Documenting Geographically and Contextually Diverse Language Data Sources

Angelina McMillan-Major[1]*, **Francesco De Toni**[2]*, **Zaid Alyafeai**[3], **Stella Biderman**[4,5] **Kimbo Chen**[6,7], **Gérard Dupont**[8], **Hady Elsahar**[9,10], **Chris Emezue**[9,11], **Alham Fikri Aji**[12], **Suzana Ilić**[13], **Nurulaqilla Khamis**[14], **Colin Leong**[9,15], **Maraim Masoud**[7], **Aitor Soroa**[16], **Pedro Ortiz Suarez**[17], **Daniel van Strien**[18], **Zeerak Talat**[12]*, **Yacine Jernite**[18]
University of Washington[1], Australian National University[2], ARBML[3], Booz Allen Hamilton[4], EleutherAI[5], BigScience[6], Independent Researcher[7], Mavenoid[8], Masakhane[9], Meta FAIR[10], Mila[11], Mohamed Bin Zayed University of Artificial Intelligence[12], University of Innsbruck[13], Faculty of Electrical Engineering, Universiti Teknologi Malaysia[14], University of Dayton[15], University of the Basque Country[16], Common Crawl Foundation[17], Hugging Face[18]
aymm@uw.edu, francesco.detoni@anu.edu.au, z@zeerak.org

Abstract Contemporary large-scale data collection efforts have prioritized the amount of data collected to improve large language models (LLM). This quantitative approach has resulted in concerns for the rights of data subjects represented in data collections. This concern is exacerbated by a lack of documentation and analysis tools, making it difficult to interrogate these collections. Mindful of these pitfalls, we present a methodology for documentation-first, human-centered data collection. We apply this approach in an effort to train a multilingual LLM. We identify a geographically diverse set of target language groups (Arabic varieties, Basque, Chinese varieties, Catalan, English, French, Indic languages, Indonesian, Niger-Congo languages, Portuguese, Spanish, and Vietnamese, as well as programming languages) for which to collect metadata on potential data sources. We structure this effort by developing an online catalogue in English as a tool for gathering metadata through public hackathons. We present our tool and analyses of the resulting resource metadata, including distributions over languages, regions, and resource types, and discuss our lessons learned.

1 Introduction

Current trends in developing large language models (LLM) require the use of vast amounts of data (Brown et al., 2020; Gao et al., 2020; Rae et al., 2021). Typically, this data is collected from online sources, ranging from highly edited and structured text such as Wikipedia to the myriad text and audiovisual components of web pages, e.g., collected by the Common Crawl Foundation.[1] However, recent research has raised concerns about the creation and use of such data resources. For instance, Wikipedia is highly biased in terms of the topics covered and the demographics of its contributors, particularly along gender, race, and geographic lines (Barera, 2020), resulting in concerns of representation in the technologies developed on Wikipedia-derived data. Data from Common Crawl has similarly been shown to correlate with country-level population density, relative access to the internet, and per capita GDP (Dunn, 2020) and to contain significant amounts of hate speech and sexually explicit content (Luccioni and Viviano, 2021). Irrespective of the data source, typical web-crawling collection practices have no structures for supporting informed consent beyond websites' own policies that users rarely read (Cakebread, 2017; Obar and Oeldorf-Hirsch, 2020).

Several documentation schemas for natural language processing (NLP) datasets (Bender and Friedman, 2018; Gebru et al., 2021; Holland et al., 2018; Stoyanovich and Howe, 2019; McMillan-Major et al., 2023) have been proposed to aid NLP researchers in documenting their datasets (Gao et al., 2020; Biderman et al., 2022; Gehrmann et al., 2021; Wang et al., 2021) and to retrospectively document and analyze datasets that were developed and released by others without thorough documentation (Bandy and Vincent, 2021; Kreutzer et al., 2022; Birhane et al., 2021; Dodge et al., 2021). Data docu-

* Corresponding Authors: Angelina McMillan-Major, Francesco De Toni, Zeerak Talat.

[1] http://commoncrawl.org/

mentation to support transparency has gained traction, following calls for a reevaluation of the acquisition and use of data in machine learning (ML) at large (Birhane and Prabhu, 2021; Jo and Gebru, 2020; Paullada et al., 2021; Gebru et al., 2021; Bender et al., 2021). Building on this work, we propose a documentation-first and human-centered method for data collection for NLP that emphasizes consent, representation, self-determination, and privacy. Using this method, we create a data catalogue for training multilingual LLMs that promotes responsible data collection and data subjects' rights to control over their own data. We conclude that starting documentation processes during the data collection phase can contribute to building a more representative dataset and allows for early identification of ethical concerns. Our contributions consist of the data catalogue tool,[2] which remains openly available for use in collecting metadata and for searching existing entries, as well as the human-centered methodology of data collection in collaboration with language communities for representative language modeling and other NLP tasks.

1.1 Research Context

Our work was situated within a large-scale global coalition of experts in NLP and related fields dedicated to researching questions related to language modeling known as the BigScience Workshop.[3] The BigScience Workshop was started as an open collaboration of international researchers by Hugging Face, GENCI (Grand Equipement National de Calcul Intensif), and IDRIS (The Institute for Development and Resources in Intensive Scientific Computing) and was dedicated to open research of NLP, social sciences, and the legal, ethics and public policy of large language models. While this coalition (henceforth *the workshop*) had many working groups with different foci determined by the research interests of the participating researchers, one of its primary goals was to train and publicly release a multilingual LLM. Key to this endeavor was the creation of a dataset to train the model on.

Bearing in mind the limitations of prior large-scale data collection efforts, we aimed to intentionally curate our dataset for *representativeness*. We defined representativeness based on the intersection of geographic and sociolinguistic contexts. This means that, for each target language, we aimed to collect data for the relevant dialects and regions where that language is spoken. Like most language modeling endeavors, we relied on commonly used web sources for collection, but we also highlighted the need for other formats, including books, audio from radio programs and podcasts, and others. Starting from this goal and the coalition members' languages of expertise, we identified 13 language groups to target for inclusion in the model training: Arabic varieties, Basque, Chinese varieties, Catalan, English, French, Indic languages, Indonesian, Niger-Congo languages, Portuguese, Spanish, and Vietnamese, as well as programming languages. In addition to coalition members speaking many of these languages themselves, we were also motivated to intentionally select data resources for these languages in order to improve the resulting language model's performance in generating these languages. Programming languages were included in the design of the language model, but because they are not natural languages with communities of use, we did not organize a specific hackathon to collect entries for them in the catalogue (see §5).

1.2 Overview

We prepared for the challenges of responsible dataset creation by focusing our efforts on documenting potential sources prior to their collection. Meanwhile, other working groups on data governance and data tooling created pipelines for hosting and processing data. In the next sections, we compare our documentation effort (henceforth *the catalogue*) to already developed catalogs in linguistics and NLP (§2). In §3 we present our catalogue and associated online form[4], including our process for designing the catalogue.

We developed the online submission form to facilitate public hackathon events for collecting metadata for language resources from specific regions (§4). While the form prioritizes submitting entries for the target languages, we made it possible for entries for any language to be submitted as the catalogue remains open for submissions and browsing after the end of the hackathons. Although the catalogue is a living documentation effort, we present the results obtained after the initial documentation effort (§5). We then discuss lessons learned in creating the catalogue, its potential use as a model for data documentation endeavors in NLP, and the limitations of our approach, suggesting improvements for future data documentation efforts (§7). Finally, we consider the ethical implications of our approach, especially with regard to data licensing and personally identifiable information (§8).

2 Related Work

Since the early 90s, NLP data organizations have maintained catalogs for datasets and tools in order to support language research.[5] While the metadata for these

[2] Available at `https://bigscience.huggingface.co/data-catalogue-form`

[3] `https://bigscience.huggingface.co/`

[4] See Footnote 2 for URL.

[5] Organizations include the Linguistic Data Consortium (LDC), The Southern African Centre for Digital Language Resources (SADiLaR),

catalogs are openly available, accessing the language resources (e.g., annotated corpora and lexicons) and supporting tools may require paying for a license to the resource or for membership to the catalog. The fees support the creation, licensing, storage, and maintenance of new datasets and language research initiatives. The LDC, for example, currently provides access to 1016 datasets.[6]

Open source dataset catalogs have also been constructed as supporting technical infrastructure in the context of NLP and ML libraries. The Natural Language Toolkit (NLTK), developed since 2001, is a Python package with utilities for NLP tasks that includes access to widely used corpora such as the Brown Corpus (Kučera and Francis, 1967) as well as features for adding datasets and using datasets locally (Bird et al., 2009). The Hugging Face Datasets library (Lhoest et al., 2021) and Tensorflow library (Tensor Flow Authors, 2021) both provide tools for loading datasets from remote and local repositories and include catalogs of directly accessible datasets. SADiLaR provides its own catalog of annotated language datasets and processing tools, with links for downloading resources that are licensed for distribution. Other catalogs of NLP datasets do not provide access to the datasets themselves, but provide information about uses and categories. For example, Papers with Code links academic publications that use the same dataset with information about the dataset.[7] Masader similarly provides metadata about Arabic-language NLP datasets without hosting the data (Alyafeai et al., 2022).

Our work is an effort to merge the careful and well-established data collection and documentation practices from organizations such as the LDC with the collaborative, open source tools for dataset construction. While large-scale NLP research requires vasts amounts of data, the work that goes into curating, documenting, and maintaining the data is often undervalued (Sambasivan et al., 2021), resulting in data collections that are often too large to document post-hoc (Bender et al., 2021) and contain significant quantities of unwanted media (Luccioni and Viviano, 2021). We provide an alternate approach to data collection and management in NLP; this approach prioritises documentation in the data creation process, engages communities to inform data curation, and contributes to a more representative dataset.

3 The Catalogue

The primary goal of the catalogue (see appendix A for screenshots of the form) was to support the creation of a training dataset for language modeling that integrated with the efforts of the other working groups and aligned with the values defined by the workshop governance. We surveyed each working group to identify their particular metadata needs, resulting in almost 40 categories of metadata. Aiming to balance the information needs of the working groups with the effort required to submit a resource and its metadata to the catalogue, we grouped and prioritized the categories. We further prioritized metadata that are applicable across as many languages and data sources as possible. We did not make use of existing metadata formalisms as we expected that they would discourage submissions to the catalogue by those unfamiliar with them. Instead we envisioned our metadata collection as an upstream process that would be flexible enough to contribute to many different kinds of downstream annotation or metadata labeling tasks.

We created an openly accessible form in English for submitting metadata for potential sources for the identified language groups.[8] We used an iterative approach to collectively develop questions that elicit the metadata, descriptions of the information being requested, and answer prompts to support efficient documenting. Wherever possible, we formatted the questions as multiple choice questions with an optional free-form field, should the pre-existing options be insufficient. After building the online form, we tested the form with actual examples, i.e., the *Le Monde* newspaper and its publishing company *Group Le Monde* to ensure its validity.

3.1 The Catalogue Submission Form

Testing the form using the *Le Monde* newspaper example helped us update our form by surfacing discrepancies in specific questions for certain resource types, particularly concerning data processing. With this consideration in mind, we defined the following resource types: **primary source**, a single source of language data (text or speech), such as a newspaper, radio, website, or book collection; **processed language dataset**, a processed NLP dataset containing language data that can be used for language modeling; and **language organization or advocate,** an organization or person holding or managing language sources of various types, formats, and languages. We follow Jernite et al. (2022) in distinguishing between **data subjects** (those talked to or about in the data), **data creators** (those who create the text, audio, or video data), and **data custodians** (those who own or manage the data). We distinguish between a data custodian, who is responsible for handling requests for the data, and language organizations, that may ultimately hold the rights to the data but do not handle day-to-day requests, though in many cases the data custodian and the language organization of a resource are the same entity.

the European Language Resource Association (ELRA), the Chinese LDC, the LDC for Indian languages (LDCIL), and CLARIN.

[6] LCD Catalog by Year, accessed April 18, 2023.

[7] `https://paperswithcode.com/datasets`

[8] We built the form using Streamlit.

For all resource types, the form requests information about the languages and locations of the resource's data creators as well as contact information for a representative, owner, or custodian of the resource. Further questions are added for primary sources and processed datasets, including the availability of the resource and legal considerations for using the data, such as licenses, the type of data it contains, and the medium of the data.

3.1.1 General Information

The form first requests the source type, and then updates the questions once a type is selected. The following questions in the section request general information (e.g., a resource name, a unique identifier for searchability, and the resource's webpage). The form provides space for a description to display when searching the catalogue.

3.1.2 Languages and Locations

We designed the *Languages and Locations* section to accommodate various degrees of granularity in order to support and evaluate our goal of representativeness, and maximize the usability of the catalogue beyond the consortium's immediate use-case. The authors of each entry can specify what languages are represented in the resource by choosing from drop-down lists of our target language groups, with additional sub-lists for languages in the Indic and Niger-Congo families, and other languages as defined by the BCP-47 standard (Phillips and Davis, 2009). The form also provides space for submitting comments about the language variety in the resource, such as whether it contains language data that exhibits dialectal variation or code-switching. Similarly, authors can add information about the geographical origin of the data (i.e., the primary location of the language creators whose data is captured in the resource) using a drop-down list of macroareas ranging from world-wide to continents to regions (such as Western Africa or Polynesia) in addition to specific countries, nations, regions, and territories.

3.1.3 Representative, Owner, or Custodian

Responsible dataset creation includes respecting the rights of the data custodian, the person or organization that owns or manages the data source. The form allows for linking the resource being submitted to an existing organization in the catalogue via a drop-down list. If the data custodian is not already in the catalogue as a language organization, the remaining questions elicit their name, type, location, and contact information. This information supports our own and future catalogue users' efforts to understand local legal structures, communicate with data custodians about data use, and request permission for uses beyond those granted by licenses.

3.1.4 Availability of the Resource

For primary sources and existing datasets, the form requests information about how to obtain the data, i.e., through a link or contacting the data custodian. Depending on the response, the form asks for the URL to download the data or the data custodian's contact information. In characterizing the licenses or terms of use, the form asks whether the resource is accompanied by an explicit license. If the license or terms are known, the submitter may select a description such as public domain, research use, non-commercial use, or do not distribute. Submitters can also select relevant licenses from a drop-down list of frequently used licenses, or input the terms or license text into the form. If the licensing terms are unknown or unclear, the form requests that the submitter gives their best assessment of whether the data can be used to train models while respecting the rights and wishes of the data subjects, creators, and custodians.

3.1.5 Primary Source Type

The form allows for characterizations of the resource data for both primary sources and processed language datasets. We provide options for two kinds of resource descriptions. **Collections** may contain books or publishers, scientific articles and journals, news articles, radio programs, movies and documentaries, podcasts, or a user-suggested response. **Websites** may include social media, forums, news or magazine websites, wikis, blogs, content repositories, or a user-suggested response.

If the submission is a processed language dataset, the section appears in the form as *Primary Sources of the Processed Dataset*. If the dataset contains original data, no further questions appear. If the data is a collection of primary sources, the form presents questions about those sources, such as if they are openly available or have accessible documentation. Users may link the processed dataset to primary sources already documented in the catalogue or provide original descriptions of those primary sources. The final question concerns the licensing information of the primary sources, as these may differ from the dataset itself. See §8.1 for further discussion.

3.1.6 Media Type, Format, Size, and Processing

The final section of the form addresses the technical aspects of the resource. A submitter may indicate the medium of the data (text, audiovisual data, images, or a combination thereof) and details about the data format (the file type or distribution format). If the data includes text, the form asks if the text was transcribed. While most datasets appear with metadata about the size of the data given by mega- or gigabytes, primary sources often do not have this information available. Instead, we asked submitters to provide an estimate of the amount

of data using a descriptive unit of data, i.e., articles, posts, episodes, books, webpages, or a user-provided unit. The form then asks for the number of instances in the resource using the provided unit and the average number of words in the unit using ranges of magnitudes of 10. This information was useful to the coalition's data processing working groups, but it proved difficult for the submitters to estimate, unless already available in the source metadata. On completion, submitters could review their responses as it will be saved (in a JSON format) before submitting their entry to the catalogue.

4 Additional Features

There are two other modes for interacting with the catalogue: a validation mode, for validating submitted entries, and a visualization mode, for filtering and mapping specific submitted entries. Because we intended to make the catalogue openly available on the web past the end date of the workshop, we included the validation functionality to allow users to confirm that metadata for submitted entries was correct and could be updated if ever the information was no longer correct (e.g., if a license for an entry changed). The purpose of the visualization mode was to support later users of the catalogue in seeing the general distribution of submitted resources of the catalogue across languages and geographic regions and searching for specific resources within those categories.

To validate an entry, the validator can confirm the previously submitted metadata or edit and resubmit the entry. The catalogue then saves both the original and the validated submission. The visualizations include a pie chart detailing the proportion of entries by language and an interactive map which shows the number of submitted entries for a region or country as defined by the location of the data creators or data custodians. In Figure 1, the color gradient indicates the number of entries by country and location markers indicate regions that can be examined for more details. Both the map and a pie chart can be filtered using one of the many properties produced by the form, e.g., the resource, license, or media type. Entries returned by the filter can be selected to display their descriptions.

5 Community Hackathons

With the catalogue submission form developed, we could begin to collect and document potential data sources for review prior to developing the full dataset towards the workshop's LLM goal. Whereas prior data collection processes utilized automatic methods for collecting as much data as possible, we wanted our collection process to prioritize sources that were created by language communities and that were determined by language communities to be representative of their language use. In order to center the metadata collection for as many languages as possible around communities who speak those languages, we decided to crowdsource our metadata collection by organizing community hackathons.[9] To do so, we reached out to regional community organizations focused on ML and NLP to collaborate in leading local hackathons and put out a similar call within the workshop for individuals who spoke one or more of the listed languages. The task for each hackathon was for participants to use the catalogue submission form to submit as much metadata as they could find on potential data sources for their language or languages. We developed a guide[10] with instructions and suggestions for the hackathon participants for each section of the catalogue submission form. A coalition member and/or a collaborating organizer from a partner organization was available to interact with participants and answer questions arising while filling the form and to discuss details about potential resources or institutions.

In total, we organized 6 hackathons for specific communities and regions of the world based on the availability of organizers and their familiarity with the communities, namely African languages in collaboration with Masakhane,[11] Asian languages with Machine Learning Tokyo,[12] Basque, English in the Americas and Europe, English in the Indo-Pacific Region, and Spanish in Latin America with LatinX in AI.[13] The hackathons took place online in October-December, 2021, lasting one to six hours. We announced hackathons using social media, in coordination with the relevant partner organizations. Because we advertised primarily to members of the workshop, social media followers of the workshop, and members of the partner organizations, the hackathons attracted participants who were generally interested in language modeling and specifically wanted to support the workshop goals of having greater language representation in the to-be-trained workshop language model. No further incentives were used to encourage participation. During the hackathons we only collected a name and e-mail. After the hackathons, we sent a 10-question survey to all participants to collect further information.

[9] Because programming languages are not natural languages with communities of speakers or signers, we did not organize a hackathon focused on programmming languages.

[10] Available at `https://github.com/bigscience-workshop/data_sourcing/blob/master/sourcing_sprint/guide.md`.

[11] `https://www.masakhane.io/`

[12] `https://www.mlt.ai/`

[13] `https://www.latinxinai.org/`

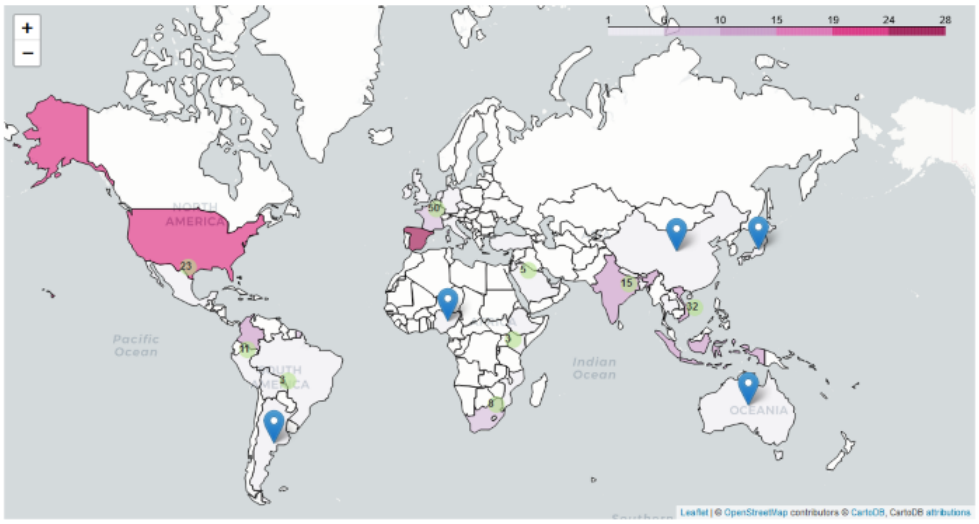

Figure 1: Geographical visualization of the locations of entries' data custodians. The color gradient indicates the number of entries by country and location markers indicate regions that can be examined for more entries and details.

6 Results

6.1 Hackathon Participation

Forty-one participants submitted descriptions of resources to the catalogue during the hackathons, of whom 11 responded to the survey. The first survey questions focused on participants' professional context, i.e., the country they are located in, their field of study and current stage in their career. The respondents were from diverse geographical location and career stages. Four respondents were located in Spain, with 3 in the Basque Country, while the remaining respondents were located in France, Japan, Kenya, Singapore, Sweden, Taiwan, and the USA. Respondents' career stages ranged from undergraduate student to a senior level position in industry, though most (7) listed an academic position. The most common research interests were NLP (8), data science (5), and linguistics (4). Other interests included library and/or information science, ethics/safety, recommendation systems, vision, creative AI, and optimization and compression techniques.

The remaining questions concerned participants' experiences before and during the hackathons. Most participants became aware of the hackathons through the coalition's internal channels or the communities and organizations that collaborate with us. Only two respondents listed social media as their entry point. Most respondents (6) only submitted resources for languages that they were fluent or advanced speakers of, while three respondents contributed resources that covered almost all of the target languages, most of which they had no familiarity with. In describing their motivations for participating in the hackathons, the most common reasons included developing the training dataset, supporting under-resourced languages in general, and improving the coverage of a particular language.

6.2 Gathered Resources

After the sixth and final hackathon, the catalogue contained 192 entries with 955 different language tags.[14] The most frequent language tags were those of the target language groups. Figure 2 shows the distribution of the target language groups across entries.[15] English is the most frequent language across all entries. For Arabic, the most frequent varieties are Modern Standard Arabic (13) and Classical Arabic (5). All other variants have 2 or fewer entries. The most frequent Indic languages are Hindi (15), Bengali (11), Telugu (9), Tamil (9), and Urdu (8) and the most frequent Niger-Congo languages are Swahili (9), Igbo (7), Yoruba (6), and isiZulu (4), with other languages having no more than 3 entries.

On the other end of the spectrum, 380 languages were tagged only in 1 or 2 entries. However, some of these languages belong to broader target language groups: i.e., 10 languages from the Niger-Congo group (Sesotho, Kirundi, Bambara, Kinyarwanda, Chi Chewa, Wolof, Twi, Lingala, ChiShona, and Kikuyu), and 12 varieties of Arabic (Algeria, Djibouti, Gulf, Egypt, Levant, Libya, Mauritania, Morocco, North Africa, Somalia, South Sudan, Sudan). Digitally accessible resources for these language varieties are less common than digital resources for languages with more frequent use on the internet, in part due to the smaller sizes of the com-

[14] The list of language tags includes both Arabic (generic tag) and specific varieties of Arabic (e.g. Classical Arabic). The form remains open and new entries have been added since the final hackathon. At present, the there are 252 entries in the catalogue.

[15] Due to multilingual resources, the percentages exceed 100%.

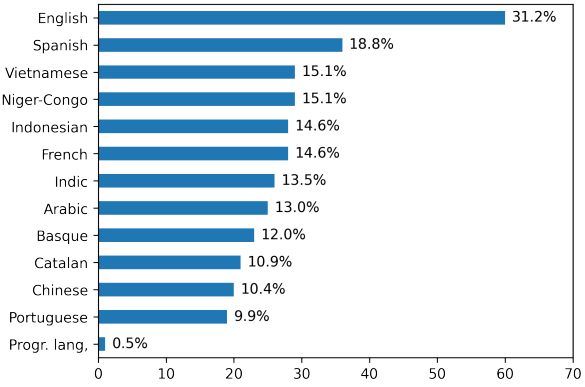

Figure 2: Relative distribution of the target languages in absolute values and as percentages of the total number of entries.

Language location	#	Percentage of all entries
Africa	18	9.38%
Americas*	3	1.56%
Asia	61	31.77%
Europe	46	23.96%
Latin America and the Caribbean	17	8.85%
Middle East and North Africa	4	2.08%
North Africa	2	1.04%
North America	11	5.73%
Oceania	5	2.60%
World-wide	21	10.94%

* entries not specifying if N. Am. or Lat. Am. and the Car.

Table 1: Distribution of language locations according to data creators (not custodians) over geographic regions (only first location for each entry).

munities using these languages and in part due to the numerous sociopolitical factors that have led to the valuation and resource allocation towards some languages (usually associated with colonial powers) over others. Excluding these, 358 languages were tagged only once or twice.

The submissions to the catalogue show a clear bias towards certain languages: English and Spanish submissions accounted for about half of the target languages recorded by the end of the hackathons. On the other hand, Chinese is included in fewer entries than languages that have fewer speakers, e.g., French, Spanish and Vietnamese (see Eberhard and Fennig 2021). This imbalance is the result of the varying availability of sources across different languages and the linguistic expertise of the coalition and hackathon participants.

We did not require users to adhere to a strict taxonomy of geographic location (e.g., continent → country → region) when providing geographic locations of a source. The submitters could freely label their submissions by macroscopic area (e.g., a continent or macroregion within a continent), country, region within a country or some combination of these. These labels are then saved in a list of location tags for each entry. We made this design decision to simplify the process of selecting geographic location for submitters while avoiding nested questions with increasing geographic granularity, providing flexibility in geographic labelling. For example, it may make more sense to label resources in Arabic as from *Middle East and Northern Africa*, rather than from *Africa* and *Asia*, even though *Middle East and Northern Africa* does not denote a continent in geographic terms. As a result, the catalogue does not conform to a particular taxonomy but can provide a frequency distribution over the location tags.

Location	Languages			
	En.	Fr.	Sp.	Port.
Africa*	6	4	0	1
Americas†	0	1	2	1
Asia	10	0	0	1
Europe	13	13	11	5
Latin America and the Carib.	3	0	15	2
North America	13	1	2	1
Oceania	5	0	0	0
World-wide	16	11	10	11

* including entries from North Africa; no entries from Middle East were recorded for these languages
† entries not specifying if N. Am. or Lat. Am. and the Car.

Table 2: Distribution of entries in English, French, Spanish and Portuguese across continents.

We focus our analysis of the geographic distribution of the recorded languages on continents and macroregions (i.e., usually the first geographic area provided). For the small number of cases where only a country was provided, we manually assigned the information to their respective continent or macroregion. We see in Table 1 that more than half of the primary language locations of the entries are located in Asia and Europe.

We further manually grouped locations into continents and macroregions and investigated how regional varieties of English, French, Spanish and Portuguese entries are represented (see Table 2). We see that these languages are well represented in their European varieties. However, each language also has a number of entries from other geographical areas, which are language specific, and several entries that were tagged as 'Worldwide' (entries that include examples of a target language from multiple geographies or multilingual sources).

Primary sources were the most common source type entered. Of the 192 entries, 98 (51%) are primary sources, 64 (33%) are processed datasets, and 30 (16%) are organizations (see Table 3 for distributions of source types across target language groups). With the ex-

ception of Catalan, Indic and Vietnamese, the target language groups have more primary sources than secondary sources.

Languages	Types		
	Primary	Processed	Org.
Arabic	13	3	9
Basque	15	0	8
Catalan	1	14	6
Chinese	9	4	7
English	29	13	18
French	13	4	11
Indic	8	11	7
Indonesian	15	8	5
Niger-Congo	11	5	13
Portuguese	7	3	9
Programming	1	0	0
Spanish	17	2	17
Vietnamese	8	15	6

Table 3: Distribution of the target languages in the catalogue across source types.

The largest share of sources recorded are stewarded by non-commercial entities (see Table 4). University and research institutions are the most frequent custodian type (23.44%), followed by commercial entities (21.35%) and nonprofit entities/NGOs (13.5%). Twenty-four (12.5%) records do not specify a custodian.

Custodian type	#
University or research institution	45
Commercial entity	41
Nonprofit / NGO	26
Not Specified	24
Private individual	20
Government organization	17
Library, museum or archival institute	16
Community (incl. online)	2
Startup	1

Table 4: Distribution of custodian types.

In terms of the custodians' geographic diversity, 28 catalouge entries do not record a custodian location while the remaining 164 do. While the custodian locations reflect the diversity of the catalogue, they also show that an outsized share are located in the USA and European countries (see Table 5). All the other locations were only recorded once (Bolivia, Burundi, Czech Republic, Ethiopia, Hong Kong, Ireland, Italy, Kenya, Luxembourg, Mexico, Netherlands, Peru, Saudi Arabia, Scotland, Thailand, Turkey, United Arab Emirates) or twice (Argentina, Bangladesh, Brazil, Japan, Jordan, Mozambique, Nepal, Nigeria, Taiwan).

The license metadata suggests that the hackathon

Custodian location	#	Custodian Location	#
Spain	27	France	9
USA	22	South Africa	6
Vietnam	14	UK	5
Indonesia	14	Australia	4
India	11	Germany	4
Colombia	10	China	3

Table 5: Top 12 most frequent custodian locations.

participants made efforts to submit sources with open licenses or without copyright (see Table 6).[16] Public domain or open license account for 37% of entries and another 37% are entered as not having licenses.

Licensing properties	#	Percentage of all entries
Missing	71	37%
Open license	56	29%
Copyright	30	16%
Non-commercial use	18	9%
Public domain	18	9%
Research use	10	5%
Multiple licenses	7	4%
Do not distribute	2	1%

Table 6: Distribution of licensing properties.

The hackathon submission entries contained primarily text data, as shown in Table 7. Two thirds of the resources contain only text data, while 5% contained text and image data, 4% contained text and audiovisual data, and another 6% contained text, image, and audiovisual data combined. Only 3% of the resources contained solely audiovisual data. A further 16% of the resources are missing information about the media types contained within the resource. This may suggest that the resources were not accessible or did not provide sufficient documentation for the hackathon participants to determine the resource media types.

Media type	#	Percentage of all entries
Text only	128	66%
Text and image	9	5%
Text and audiovisual	7	4%
Text, image and audiovisual	11	6%
Audiovisual	5	3%
Missing	32	16%

Table 7: Distribution of media types.

After the hackathons, the resources within the catalogue data were downloaded and used to develop the

[16] Entries may have multiple license properties.

BigScience ROOTS Corpus. Details on the data processing methods for the dataset and the resulting data metrics may be found in Laurençon et al. (2022). While the resources from the catalogue were ultimately used in collaboration with other data sources such as the OSCAR version 21.09 corpus (Ortiz Suárez et al., 2019) in order to meet the data quota for training an LLM, the catalogue could continue to grow to provide more metadata on resources used for NLP tasks and support documentation efforts for future data collection projects.

7 Discussion

The result of our efforts is an openly available catalogue of 192 data sources, with each of our target natural language groups constituting at least 10% of the total submitted entries.[17] The majority of these resources are primary and processed resources, with data custodian primarily located in the Americas, Europe and Australia. The sources recorded in the catalogue were used as a core component of the training dataset for training the LLM developed by the coalition. In addition, the development of the catalogue serves as an opportunity for methodological reflection on documentation-first and human-centered data collection in NLP. In this section we discuss lessons learned from creating and crowdsourcing the catalogue and present recommendations for future data collection efforts.

7.1 Centering the Human

A human-centered approach to data is one that is focused on "human values such as privacy, human rights, and ethics", is engaged in "asking ... what [technology] should do", and is committed to "acknowledging and addressing the individuals, organizations, and communities behind ... data" (Shah et al., 2021, p. 794). Our collection methodology consists of engaging with language communities to prioritize the collection of resources that those communities deem are representative of their language, as opposed to automatic collection and language identification methods. Additionally, the form dedicates multiple sections to the individuals and groups that produced and hold the data while the hackathons made the data curation process more accessible to members of the coalition not working in the data-sourcing working group. Our methodology also centers humans by collecting information on the rights of data holders and owners (Jernite et al., 2022) prior to collecting the actual data. This affords making informed decisions with respect to privacy and ethical considerations as well as data curation choices for content. However, the methodology has several limitations. First, it only addresses the needs of immediate users of the catalogue. Serving less immediate data consumers and connecting them with data producers would require additional infrastructure. Second, the methodology does not protect the rights of data holders and owners from future malicious users of the catalogue, as it does not embed a data governance structure within it. We discuss these risks further in Section 8.

7.2 Representativeness

Representativeness across languages Our catalogue only covers a fraction of world languages, largely reflecting the languages and contexts of the coalition members; missing are signed languages, some of the most widely spoken languages, and most underrepresented languages. Moreover, the distribution of target language entries in the catalogue is not uniform. While the efforts of the hackathon resulted in diverse resources for the languages covered, especially for English, French, Spanish, and Portuguese, the variation in success across languages emphasizes the need to actively include collaborators who sign and speak underrepresented languages and supporting them in leadership positions in NLP research.

As our results evidence, our definition for success, namely broad geographical representation, had direct impacts on our ability to evaluate the catalogue. For instance, the loosely structured ontology for recording dialects and geographical locations on one hand provided users with flexibility to adapt data entry to each source. On the other, it becomes more difficult to analyze information across the catalogue. Aggregating dialects and geographical locations of data posed a challenge because sources may include examples from multiple dialects and/or regions, resulting in significant difficulties in creating a classification protocol that was applicable to all sources. Furthermore, information about the geographical location of the languages in the sources may not be easily accessible or available to submitters.

Representativeness beyond language diversity Our effort focused on ensuring geographic variation and representativeness of target languages. However, from the perspective of linguistics, representativness encompasses a broad set of variables. For example, Biber (1993) identifies 8 hierarchical parameters that define the representativeness of a corpus: the primary channel (written, spoken, or scripted speech);[18] the format (published or not published); the setting (institutional, other public, or private); the addressee (plurality, pres-

[17] These numbers were calculated at the conclusion of the hackathons. The catalogue shows 252 entries at the time of publication.

[18] We note that this framework also fails to consider signed languages.

ence, interactiveness, and shared knowledge); the addressor (demographic variation and acknowledgement); factuality; purposes; and topics. While some catalogue submissions include speech data, e.g., the Global Voices dataset,[19] the majority of the entries are written texts from the internet and book archives. Language from private settings, e.g., medical consultations, is therefore not present in the catalogue. The content of the sources mostly have asynchronous and unspecified addressees (i.e., the addressees are readers of physical and digital written sources as opposed to participants in a face-to-face dialogue), and a variable degree of interactiveness (more for social media, less for books), and shared knowledge. The catalogue captures neither factuality of the sources (e.g., as determined heuristically by classifying the genre as fiction or non-fiction or by analyzing the content against a knowledge base), their purposes or topics, nor demographic variables. While an analysis of demographic variation is beyond the scope of this paper, we assume that the catalogue does not proportionally represent demographic categories such as age and socioeconomic status, as internet participation is skewed towards certain demographics (Ranchordás, 2022).

7.3 Challenges in creating the catalogue

Some of the limitations of the catalogue are consequences of the challenges faced in crowdsourcing. The origin of these challenges was the need for a crowdsourced data collection process that met the goals of human-centeredness and representativeness. In our analysis, we focus on three items at the intersection of crowd participation and catalogue design: recruiting volunteers, creating entries, and tracking them.

Recruiting volunteers The motivations of volunteer participants in projects like the catalogue have previously been explored in citizen science and crowdsourcing research across disciplines, e.g., astronomy (Raddick et al., 2013), biology (Berger-Wolf et al., 2017) and history (Causer and Terras, 2014). Such studies often find that a small number of volunteers contribute the majority of data, while a large number of volunteers only contribute once (Segal et al., 2016).[20] These observations emphasize the importance of a large number of contributors for our hackathons. However, given the scope of our project, the 41 participants fell short of our goal. Despite public advertising, our participant survey suggests that the majority of participants were partner organizations or members of our coalition. To address this issue, future hackathons can perform more outreach through partner organisations, and sustain a long period of promoting the events. The actual and perceived difficulty in contributing may have further hampered participation. Additionally, motivations to volunteer for data-related work may have suffered given the broader under-valuing of such work in NLP and ML (Sambasivan et al., 2021).

Creating catalogue entries In the participant survey, we asked respondents to detail challenges in contributing to the catalogue. Participants noted difficulties with finding appropriate resources, specific metadata, and catalogue infrastructure. The appropriateness concern grew from the potential for conflict around the use of data for training ML models. When respondents submitted a resource, they further detailed difficulties in describing certain metadata. For instance, primary sources often lack licensing metadata (see §5.2). Other difficult-to-obtain metadata include information about the data custodian; amount, type and format of data; and curation rationales. Libraries and archives face similar challenges and creating metadata to describe collections is one of their core missions. However, Padilla et al. (2019) found a gap between the detail of metadata at the item and collection level, suggesting that addressing this challenge may require new infrastructure. Respondents also requested features for the catalogue's technology, e.g., fuzzy-search and visualization (detailing relations between sources). For future hackathons, respondents suggested language-specific communication channels for sharing resources and information, more accessible times for the events, and support for uploading CSV files.

Tracking entries At this stage of the catalogue, the infrastructure for verifying information and moderation of submitters is underdeveloped. There is currently a system in place which notifies a submitter when their submission has been verified by another user. Future development of the catalogue could restructure the submission system to allow subscription to updates to submissions, or make edit histories available with associated functionality for explaining and discussing changes. The inclusion of discussion functionality however would also require an active moderation team to ensure that discussions are respectful and relevant to the catalogue.

7.4 Recommendations

Based on our experiences, we provide recommendations for future efforts on designing tasks with community participation that engage a broader data ecosystem, and uses catalogues for language sources in NLP. Completing an entry for the catalogue proved to be a complex task, as it requires domain knowledge of potential sources (or how to identify them) and understanding how to identify the necessary metadata. Future efforts can make submitting to the catalogue more inclusive by

[19] `https://globalvoices.org/`

[20] A similar pattern arose in EleutherAI's Evaluation Harness (Gao et al., 2021) and Google's Big Bench (Srivastava et al., 2023).

breaking down tasks for creating and reviewing entries into subtasks. Future efforts may also recruit volunteers for recording and correcting metadata about language variety or licenses, where these are inconsistent or missing in the catalogue. Crowdsourcing-task designers in the cultural heritage sector propose defining differentiated roles, e.g., submitters and reviewers, to streamline voluteer efforts (Ridge et al., 2021).

We also recommend that future efforts establish collaborations with data custodians that have existing processes for describing and curating data, e.g., libraries and archives, as these can ease the burden of access to (meta)data while supporting the development of standards for metadata and ethical best practices (Jo and Gebru, 2020). Although selecting and implementing a standard is a political process with many stakeholders, it can afford a machine-readable schema providing ease of aggregation across records. One such example, DataCite[21] provides a core metadata schema that has been adopted across many data and software repositories.[22]

Finally, crowdsourced catalogues of language data may also find use in education settings, e.g., courses on data selection and management for NLP.[23] In our efforts to build the catalogue, we relied on volunteer researcher hours, however within a classroom setting, students could search for entries, submit and review metadata as a part of classroom exercises. Such an exercise could provide students with experiences of the challenges and ethical considerations of language data curation.

8 Ethical Considerations

Beyond the limitations outlined in §7, future users of the catalogue and the data it references should be aware of a number of ethical considerations relevant to it. Whilst the catalogue is open, the data it registers have their own licenses and usage restrictions that users must abide by (e.g. licenses that preclude commercial uses of data). For instance, appropriately handling personally identifiable information (PII) must be included in plans for the catalogue, with attention to the detection and implications of different types of PII. In the following sections we reflect on these topics, based on lessons learned during the catalogue development.

[21] `https://schema.datacite.org/`

[22] While it is possible to convert between the majority of Datacite's schema and the catalogue, the catalogue lacks some fields (e.g., PublicationYear) required by DataCite. The requirement of a fixed publication date presents a challenge for living data sources, which we sought to include. A possible solution can be to clarify the dataType field for different resources, to allow for collecting this information at different granularity. For example, the 'Collected' dataType allows specifying the "date or date range" for a resource (DataCite Metadata Working Group, 2021).

[23] Thank you to Emily M. Bender for suggesting this additional use case.

8.1 Licensing

Instances of automatically collected data from the internet have been shown to disregard licenses and copyright terms defined by the original data owners (Bandy and Vincent, 2021). Currently, the submission form includes a section that requests the licensing terms for the primary data source of an entry and whether the submission respects the terms of the primary source. The catalogue also accepts and makes visible submissions that do not adhere to the licensing terms of their primary data source. This limitation in the catalogue design may have undesired consequences of facilitating access to resources that violate licensing terms. Future catalogues may allow the submitter to view the entry, but hide it from others. If the resource were to remedy the licensing issues, the submitter could then update the catalogue entry and make it globally visible. A data governance structure, e.g., the one proposed by Jernite et al. (2022), would be necessary for the removal of entries when they are mislabeled as respecting licensing terms but in fact violate them.

8.2 Personally Identifiable Information

The first version of the form requested that submitters specify the kinds of PII contained by an entry, if any; however, because a third of the entries indicated that the amount and type of PII was unknown or was left blank, we decided to move forward under the assumption that all data sources have some kind of PII and that properly addressing PII documentation and identification would be better handled by a targeted investigation. We initially included this metadata so that it could act as a foundation for privacy-preserving data processes and support data subjects' right to be forgotten. On the basis of the US Health Insurance Portability and Accountability Act of 1996 and the EU General Data Protection Regulation,[24] we define three categories of PII:[25] **General PII** includes information such as names, physical and email addresses, website accounts with names or handles, dates (birth, death, etc.), full-face photographs and comparable images, and biometric identifiers (fingerprints, voice, etc.). **Numeric PII** includes identifying numbers, e.g., contact information, vehicle and device identifiers, serial numbers, IP addresses, medical or health plan numbers, and any other uniquely identifying numbers. **Sensitive PII** includes descriptions of racial or ethnic origin, political opinions, religious or philosophical beliefs, trade-union membership, genetic data, health-related data, and data concerning a person's sex life or sexual orientation. We asked submitters

[24] HIPAA and GDPR

[25] While not all data sources in the catalogue are under the jurisdiction of these regulations, they provide a starting point for examples of information that may lead to the identification of an individual.

to determine whether data sources were likely to contain any of the PII described above on a scale from very likely to none.

If an entry had a likelihood of containing PII, the submitter was asked to select the kinds of information that might occur from the examples above. We advised submitters to assume that entries contained PII unless there was a good cause to believe otherwise, in which case we asked the submitter to justify their belief. Considering common sources, we predicted two likely justifications for the absence of PII: the data was fictional or general knowledge not written by or referring to private persons. These options appeared as prepopulated answers, but the submitter could also provide their own.

Contains PII	#	Percentage of all entries
Yes	84	44%
Unclear	48	18%
Answer Missing	30	16%
No	25	13%
Yes (text author's name only)	18	9%

Table 8: Distribution of entries with PII or sensitive information.

Our analysis of PII metadata showed that more than half of the catalogue contained PII (see Table 8). Another 34% of the catalogue had unclear information or missing metadata about PII, and only 13% of the catalogue had no PII (according to the the catalogue entries). With just 13% of entries clearly indicating no PII, we removed PII as a category in the form, assuming that each entry should be considered to contain PII when preprocessing the training dataset. This decision represents a conservative approach; it also highlights a practical limitation to data sourcing efforts with regard to PII. Jernite et al. (2022) propose data sourcing, governance, and tooling as the three components of distributed and people-centric handling of PII. Data sourcing decides what data to prioritize based on identified privacy risks and impacts on stakeholders. However, as our catalogue shows, crowdsourcing informative metadata about PII presents challenges when submitters are unable to accurately estimate the presence of PII in the sources. As a result, decision-making about sources and PII is relegated to the data tooling stage, where PII are filtered from the data. This indicates a need for new models of data sourcing that can optimize the process of handling data. These should involve closer integration of data sourcing and tooling during data collection, e.g., automatic scanning for PII in the sources and metadata proposed to the catalogue.

Efficient PII handling is, however, dependent on the quality of non-PII metadata collected. This is especially the case for metadata about language varieties and geographical locations. For example, disparities in detection rates have been shown for names depending on their ethnic and geographic origin, with lowest performance for Black American and Asian/Pacific Islander names in datasets from US institutions (Mansfield et al., 2022). Accurate metadata on the language varieties included in training datasets can therefore inform improved methods for PII identification and anonymization.

9 Conclusion

We have presented our design processes, our human-centered metadata collection efforts, and our resulting successes and challenges in creating a data catalogue targeting 13 language groups. Next steps for the catalogue include translating the form into more languages, filling in missing information for existing entries, and adding more entries to continue efforts toward greater representation across languages and regions. We also plan to update the interactive aspects of the catalogue with more advanced features, i.e., the survey respondents' recommendations and automated screening of new submissions to avoid duplication of entries. The resources within the catalogue (both collected during the hackathons and submitted later) have contributed to the development of the BigScience ROOTS Corpus and the subsequent training of the BLOOM open-access multilingual language model (Laurençon et al., 2022; Scao et al., 2023).

This work produced the data catalogue form, the submission website, and the human-centered methodology of data collection in collaboration with language communities for representative language modeling and other NLP tasks. The catalogue tool remains openly available for use in collecting metadata towards new dataset development projects and for searching existing entries for specific languages and regions. The catalogue form is available for adaption and translation for future documentation and metadata collection efforts to build on. We also discuss a number of challenges, ethical considerations, and recommendations for representative data collection efforts to continue to engage with, particularly in relation to licensing and personally identifiable information. We expect that the form may need to be updated as documentation requirements for NLP and ML systems become regulated and official documentation standards are developed. Scaling the hackathon collection methodology to support larger data collection efforts as well as smaller language communities will require further research and collaboration efforts. Despite these challenges, we hope to encourage others to follow conscientious documentation practices prior to releasing data collections, especially for large-scale NLP applications.

Acknowledgements

The authors would like to thank the hackathon collaborating organizers (Masakhane, Machine Learning Tokyo, and LatinX in AI) and the hackathon participants for their time and efforts. This work would not have been possible without them. The authors would also like to thank the many collaborators and journal reviewers who provided feedback on this paper across its many iterations.

This work was completed while Angelina McMillan-Major was an intern at Hugging Face. Francesco De Toni conducted work on this project while he was a postdoctoral research fellow at the University of Western Australia and at the Forrest Research Foundation. Zeerak Talat conducted work on this project while at Simon Fraser University and while visiting MilaNLP at Bocconi University. Pedro Ortiz Suarez worked on this project while being a PhD student at Inria and at Sorbonne Université, and later a postdoctoral researcher at the Mannheim University. Suzana Ilić conducted the work on this paper while she was working as a technical program manager at Hugging Face. Daniel van Strien conducted work on this project while working on the Living with Machines project at the British Library.

The BigScience Workshop was granted access to the HPC resources of the Institut du développement et des ressources en informatique scientifique (IDRIS) du Centre national de la recherche scientifique (CNRS) under the allocation 2021-A0101012475 made by Grand équipement national de calcul intensif (GENCI).

References

Alyafeai, Zaid, Maraim Masoud, Mustafa Ghaleb, and Maged S. Al-shaibani. 2022. Masader: Metadata Sourcing for Arabic Text and Speech Data Resources. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6340–6351, Marseille, France. European Language Resources Association.

Bandy, John and Nicholas Vincent. 2021. Addressing "Documentation Debt" in Machine Learning: A Retrospective Datasheet for BookCorpus. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Curran.

Barera, Michael. 2020. Mind the Gap: Addressing Structural Equity and Inclusion on Wikipedia. https://rc.library.uta.edu/uta-ir/handle/10106/29572.

Bender, Emily M. 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*, 6:587–604.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Berger-Wolf, Tanya Y, Daniel I Rubenstein, Charles V Stewart, Jason A Holmberg, Jason Parham, Sreejith Menon, Jonathan Crall, Jon Van Oast, Emre Kiciman, and Lucas Joppa. 2017. Wildbook: Crowdsourcing, computer vision, and data science for conservation. *arXiv preprint arXiv:1710.08880*. Presented at the Data For Good Exchange 2017.

Biber, Douglas. 1993. Representativeness in Corpus Design. *Literary and linguistic computing*, 8(4):243–257.

Biderman, Stella, Kieran Bicheno, and Leo Gao. 2022. Datasheet for the Pile. *arXiv preprint arXiv:2201.07311*.

Bird, Steven, Edward Loper, and Ewan Klein. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc.

Birhane, Abeba and Vinay Uday Prabhu. 2021. Large image datasets: A pyrrhic win for computer vision? In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1536–1546.

Birhane, Abeba, Vinay Uday Prabhu, and Emmanuel Kahembwe. 2021. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Cakebread, Caroline. 2017. You're not alone, no one reads terms of service agreements. *Business Insider*.

Causer, Tim and Melissa Terras. 2014. Crowdsourcing Bentham: Beyond the traditional boundaries of academic history. *International Journal of Humanities and Arts Computing*, 8(1):46–64.

DataCite Metadata Working Group. 2021. *DataCite Metadata Schema Documentation for the Publication and Citation of Research Data and Other Research Outputs v4.4*. DataCite.

Dodge, Jesse, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dunn, Jonathan. 2020. Mapping languages: the Corpus of Global Language Use. *Language Resources and Evaluation*, 54:999–1018.

Eberhard, Gary F., David M.and Simons and Charles D. Fennig. 2021. *Ethnologue: Languages of the world*, 24 edition. SIL International.

Gao, Leo, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The Pile: An 800GB Dataset of Diverse Text for Language Modeling. *arXiv preprint arXiv:2101.00027*.

Gao, Leo, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2021. A framework for few-shot language model evaluation. https://doi.org/10.5281/zenodo.5371628. V0.0.1.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Gehrmann, Sebastian, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Chinenye Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Mihir Kale, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly Nikolaev, Andre Niyongabo Rubungo, Salomey Osei, Ankur Parikh, Laura Perez-Beltrachini, Niranjan Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou. 2021. The GEM benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120, Online. Association for Computational Linguistics.

Holland, Sarah, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 2018. The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards. *arXiv preprint arXiv:1805.03677*.

Jernite, Yacine, Huu Nguyen, Stella Biderman, Anna Rogers, Maraim Masoud, Valentin Danchev, Samson Tan, Alexandra Sasha Luccioni, Nishant Subramani, Isaac Johnson, Gerard Dupont, Jesse Dodge, Kyle Lo, Zeerak Talat, Dragomir Radev, Aaron Gokaslan, Somaieh Nikpoor, 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, page 2206–2222, New York, NY, USA. Association for Computing Machinery.

Jo, Eun Seo and Timnit Gebru. 2020. Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 306–316, New York, NY, USA. Association for Computing Machinery.

Kreutzer, Julia, Isaac Caswell, 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 Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, 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. 2022. Quality at a Glance: An Audit of Web-

Crawled Multilingual Datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Kučera, Henry and Winthrop Nelson Francis. 1967. *Computational analysis of present-day American English*. Brown University Press, Providence, RI.

Laurençon, Hugo, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, Jörg Frohberg, Mario Šaško, Quentin Lhoest, Angelina McMillan-Major, Gerard Dupont, Stella Biderman, Anna Rogers, Loubna Ben allal, Francesco De Toni, Giada Pistilli, Olivier Nguyen, Somaieh Nikpoor, Maraim Masoud, Pierre Colombo, Javier de la Rosa, Paulo Villegas, Tristan Thrush, Shayne Longpre, Sebastian Nagel, Leon Weber, Manuel Muñoz, Jian Zhu, Daniel Van Strien, Zaid Alyafeai, Khalid Almubarak, Minh Chien Vu, Itziar Gonzalez-Dios, Aitor Soroa, Kyle Lo, Manan Dey, Pedro Ortiz Suarez, Aaron Gokaslan, Shamik Bose, David Adelani, Long Phan, Hieu Tran, Ian Yu, Suhas Pai, Jenny Chim, Violette Lepercq, Suzana Ilic, Margaret Mitchell, Sasha Alexandra Luccioni, and Yacine Jernite. 2022. The BigScience ROOTS Corpus: A 1.6TB Composite Multilingual Dataset. In *Advances in Neural Information Processing Systems*, volume 35, pages 31809–31826. Curran Associates, Inc.

Lhoest, Quentin, 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*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Luccioni, Alexandra and Joseph Viviano. 2021. What's in the Box? An Analysis of Undesirable Content in the Common Crawl Corpus. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 182–189, Online. Association for Computational Linguistics.

Mansfield, Courtney, Amandalynne Paullada, and Kristen Howell. 2022. Behind the Mask: Demographic bias in name detection for PII masking. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 76–89, Dublin, Ireland. Association for Computational Linguistics.

McMillan-Major, Angelina, Emily M. Bender, and Batya Friedman. 2023. Data Statements: From Technical Concept to Community Practice. *ACM J. Responsib. Comput.* Just Accepted.

Obar, Jonathan A. and Anne Oeldorf-Hirsch. 2020. The biggest lie on the Internet: ignoring the privacy policies and terms of service policies of social networking services. *Information, Communication & Society*, 23(1):128–147.

Ortiz Suárez, Pedro Javier, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. In *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7)*, pages 9 – 16, Cardiff, UK. Leibniz-Institut für Deutsche Sprache.

Padilla, Thomas, Laurie Allen, Hannah Frost, Sarah Potvin, Elizabeth Russey Roke, and Stewart Varner. 2019. Final Report — Always Already Computational: Collections as Data. https://doi.org/10.5281/zenodo.3152935.

Paullada, Amandalynne, 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):100336.

Phillips, Addison and Mark Davis. 2009. Tags for Identifying Languages. RFC 5646.

Raddick, M Jordan, Georgia Bracey, Pamela L Gay, Chris J Lintott, Carie Cardamone, Phil Murray, Kevin Schawinski, Alexander S Szalay, and Jan Vandenberg. 2013. Galaxy Zoo: Motivations of citizen scientists. *arXiv preprint arXiv:1303.6886*.

Rae, Jack W., Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic

Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2021. Scaling Language Models: Methods, Analysis & Insights from Training Gopher. *arXiv preprint arXiv:2112.11446*.

Ranchordás, Sofia. 2022. Connected but still excluded?: Digital exclusion beyond internet access. In Marcello Ienca, Oreste Pollicino, Laura Liguori, Elisa Stefanini, and Roberto Andorno, editors, *The Cambridge handbook of information technology, life sciences and human rights*, Cambridge Law Handbooks, page 244–258. Cambridge University Press.

Ridge, Mia, Samantha Blickhan, Meghan Ferriter, Austin Mast, Ben Brumfield, Brendon Wilkins, Daria Cybulska, Denise Burgher, Jim Casey, Kurt Luther, Michael Haley Goldman, Nick White, Pip Willcox, Sara Carlstead Brumfield, Sonya J. Coleman, and Ylva Berglund Prytz. 2021. Choosing tasks and workflows. In *The collective wisdom Handbook: Perspectives on crowdsourcing in cultural heritage - Community review version*, 1 edition. Digital Scholarship at the British Library. Https://britishlibrary.pubpub.org/pub/choosing-tasks-and-workflows.

Sambasivan, Nithya, 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*, New York, NY, USA. Association for Computing Machinery.

Scao, Teven Le, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina Mcmillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco de Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro von Werra, Leon Weber, Long Phan, Loubna Ben Allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-Shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lover-

ing, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh Hajihosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael Mckenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel de Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-Aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2023. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. Working paper or preprint.

Segal, Avi, Ya'akov Gal, Ece Kamar, Eric Horvitz, Alex Bowyer, and Grant Miller. 2016. Intervention strategies for increasing engagement in crowdsourcing: Platform, predictions, and experiments. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 3861–3867.

Shah, Chirag, Theresa Anderson, Loni Hagen, and Yin Zhang. 2021. An iSchool approach to data science: Human-centered, socially responsible, and context-driven. *Journal of the American Society for Information Science and Technology*, 72(6):793–796.

Srivastava, Aarohi, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Johan Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew M. Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, Cesar Ferri, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Christopher Waites, Christian Voigt, Christopher D Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, C. Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra,

Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodolà, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Xinyue Wang, Gonzalo Jaimovitch-Lopez, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Francis Anthony Shevlin, Hinrich Schuetze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B Simon, James Koppel, James Zheng, James Zou, Jan Kocon, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh Dhole, Kevin Gimpel, Kevin Omondi, Kory Wallace Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros-Colón, Luke Metz, Lütfi Kerem Senel, Maarten Bosma, Maarten Sap, Maartje Ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramirez-Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael Andrew Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan Andrew Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter W Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan Le Bras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Russ Salakhutdinov, Ryan Andrew Chi, Seungjae Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel Stern Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima Shammie Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven Piantadosi, Stuart Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsunori Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Venkatesh Ramasesh, vinay uday prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. 2023. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. *Transactions on*

Machine Learning Research.

Stoyanovich, Julia and Bill Howe. 2019. Nutritional labels for data and models. *A Quarterly bulletin of the Computer Society of the IEEE Technical Committee on Data Engineering*, 42(3).

Tensor Flow Authors. 2021. TensorFlow Datasets, a collection of ready-to-use datasets. `https://www.tensorflow.org/datasets`.

Wang, Boxin, 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. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

A Images of the Submission Form

In this appendix we provide screenshots of the submission form for a view of the version of the form at the time of writing, organized in order of appearance within the submission form. Section A.1 shows the *Language and Locations* portion of the submission form in Figures 3, 4, 5, and 6. Section A.2 shows the *Representative, Owner, or Custodian* portion of the submission form in Figure 7. Section A.3 shows the *Availability of the Resource* portion of the submission form in Figures 8, 9, and 10. Section A.4 shows the *Primary Source Type* portion of the submission form in Figures 11 and 12. Finally, Section A.5 shows the *Media Type, Format, Size, and Processing* portion of the submission form in Figures 13 and 14.

A.1 Languages and Locations

The *Languages and Locations* section of the catalogue submission form presents the user with a dropdown list of the languages selected as primary targets for the BigScience project. Multiple languages may be selected. A textbox also allows users to add comments about the language varieties, such as the presence of dialectal variation or code-switching. Figure 3 shows the dropdown without any languages selected.

Some of the selections refer to language families rather than individual languages, in which case a specific language within that family may be selected from a secondary dropdown list. Figure 4 shows the first dropdown with the 'African languages of the Niger-congo family' tag selected and isiZulu selected as a specific language tag within that family.

A checkbox allows users to indicate they would like to include a language outside the set of targeted languages. When the checkbox is selected, it makes another dropdown visible which allows users to select languages from a list generated from BCP-47 language tag best practices (Phillips and Davis, 2009). Figure 5 shows the checkbox selected and the language tag for Afar (ISO 639-3 language code: aar) added.

After selecting the language tags for the resource, the user may then select country and region location tags using two dropdown lists. The first dropdown list allows users to select from a list of continents, world areas, and country groups (e.g., Australia and New Zealand). The second dropdown list allows users to select from a list of individual countries, nations, regions, and territories. Figure 6 shows an example of overlapping tags with two macroscopic area tags for Oceania as well as Australia and New Zealand selected and the country tag for Australia selected.

A.2 Representative, Owner, or Custodian

The *Representative, Owner, or Custodian* section of the submission form presents the user with several questions regarding the custodian of the resource, including the name, entity type (e.g., organization, library, or individual), and contact information for the custodian. Figure 7 shows a dropdown question for whether the data custodian is already in the catalogue, a text field for the name of the data custodian if not already in the catalogue, and a dropdown question to select the entity type for the custodian.

If the submission user selects a custodian from the dropdown list of custodians already in the catalogue (e.g., Global Voices), the remainder of the questions for the *Representative, Owner, or Custodian* are no longer shown to the user. The entity type and contact information are populated with the existing information in the catalogue to reduce the submission completing time.

A.3 Availability of the Resource

The *Availability of the Resource: Procuring, Licenses, PII* section of the submission form contains three subsections related to procuring the resource (Figure 8), the license and/or terms of service for the resource (Figure 9), and personal identifying information (PII) within the resource (Figure 10; see Section 8.2 for our discussion of PII).

As shown in Figure 8, the submission form first asks users to characterize the availability of the resource with one of four possible answers: 1) yes, it has a direct download link or links; 2) yes, after signing a user agreement; 3) no, but the current owners/custodians have contact information for data queries; and 4) no, we would need to spontaneously reach out to the current owners/custodians. If the selected response indicates the data can be downloaded, the user is asked for a URL. Otherwise, the form asks users to provide the email of the person to contact to obtain the data if it is different from the contact email entered for the data custodian in the *Representative, Owner, or Custodian* section.

The first question for the resource licensing terms is simply whether or not the language data in the resource come with explicit licenses of terms of use. If the user responds yes, as is the case in Figure 9, the submission form displays a dropdown question for the user to select the best characterization(s) of the licensing status of the data: public domain, multiple licenses, copyright - all rights reserved, open license, research use, non-commercial use, or do not distribute. Users may then further specify specific licenses from a dropdown and include the terms of use or license text by coping it into a textbox area. If there are no licenses or terms of service, or if it is unclear as to what they are, the user is asked to provide their best assessment of whether the data can

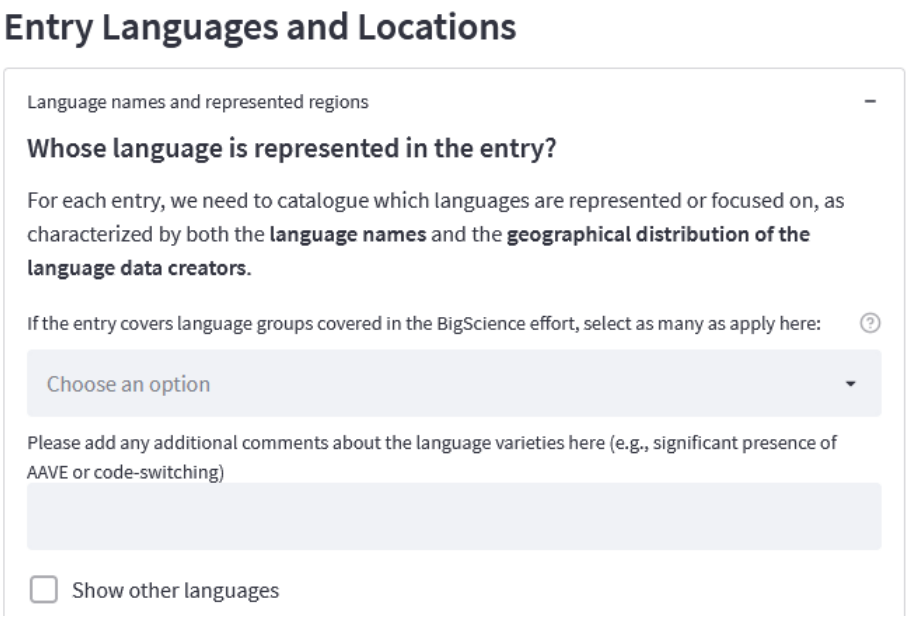

Figure 3: The *Language* section of the submission form for the catalogue.

be used to train models while respecting the rights and wishes of the data creators and custodians.

To support submission form users with identifying PII concerns, we introduced three categories of PII: general information including names, physical and email addresses, etc.; numeric information such as telephone numbers, fax numbers, social security numbers, etc.; and sensitive information such as descriptions of racial or ethnic origin, political opinions, and religious or philosophical beliefs. The form first asks submission form users whether the resource contains any of these kinds of personally identifiable or sensitive information with options for 'yes', 'yes - text author name only', 'no', or 'unclear'. If the user indicates that the resource does contain PII, as shown in Figure 10, the submission form then presents three dropdown questions for the user to indicate how likely it is that the resource contains each kind of personally identifiable or sensitive information: very likely, somewhat likely, unlikely, or none. If the user indicates no or unclear when responding to whether or not the resource contains PII, the submission form presents options for explaining why there may not be PII in the data. The options include that the data only contains general knowledge not written by or referring to private persons, that the data consists of fictional text, and other, in which case the user can provide their own explanation in a textbox.

A.4 Primary Source Type

The questions asked in the *Primary Source Type* section of the submission form depend on whether resource being submitted is an original data source or an existing dataset that has been processed and released for ML or NLP tasks. Figure 11 shows the questions posed in the event that the resource is an original data source. Figure 12 shows the questions asked if the resource is an existing dataset.

The first dropdown of the questions for original sources allows users to describe the resource as either a collection, website, or some other user-provided description. The second dropdown provides a list of further categorize the collection or website, or provides a textbox for the user-provided description to be clarified. In Figure 11, 'collection' is selected for the resource type and 'books/book publisher' is selected for the kind of collection.

Because we assume that processed datasets are already collections, we instead focus the questions for processed datasets on the primary sources from which the dataset was created. The form provides users with the option of stating that the data was created for the purpose of including it in the dataset or that the data was taken from other primary sources. If the data was taken from other primary sources, as shown in Figure 12, the form the provides four options for describing whether the primary sources are available to investigate: 1) yes because the sources are documented; 2) yes because the sources are fully available; 3) no because they are private; and 4) no because the data sources are secret. The submission form user may then select the primary sources from a dropdown if they are already entered in the catalogue to link the primary sources and the processed dataset. A second dropdown then allows users to categorize the primary sources as websites or collections of data sources like when submitting an original

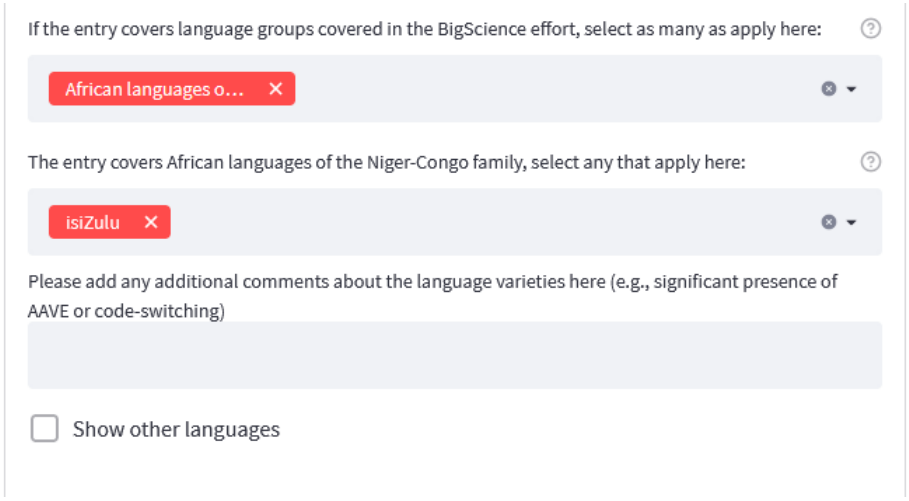

Figure 4: The *Language* section of the submission form for the catalogue with the 'African languages of the Niger-congo family' tag selected and the isiZulu language tag selected.

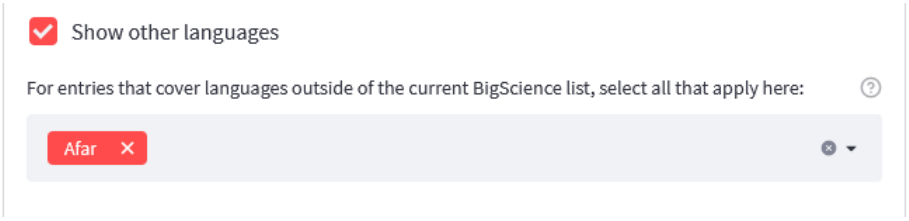

Figure 5: The *Language* section of the submission form for the catalogue with the checkbox for other languages selected and the language tag for Afar (ISO 639-3 language code: aar) added.

data source. Finally, the submission form presents the users with several options for determining the agreement between the license of the processed dataset and the license of the source data: 1) the license is unknown to the submission user; 2) the source data has an open license; 3) the dataset has the same license as its source data; 4) the dataset curators obtained consent from the source data owners; and 5) the source data license disallows re-use.

A.5 Media Type, Format, Size, and Processing

The *Media Type, Format, Size, and Processing* section contains questions concerning the technical aspects of digitizing physical data sources and processing digital data sources for language modeling. Figure 13 shows the questions concerning the media type of the data and Figure 14 shows the questions regarding the amount of data in the resource.

To categorize the data type(s) within the resource, the form allows users to select tags indicating that the data is primarily text, audiovisual (from either video or audio recordings), and/or image data. If the data is primarily text, users can then select several format tags for the data including plain text, HTML, PDF, XML, mediawiki, or other. Similarly, if the data is primarily audiovisual, users can select the format tags from mp4, wav, video, and other, and if the data is primarily images, the presented formats are JPEG, PNG, PDF, TIFF, and other. If the media type tag for text was selected (but not audiovisual or image types), the submission form then asks users to select whether the text was transcribed from another media format and, if so, whether that media format was audiovisual or images. Figure 13 shows these additional questions when the media type tag for text is selected.

Bytes are difficult to estimate, so the submission form instead asks users to define an instance unit for the resource and then estimate the resource size in terms of that unit. Figure 14 shows the three dropdown questions we designed to help users with their estimations of the amount of data in the resource. The first drop down allows users to select their definition of a data instance within the resource from either an article, post, dialogue, episode, book, or other. Users are then prompted to select an estimate the number of instances in the resource on the order of hundreds, thousands, tens of thousands, hundreds of thousands, or millions. Additionally, users may select an estimate of the number of words per in-

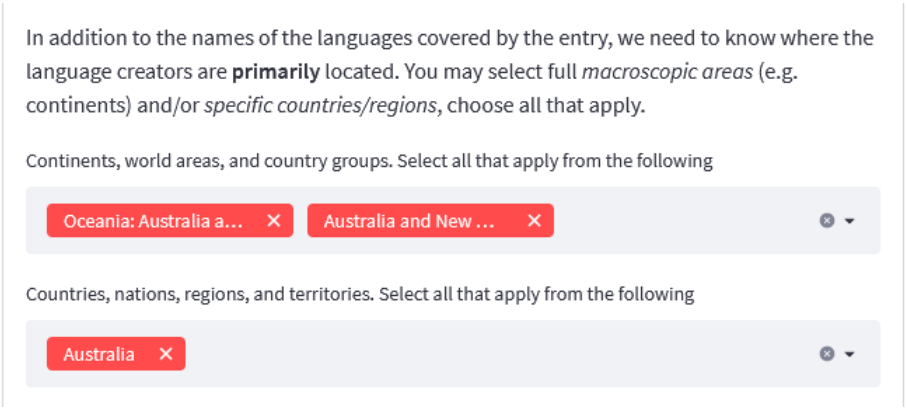

Figure 6: The *Location* section of the submission form with two macroscopic area tags for Oceania as well as Australia and New Zealand selected and the country tag for Australia selected.

stance in similar ranges. Submission form users were encouraged to select their best estimates for these questions even if they were uncertain.

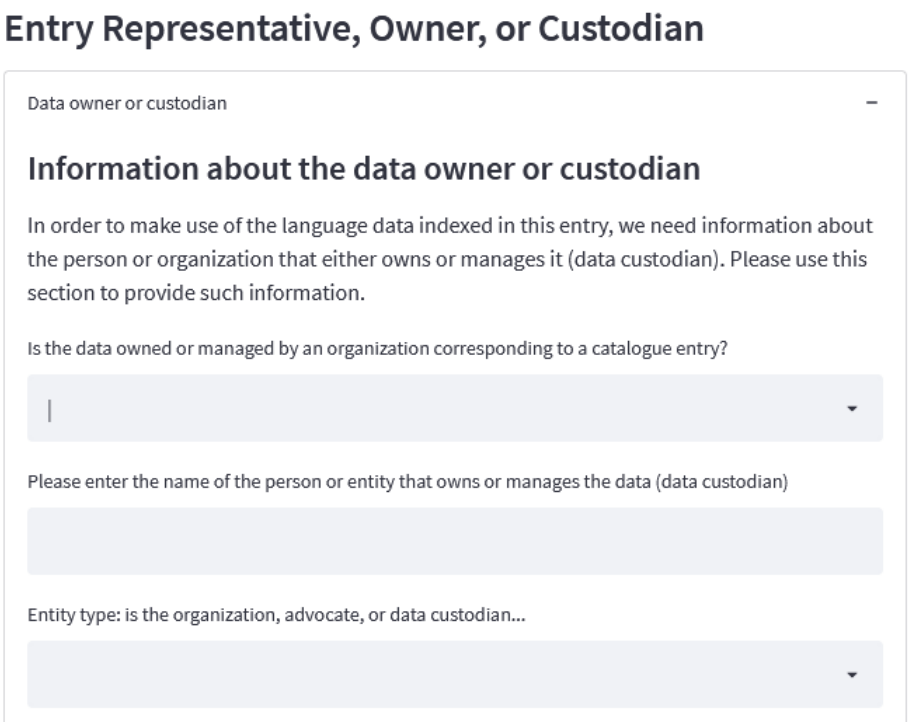

Figure 7: The *Representative, Owner, or Custodian* section of the submission form for the catalogue.

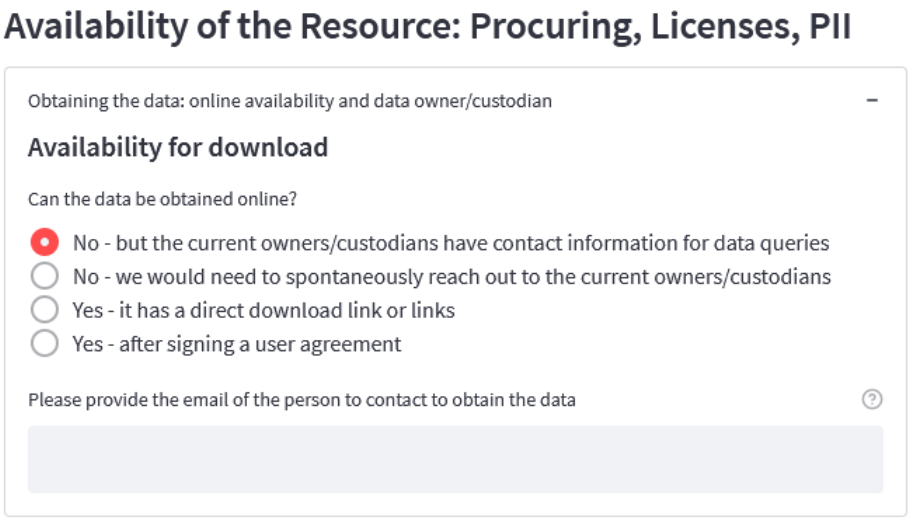

Figure 8: Options for describing whether a resource may be downloaded in the *Availability of the Resource: Procuring, Licenses, PII* section of the form.

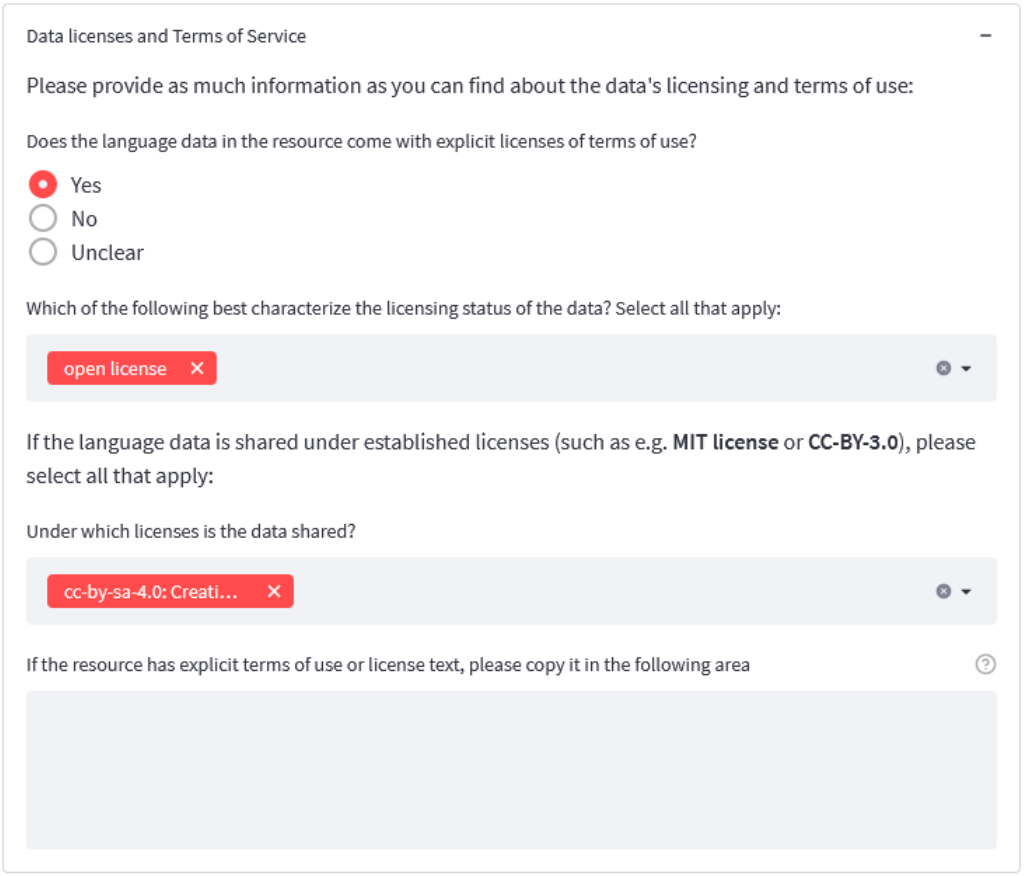

Figure 9: The questions for characterizing the licensing terms of the resource in the *Availability of the Resource: Procuring, Licenses, PII* section of the form with the tag for an open license and the CC-BY-SA-4.0 tag selected.

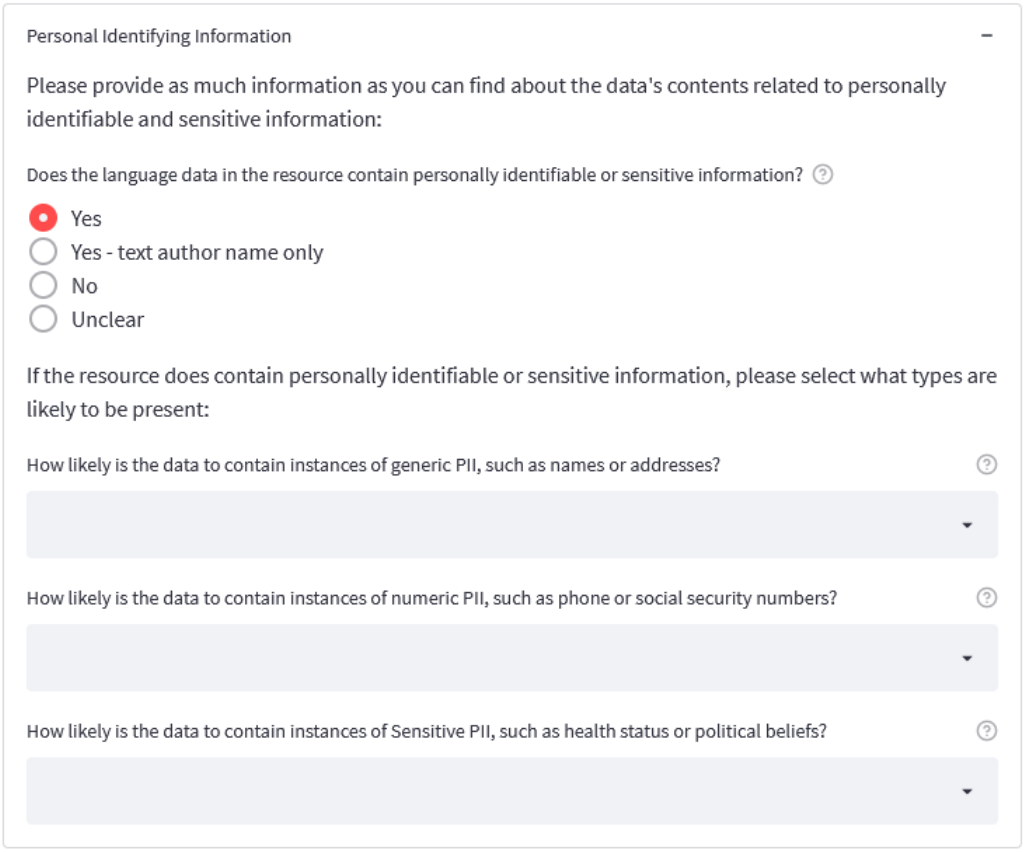

Figure 10: The questions for characterizing the types of PII in the resource in the *Availability of the Resource: Procuring, Licenses, PII* section of the form.

Figure 11: The *Primary Source Type* section of the submission form for original data source with 'collection' selected for the resource type and 'books/book publisher' selected for the kind of collection.

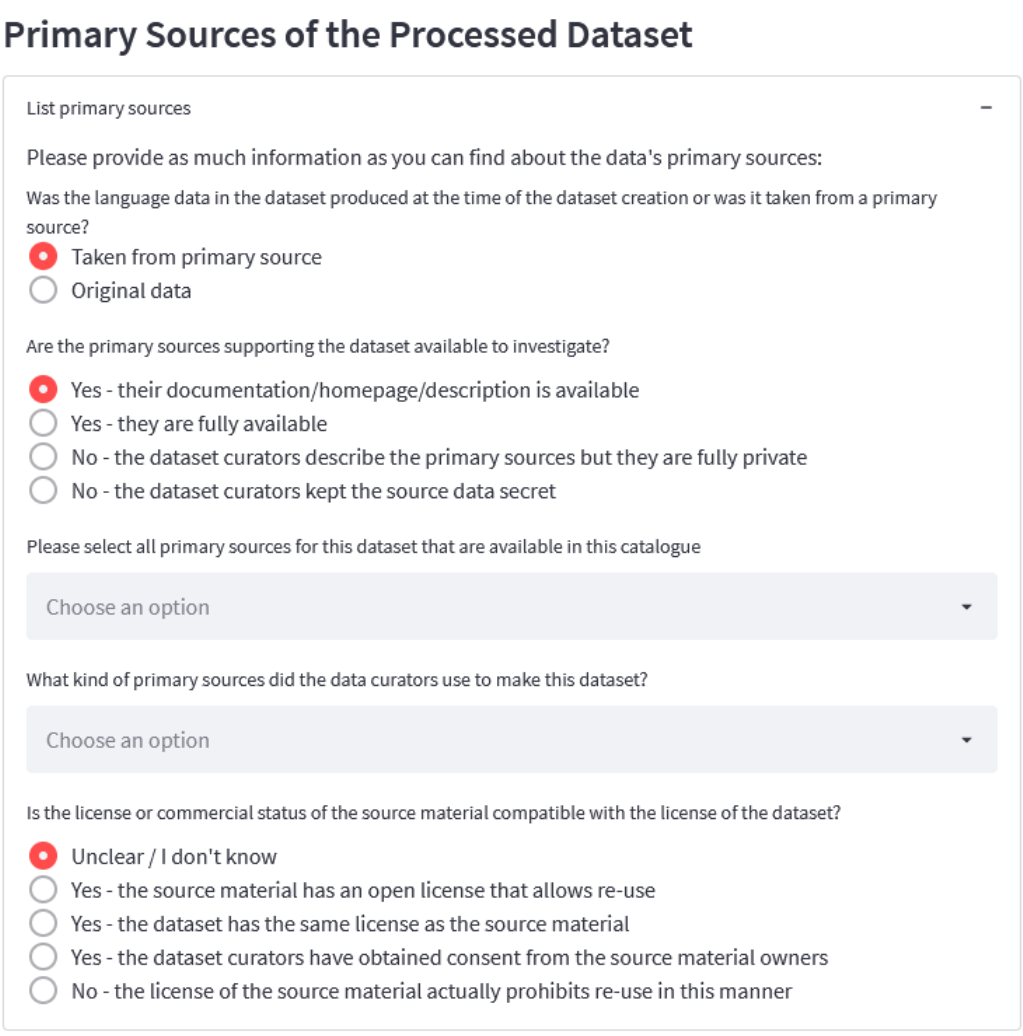

Figure 12: The *Primary Source Type* section of the submission form for processed datasets.

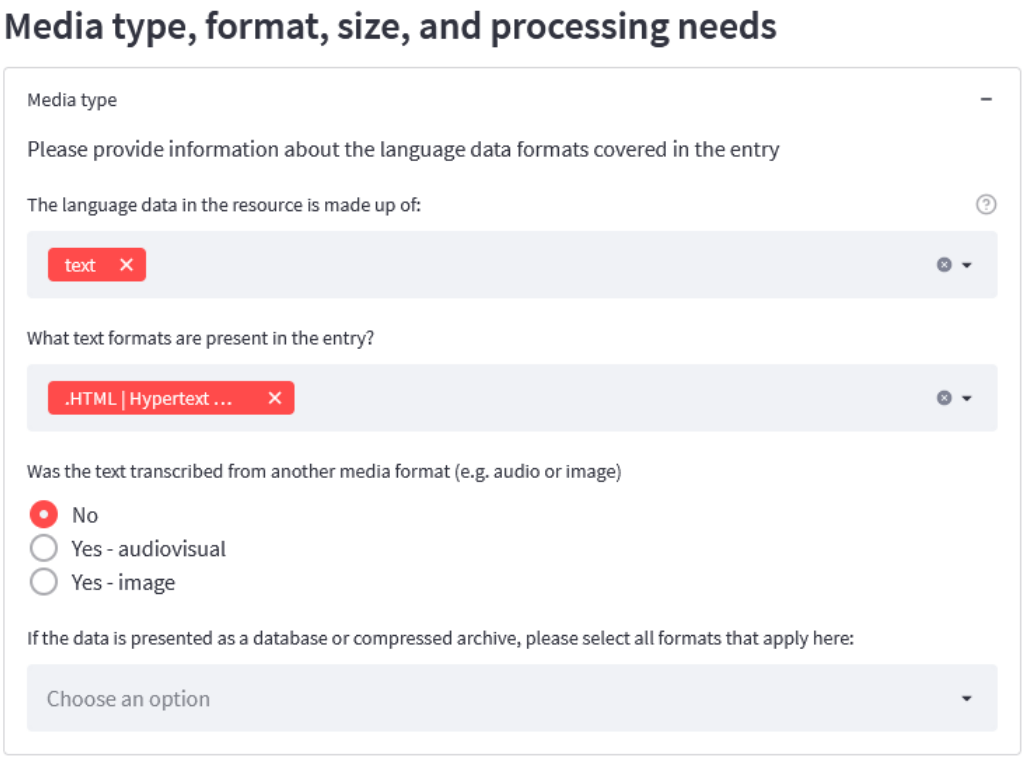

Figure 13: The subsection for estimating the media types in the *Media type, format, size, and processing needs* section of the submission form with tags for 'text' and 'HTML' selected.

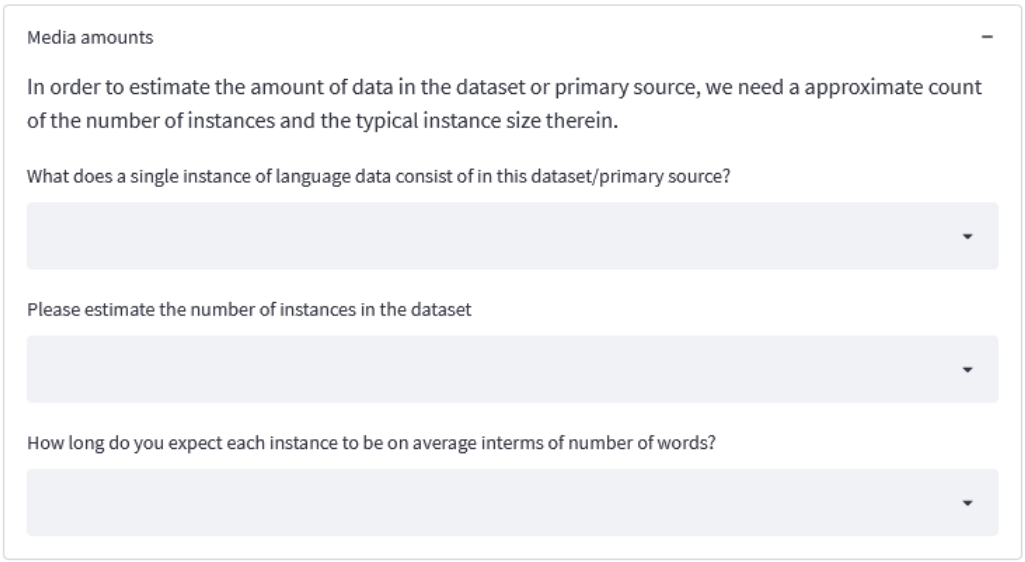

Figure 14: The subsection for estimating the media amounts in the *Media type, format, size, and processing needs* section of the submission form.