DATA PROVIDER-HOST AGREEMENT

v0.1 (input DATE)

1. PREAMBLE

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

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

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

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

BigScience takes an alternative approach of identifying Data sources for a training corpus. Specifically, our participants built an annotated catalog of high-quality language resources to cover the diversity of languages and social contexts that should make up such a training corpus. There are two essential parties in charge of making this data available, under the auspices of BigScience: First, the **Data Providers**, any institution willing to license datasets of interest purely for research purposes on a royalty free basis; and Second, the **Data Hosts**, institutions willing to contribute their technical capabilities in order to host the data provided, enabling society to access it. These are the champions of data sharing and openness in research.

This License governs the use of Data as informed by the BigScience Ethical Charter and the values set forth in the BigScience workshop. These establish the perspective informing this license that text and language are above all human-centric data. This means that data subjects have inherent rights and protections, and interests that exist outside of its Machine Learning context, and which we also need to account for.

Although the BigScience community does not aim to impose its values on potential users of the Data, it is determined to take tangible steps towards protecting the community from inappropriate uses of the work being developed by BigScience.

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

2. DEFINITIONS

"Agreement" means this Agreement including all its Exhibits.

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

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

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

"**Dataset**" means one specific collection of Data that the Data Provider has the necessary rights for sharing under this agreement.

"Processed Dataset" is a Dataset that is further processed via Data transformations, including additional modifications to one dataset (e.g., personal information removal, additional annotations, extracted text, subsetting by language, removal of individual data points), dataset combinations, etc.

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

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

"Derived Work" means any artifact created using Data covered by this Agreement.

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

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

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

3. PURPOSE, RIGHTS GRANTED & SCOPE

The Data Provider grants to the Data Host a non-exclusive, non-transferable, non-sublicensable, irrevocable, perpetual, royalty-free and worldwide license to use (that is access, store, prepare, process, label and/or share) the agreed upon Data (see List of

Datasets in Exhibit A) in accordance with the use case scenarios and further (re)distribution policy, as stated in Annex III (see below).

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

5. DATA HOST RIGHTS AND OBLIGATIONS
    a. No rights are granted to the Data Host with respect to the Data other than those stipulated in this Agreement, except any exceptions or limitations provided by law.
    b. The Data Host will hold harmless the Data Provider against any claims, demands, suits or damages arising from the use of the Data in accordance with the purpose set out in this Agreement.
    c. The Data Host acknowledges and agrees that the following disclaimers apply to all End-Users and other entities who have access to the Data The Data is provided “as-is”.
    d. In case of becoming knowledgeable of any actual or suspected intellectual property rights infringement, damages or claims associated with the Data the

Data Host will promptly notify the Data Provider and stop the further usage of the Data until the issue in question can be resolved.

e. The Data Host will not assert rights over any Data (excluding Meta-Data) made available by way of this Agreement.
f. The Data Host will undertake commercially reasonable efforts to appropriately attribute the Data Provider as the source of the Data.
g. The Data Host is allowed to create and publish research (including benchmarks, performance indicators and/or scientific insights) gained using the data under this agreement, for the purposes of BigScience's research scope.
h. The Data Host will undertake commercially reasonable efforts to remove personal data and information from the Dataset before using the Dataset Notwithstanding the latter undertaking, the Data Provider should, under Section 4(j) of this agreement, inform the Data Host in case the former is aware of the existence of personal data under the licensed dataset(s).

6. LIMITATIONS

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

7. FEES AND COSTS

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

8. SECURITY

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

9. TERM AND TERMINATION

a. This Agreement is valid from the date the involved parties agree, or by default, from the moment it is signed by all the involved parties.
b. The term of this Agreement shall be from the Agreement Date until the last to expire of the Data Provider's intellectual property rights or any related rights on the licensed Dataset, strictly for the purpose of this Agreement.

c. This Agreement can be terminated by either party immediately in case the other party breaches this Agreement upon due notice of it and the breach is not remediated within 30 days.
d. In case either party would like to voluntarily terminate the Agreement, it shall act in good faith and provide the other party with (i) a reasoned statement justifying the decision; (ii) a 60 days pre-advice; (iii) and, give the other party the opportunity to negotiate any new terms and conditions for the sake of the Agreement's continuity, and beyond, for the sake of BigScience's research goals.
e. Upon termination of this Agreement for any reason, the Data Host and Data Provider cease to use the Data and Processed Datasets and for the purposes set out in this Agreement within 14 days and upon that delete the Data and Processed Datasets immediately (respecting any holding periods or processing information storage required by the law) . This does not affect already completed or created Derived Work before the termination of this Agreement.

10. FORCE MAJEURE

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

11. CONFIDENTIALITY

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

12. ENTIRE AGREEMENT

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

13. MODIFICATION AND AMENDMENT

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

14. DISPUTE RESOLUTION

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

15. SURVIVAL

The provisions set forth in section 6(b) (Limits of Liability), 10 (Term and Termination), 12 (Confidentiality), 15 (Governing Law), 16 (Survival) and Exhibit A (Section Restrictions of

Use in the Dataset section) shall survive the termination of this agreement and continue to bind both parties.

16. SEVERABILITY

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

17. NO ADDITIONAL TERMS

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

18. FULL UNDERSTANDING

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

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

Annex

DATA PROVIDER SCHEDULE (EXHIBIT A)

I. Data Provider Information

Data Provider Name

Data Address

Data Provider Contact Information

Data Provider Name

Special Conditions (*if applicable*)

Data Management Plan

II. Datasets

List of Datasets

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

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

III. Field of Use

Scope / use cases:

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

IV. Data Distribution Policy

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

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

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

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

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

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

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

RAIL Model License (EXHIBIT B)

Include once finished:
https://docs.google.com/document/d/117RhytMYC9HS-1NmWHEn9XBK7vJ5kdv9OcG6AV69Vec

Potential further clauses:

X. CONFLICT RESOLUTION
In the case of any dispute, the parties shall attempt to resolve the issue by negotiation first. In the case such negotiations cannot resolve the issue within six months either party may bring the issue to the applicable court of law.

X. NO WAIVER
The failure of the Data Host or Data Provider to enforce or execute any right or provision of this Agreement shall not constitute a waiver of that right or provision.

X TITLES
Headings and Section titles in this Agreement are only for convenience and are not to be considered in construing this Agreement.