available by the Data Provider.

**“Meta-Data”** means supplementary information of the Data, for example, summaries or visualizations of the data, restricted excerpts, authorship information and high-level statistics (i.e. word counts)).

“**Dataset**” means one specific collection of Data on which the Data Provider has the necessary rights enabling the latter the sharing of it under this agreement.

**“Processed Dataset”** is a Dataset produced via Data transformations, including additional modifications to one dataset (including PI(I) removal, additional annotations, extracted text, subsetting by language, removal of individual data points), dataset combinations, etc.

“**Data Host**” means a legal entity permitted to process, prepare, and manage subsequent 3rd party access to the Data of the Data Provider under the scope of this agreement.

“**Data Provider**” means the individual or legal entity granting permission to the Data Host to access and further manage the Data for the purpose of this Agreement.

**“Derived Work”** means any artifact created using Data covered by this Agreement.

“**Parties**” means any individual or entity entering into this Agreement.

“**Third Parties**” means individuals or legal entities that are not controlled by any of the involved parties in this Agreement.

**“User”** means individual and/or legal entity having access to the data provided by the Data Provider and hosted by the Data Host for the purpose of this Agreement.

## 3. PURPOSE, RIGHTS GRANTED & SCOPE

The Data Provider grants to the Data Host a non-exclusive, non-transferable, non-sublicensable, irrevocable, perpetual, royalty-free and worldwide license to use (that is access, store, prepare, process, label and/or share) the agreed upon Data (see List of Datasets in Exhibit A) in accordance with the use case scenarios and further (re)distribution policy, as stated in Annex III (see below).

## 4. DATA PROVIDER RIGHTS AND OBLIGATIONS

a. The Data Provider warrants that it is the owner of the Data or has the necessary rights to enter into this Agreement regarding the Data listed in the Dataset section (see Exhibit A).
b. The Data Provider will provide Data Host with valid contact information in order to settle any queries or issues related to the Data.

c. The Data Provider will provide the Data Host access to the Data in a suitable format agreed upon by both parties.
d. The Data Provider shall not be subject to any damages or liabilities for any malfunction, error or omission in the Data. In case the Data Provider becomes aware of, it will diligently inform the Data Host in order to implement the proper modifications. From its side, the Data Host will do the same.
e. In case the Data Provider is informed about the application of restrictions of any kind the Data Provider will notify the Data Host. For instance, in case of becoming knowledgeable of any actual or suspected intellectual property rights infringement, damages or claims associated with the Dataset the Data Provider promptly notifies the Data Host such that the further infringing usage of the Dataset can be stopped.
f. The Data Provider acknowledges that the Data does not contain any malicious source-code that adversely affect, alter, damage or destroy the proper functioning of any software, operating system and/or hardware this may include but is not limited to viruses, trojan horses ransomware, back doors and spy software.
g. The Data Provider shall inform the Data Host in case the Data Provider becomes aware that the dataset does not comply with relevant regulations and laws, such as personal data-related regulations.
h. The Data Provider hereby disclaims any representations and warranties of any kind, express or implied, including without limitation any warranties of fitness for the purpose set out in this Agreement or beyond regarding the Data. The Data Provider does not guarantee the accuracy, adequacy or completeness of the Data.

5. DATA HOST RIGHTS AND OBLIGATIONS
a. No rights are granted to the Data Host with respect to the Data other than those stipulated in this Agreement, except any exceptions or limitations provided by law.
b. The Data Host will hold harmless the Data Provider against any claims, demands, suits or damages arising from the use of the Data in accordance with the purpose set out in this Agreement.
c. The Data Host acknowledges and agrees that the following disclaimers apply to all End-Users and other entities who have access to the Data The Data is provided “as-is”.
d. In case of becoming knowledgeable of any actual or suspected intellectual property rights infringement, damages or claims associated with the Data the Data Host will promptly notify the Data Provider and stop the further usage of the Data until the issue in question can be resolved.
e. The Data Host will not assert rights over any Data (excluding Meta-Data) made available by way of this Agreement.
f. The Data Host will undertake commercially reasonable efforts to appropriately attribute the Data Provider as the source of the Data.
g. The Data Host is allowed to create and publish research (including benchmarks, performance indicators and/or scientific insights) gained using the data under this agreement, for the purposes of BigScience’s research scope.

h. The Data Host will undertake commercially reasonable efforts to remove personal data and information from the Dataset before using the Dataset Notwithstanding the latter undertaking, the Data Provider should, under Section 4(j) of this agreement, inform the Data Host in case the former is aware of the existence of personal data under the licensed dataset(s).

6. LIMITATIONS

a. This agreement grants access to the Data exclusively for the purpose and chosen Data Access Policy (see Exhibit A) stated in this Agreement and does not extend to any other purpose nor does it apply to any other data not listed in the Dataset section (see Exhibit A).
b. If the Data Provider complies to the data management plan the Data Provider holds the Data Host free and harmless of any action, recourse or claims made by any third party due to the non-observance by the Data Provider of its obligations under this Agreement and intellectual property and/or personal data related 3rd party claims.
c. Both Data Provider and Data Host will not be liable for any processing activities of the Data under this agreement by any User having access to it under the framework of BigScience.

7. FEES AND COSTS

Neither party will charge any fees, royalties or costs associated with implementing this Agreement. All accruing costs or expenses of any party in relation to this Agreement are solely to be carried by the responsible party alone.

8. SECURITY

The Data Provider shall make reasonable efforts to provide the Data to the Data Host using up-to-date security standards (this may include but is not limited to data transmission via secure transport protocols, storage on secured servers as well as secure data processing). In case the data is made accessible via authentication the Data Host ensures that the used authentication method meets up-to-date standards.

9. TERM AND TERMINATION

a. This Agreement is valid from the date the involved parties agree, or by default, from the moment it is signed by all the involved parties.
b. The term of this Agreement shall be from the Agreement Date until the last to expire of the Data Provider's intellectual property rights or any related rights on the licensed Dataset, strictly for the purpose of this Agreement.
c. This Agreement can be terminated by either party immediately in case the other party breaches this Agreement upon due notice of it and the breach is not remediated within 30 days.
d. In case either party would like to voluntarily terminate the Agreement, it shall act in good faith and provide the other party with (i) a reasoned statement justifying the decision; (ii) a 60 days pre-advice; (iii) and, give the other party the opportunity to negotiate any new terms and conditions for the sake of the Agreement's continuity, and beyond, for the sake of BigScience's research goals.

e. Upon termination of this Agreement for any reason, the Data Host and Data Provider cease to use the Data and Processed Datasets and for the purposes set out in this Agreement within 14 days and upon that delete the Data and Processed Datasets immediately (respecting any holding periods or processing information storage required by the law) . This does not affect already completed or created Derived Work before the termination of this Agreement.

10. FORCE MAJEURE

Neither party shall be liable to the other for a failure of performance undertaken in this Agreement if prevented from doing so by any circumstances beyond its reasonable control (such as but not limited to fire, flood, drought, war, explosion, terrorism, computer hacking and viruses, acts of any government body, perils of the sea and air).

11. CONFIDENTIALITY

Each party shall treat this Agreement and all information and/or business practices of the other party it aquires or becomes knowledgeable of as confidential. Confidential information does not include any public or generally available information or any information independently obtained or available prior to entering this Agreement. Notwithstanding the foregoing, either party is allowed to reveal confidential information if it is required by law to do so.

12. ENTIRE AGREEMENT

This Agreement including its exhibits and attachments constitute the entirety of the Agreement between the parties and supersedes any prior negotiations or understanding.

13. MODIFICATION AND AMENDMENT

This Agreement can be amended or modified by mutual consent at any time. The amendment and/or modification must be put forth in writing.

14. DISPUTE RESOLUTION

Any dispute that may arise from the breach of this Agreement will be first subject to an alternative dispute resolution phase under the auspices of the BigScience Community.

15. SURVIVAL

The provisions set forth in section 6(b) (Limits of Liability), 10 (Term and Termination), 12 (Confidentiality), 15 (Governing Law), 16 (Survival) and Exhibit A (Section Restrictions of Use in the Dataset section) shall survive the termination of this agreement and continue to bind both parties.

16. SEVERABILITY

If any provision of this Agreement is held to be invalid, illegal or unenforceable, the remaining provisions shall be unaffected thereby and remain valid as if such provision had not been set forth herein. The parties agree to substitute such a provision with a valid provision most closely resembling the intent of such severed provision.

17. NO ADDITIONAL TERMS

Unless and to the extent expressly agreed to in writing between the Data Host and the Data Provider no other terms and conditions shall be binding to either party.

18. FULL UNDERSTANDING

The parties acknowledge that they fully understand and agree to all of their rights and obligations under this Agreement.

| DATA PROVIDER | DATA HOST |
| --- | --- |
| ……. | ……. |
| Name | Name |
| ……. | ……. |
| Date and Location | Date and Location |

# Annex

**DATA PROVIDER SCHEDULE (EXHIBIT A)**

I. Data Provider Information

Data Provider Name

Data Address

Data Provider Contact Information

Data Provider Name

Special Conditions (*if applicable*)

Data Management Plan

II. Datasets

List of Datasets

| |
|---|
| |

License of Datasets (*if more than one, please assign in List of Datasets*)

| |
|---|
| |

Restrictions - Please indicate of any restrictions apply to any of the above listed datasets

## III. Field of Use

Scope / use cases:

- ☐ under condition: openly released models, results, and artifacts
- ☐ under condition: use RAIL license for ML artifacts (has to be attached)
- ☐ under condition: value alignment (determined by data host)
- ☐ under condition: value alignment (data modelers sign click-through form)

## IV. Data Distribution Policy

Acknowledging the immense value and benefits that your datasets may provide, and being conscious and respectful towards the different economic interests that you may have, this Agreement offers the Data Provider a flexible set of optional frameworks for the use, re-use, and distribution of data:

- ☐ The Data Provider permits the Data Host to use the Data for the purpose set out in this Agreement. The Data Host is **not** allowed to make the Data publicly available outside of the remits of this license (this does not include Meta Data).

- ☐ The Data Provider permits the Data Host to make the Data (as a whole or in parts or processed) available to downstream users upon signing a non-dissemination agreement.

- ☐ The Data Provider permits the Data Host to make the Data (as a whole or in parts or processed) available to downstream users using a system that supports authentication/synchronization

- ☐ The Data Provider permits the Data Host to make the Data (as a whole or in parts or processed) available with modifications such as anonymizing personal and/or sensitive information about individuals.

The Data Provider permits the Data Host to use the Data for the purpose set out in this Agreement. Additionally, the Data Host is allowed to make the Data publicly available under the Data license (select one) provided by the Data Provider.

- CC BY 4.0 (Link)
- CC BY-NC-ND 4.0 (Link)
- CC BY-NC-SA 3.0 (Link)
- CC BY-NC-SA 4.0 (Link)
- CC BY-SA 3.0 (Link)
- CC BY-SA 4.0 (Link)
- CC-BY-NC 4.0 (Link)
- Microsoft Research Data License Agreement (Link)
- custom license agreement (see Attachment if applicable)
- Linux Foundation CDLA Permissive
- Linux Foundation CDLA Restrictive