

Reclaiming Open Science in the Age of AI

Panel: AI, Open Science, and the Global Digital Divide

**United Nations Open Science and Open Scholarship
Conference**

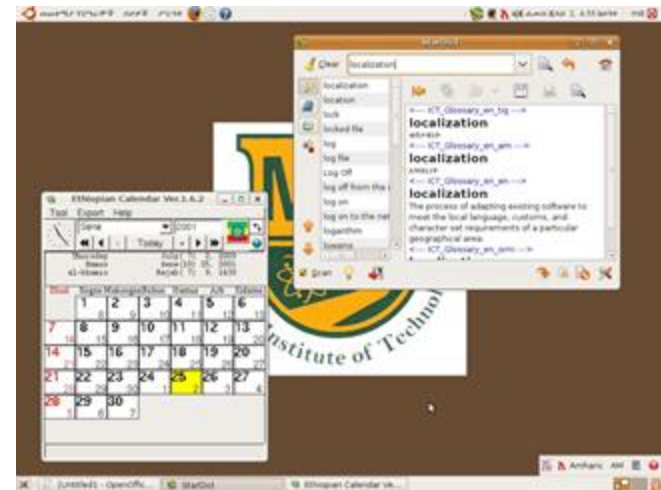
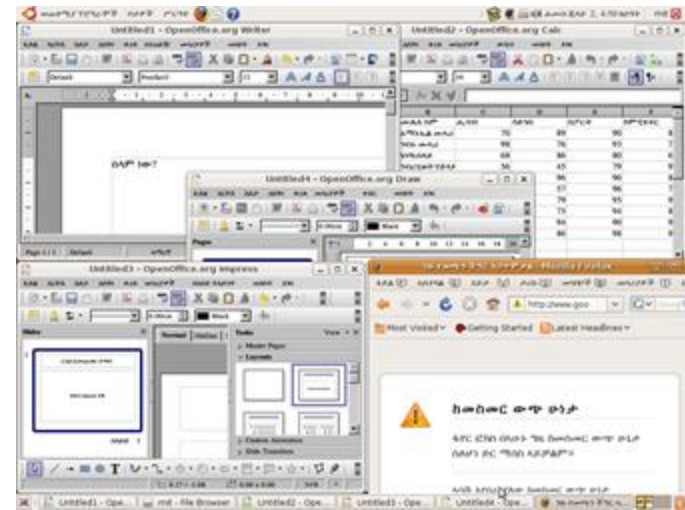


Asmelash Teka Hadgu

Undergraduate project in Tigray, Ethiopia.

- **The Goal:** Localize Ubuntu, OpenOffice, and Firefox into Amharic and Tigrinya.
- **The Enabler:** Free and Open Source Software (FOSS).

It allowed us to learn, modify, and build for our community's needs.



UNESCO Definition:

"Open science is a set of principles and practices that aim to make scientific research... accessible to everyone for the benefits of scientists and society as a whole."

Lesan a community rooted approach

High-Quality, Curated Data: We solve specific problems, we don't just scrape the web.

Deep Community Knowledge: We incorporate linguistic nuances like dialects.

Ethical Sourcing: We pay contributors fairly for their expertise.

This stands in stark contrast to the dominant paradigm:
"scale is all you need"

HornMT: An open benchmark dataset for machine translation

Freely available on GitHub and Zenodo.

Oromo (Orm)

Amharic (Amh)

Tigrinya (Tir)

Somali (Som)

Afar (Aaf)



Asmelash Teka Hadgu, Gebrekirstos G. Gebremeskel, & Abel Aregawi. (2022). HornMT – Machine Translation Benchmark Dataset for Languages in the Horn of Africa (1.0.0) [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.6369442>

The Hijacking of "Open (Science)"

A Dangerous Trend: The term "Open (Science)" is being co-opted by Big Tech.

It has become a **powerful marketing weapon** to:

- Dominate the public narrative
- Siphon funding and resources
- Consolidate market power
- Undermine genuine, local innovation

Meta's "No Language Left Behind" (NLLB)

RESEARCH



200 languages within a single AI model: A breakthrough in high-quality machine translation

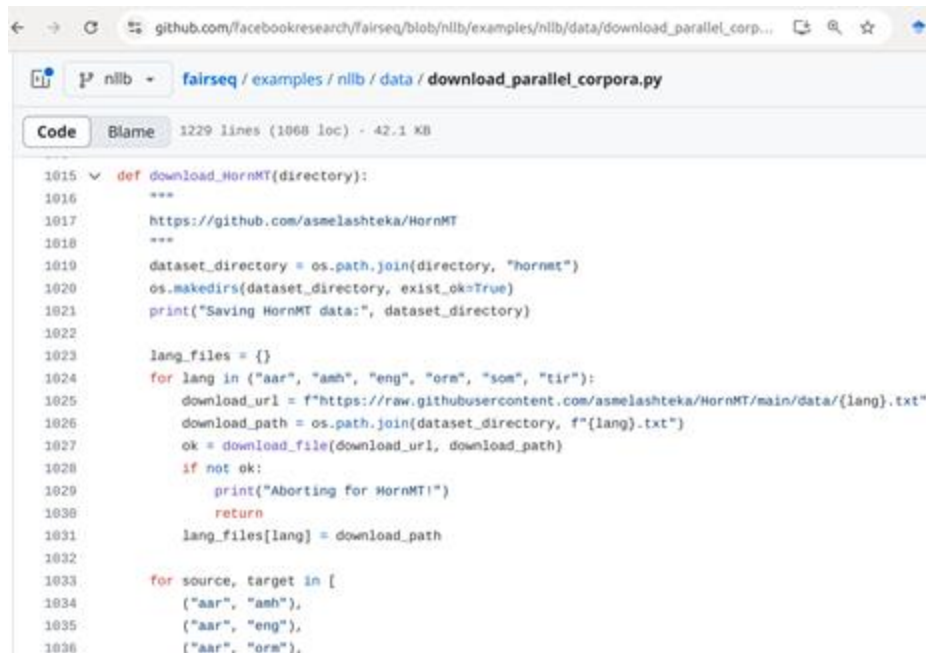


"[...] We've built a single AI model called NLLB-200, which translates 200 different languages with state-of-the-art results.

[...] NLLB-200 supports 55 African languages with high-quality results."

Companies violate the core ethos of open science

- **Scientific Integrity:** Using evaluation dataset as a training data
- **Attribution:** Failed to acknowledge or cite our HornMT dataset
- **Licensing:** Built model on data they didn't have a license for, eg., JW300



The screenshot shows a GitHub web interface for the repository `facebookresearch/fairseq/blob/nlib/examples/nlib/data/download_parallel_corpora.py`. The file is 1229 lines long, with 1068 lines of code and 42.1 KB in size. The code is a Python script that defines a function `download_HornMT(directory)`. This function is responsible for downloading the HornMT dataset, which is a parallel corpus. It sets the `dataset_directory` to `os.path.join(directory, "hornmt")` and creates it if it doesn't exist. It then defines a dictionary `lang_files` to store the download paths for different languages. The function iterates over a list of languages: `("aar", "amh", "eng", "orm", "som", "tir")`. For each language, it constructs a `download_url` pointing to a raw GitHub file, downloads it, and saves it in the `dataset_directory`. Finally, it updates the `lang_files` dictionary with the download path for each language. The script also includes a `for source, target in` loop at the bottom, which is partially visible.

```
1015 def download_HornMT(directory):
1016     """
1017     https://github.com/asmelashteka/HornMT
1018     """
1019     dataset_directory = os.path.join(directory, "hornmt")
1020     os.makedirs(dataset_directory, exist_ok=True)
1021     print("Saving HornMT data:", dataset_directory)
1022
1023     lang_files = {}
1024     for lang in ("aar", "amh", "eng", "orm", "som", "tir"):
1025         download_url = f"https://raw.githubusercontent.com/asmelashteka/HornMT/main/data/{lang}.txt"
1026         download_path = os.path.join(dataset_directory, f"{lang}.txt")
1027         ok = download_file(download_url, download_path)
1028         if not ok:
1029             print("Aborting for HornMT!")
1030             return
1031         lang_files[lang] = download_path
1032
1033     for source, target in [
1034         ("aar", "amh"),
1035         ("aar", "eng"),
1036         ("aar", "orm"),
```

The Consequence: A Widening Digital Divide

A vicious cycle:

- Big Tech releases overhyped “open” models
- Their results become the de facto standard
- Local innovators are forced to release data
- Big Tech reuses that data, reinforcing control



**Federation of community rooted startups
and organization for African languages**



Lesan



Ghana**NLP**



LELAPA AI

Open Science should be the tool to close the digital divide, not the brand that widens it.

Let's commit to a future where AI is built with and for all of humanity.

Thank you!
asme@lesan.ai