ACCELERATED DATA DISCOVERY FOR SCALABLE CLIMATE ACTION

Henning Schwabe
Henning_Schwabe@icloud.com

Sumeet Sandhu
CEO, Elementary IP LLC,
Santa Clara CA, USA
sumeet.k.sandhu@gmail.com

Sergy Yu. Grebenshchikov
sgreben@gwdg.de

ABSTRACT

According to the Intergovernmental Panel on Climate Change (IPCC), the planet must decarbonize by 50% by 2030 in order to keep global warming below 1.5°C. This goal calls for a prompt and massive deployment of solutions in all societal sectors - research, governance, finance, commerce, health care, consumption. One challenge for experts and non-experts is access to the rapidly growing body of relevant information, which is currently scattered across many weakly linked domains of expertise. We propose a large-scale, semi-automatic, AI-based discovery system to collect, tag, and semantically index this information. The ultimate goal is a near real-time, partially curated data catalog of global climate information for rapidly scalable climate action.

1 INTRODUCTION

To keep warming below 1.5°C, the planet must decarbonize by 50% by 2030, and by 100% by 2050, with net negative emissions afterwards (IPCC, 2018). This leaves just one decade for concerted climate action, which has to embrace multiple technical disciplines and multiple sectors of economy and society.

Climate action differs from established climate modeling in the highly upscaled data demand. It requires the integration of several heterogeneous tasks and datasets currently curated by separate expert communities. With growing awareness, many individuals and organizations, ranging from IT to agricultural communities, are looking for ways to contribute towards climate action. In the current setting, every expert in a given sub-domain is a novice (a 'non-expert') in an adjacent sub-domain. In order to scale climate action rapidly over the next few years, 'non-experts' need to be onboarded efficiently, while the established communities are called upon to open up and remove their legacy boundaries (D.E. Jensen et al, 2020a)

In this contribution, we propose to address the lack of data transparency and usability in the climate change context using a large-scale data discovery system tailored to machine learning tasks. The system is designed to be able to track, semantically index, and search the existing and emerging climate-related data sources. The ultimate goal is a large-scale data catalog for rapid climate action that is supported by task-specific interest groups and experts.

2 PROBLEM STATEMENT

Many user groups seek climate and climate-derived data today: legislators and activists, financial institutions analysing risk and insurance, academics and researchers modeling climate and weather, entrepreneurs and investors, professionals seeking jobs and service opportunities, consumers seeking sustainable goods and services, and climate impacted communities seeking recovery and resilience. These cohorts of users maintain different lexica reflecting their needs and experience, and usually have certain cohort-specific key variables and metrics such as Essential Variables in climate

science (WMO). To make search efficient, these lexica and variables must be normalized across cohorts.

A set of high-level climate-action tasks, which is directly or indirectly addressed by the above user groups, includes:

A. **Assessment of the current state of the environment** – including land, ocean, atmosphere, biosphere, hydrosphere - taken separately or combined and averaged over an appropriate time window.

B. **Model prediction of future changes of the environment**, using a suitably selected time horizon.

C. **Quantification of interactions and correlations** between geophysical, industrial, financial, or political sub-systems.

D. **Evaluation of the environmental impact** of specific facilities, industrial practices, investment decisions, or international policies.

Global and dynamical monitoring of atmosphere and planetary surface is performed in a number of international earth observation projects such as Copernicus (European Commission) and EOS-DIS (NASA) and generate a continuous stream of high dimensional climate-related datasets. These are complemented by a large number of data collections from ground-based observations (for example at NOAA). Socioeconomic data is collected and distributed by a range of inter-governmental institutions (e.g. OECD) but also by non-profit organizations (e.g. Max Roser et al).

All tasks can be specified and aggregated at the regional, national, or planetary scale, thus adding another dimension to the analysis. Illustrative examples of practical synthesized tasks, as well as the integration of datasets, can be found in the volunteer work in (Catalyst Cooperative, 2020) focusing on mitigating climate change and improving electric utility regulation in the United States.

Machine Learning (ML) has a proven track record in the analysis and inference on complex non-linear multiscale systems, and is seen as an effective tool to address the climate change problem (Rolnick & et al., 2019). We suggest that there is another application for ML techniques in the field: It can be the pivotal tool to mediate the accelerated access to the existing and emerging array of heterogeneous datasets, which is a prerequisite for rapid development of efficient analysis and inference tools to enable climate action on a truly global scale.

3 Sketch of a Two-Step Solution

In the first step, a Search Engine for Climate-Action Data Discovery is proposed. It is expected to ingest and index datasets that have been compiled, reviewed and released by expert communities or working groups. The search engine features include full text search, semantic search on metadata, and data discovery in numerical data related to above mentioned climate-action tasks.

The following extensions over the State of the Art are suggested:

- Metadata taxonomy for climate action tasks is developed to concisely cover the requirements of the climate action tasks identified above. To this end, we propose to combine metadata of several existing public repositories (e.g. WRI), existing taxonomies and ontologies related to environmental and sustainable development goals (L.J. McGibbney; UNEP), new metadata derived from task definitions, and align the resulting metadata set with knowledge domain-specific Essential Variables (Special Issue, 2020; WMO).
- Partial imputation of missing metadata will take advantage of the restricted context of climate-action tasks definitions (P. Cudré-Mauroux et al, 2008).
- Data discovery based on numerical signatures will enable discovery in numerical data space (R.C. Fernandez et al, 2018)

Metadata curation is supported by the creation of a collaborative network of data owners, building on existing efforts in the scientific community (D.E. Jensen et al, 2020b). Implementation and hosting can be realized by utilizing the support and digital infrastructure of established web-scale free-of-charge services or by setting up a new pro-bono service run by a non-profit foundation.

UN supported multi-stakeholder processes aim to facilitate a global data infrastructure based around data availability, quality and interoperability (D.E. Jensen et al, 2020b). In this context, a Search Engine for Climate-Action Data Discovery can be seen as a contribution to rapid prototyping and solving technical questions pertinent to ML tasks. In turn, successful ML prototypes can then be shared widely to provide essential services to the broader ecosystem of socioeconomic approaches and solutions in climate action.

In the second step, the search engine solution will be extended into a data catalog as a tool to collaboratively sort and organize datasets to shorten *time to value* for data users. Self-organized interest or expert groups dedicated to specific climate-action tasks are able to curate datasets that have been identified and annotated as suitable for the selected tasks (M.-A. Sicilia et al, 2017). Illustrative examples for a data catalog structure relevant for this proposal can be found at (CERN).

4 ACKNOWLEDGEMENTS

We would like to acknowledge the team at `https://www.climatechange.ai` and David E. Jensen who brought attention to many of the issues cited here and provided important online resources during preparation of this contribution.

REFERENCES

Catalyst Cooperative. , 2020. URL `https://catalyst.coop/`.

CERN. *ZENODO*. URL `https://zenodo.org`.

D.E. Jensen et al. The promise and peril of a digital ecosystem for the planet. *Medium*, 2020a. URL `https://medium.com/@davidedjensen_99356/building-a-digital-ecosystem-for-the-planet-557c41225dc2`.

D.E. Jensen et al. Annex 1: Priorities and processes in 2020 for a digital ecosystem for the planet. *Medium*, 2020b. URL `https://medium.com/@davidedjensen_99356/annex-processes-and-priorities-in-2020-for-a-digital-ecosystem-884b09cb8cc7`.

European Commission. *Copernicus: European Union's Earth Observation Programme*. URL `https://www.copernicus.eu/en`.

IPCC. *Special Report: Global Warming of 1.5°C*, 2018. URL `https://www.ipcc.ch/sr15/`.

L.J. McGibbney. *Semantic Web for Earth and Environmental Terminology (SWEET) Ontologies*. URL `https://github.com/ESIPFed/sweet/`.

M.-A. Sicilia et al. Community curation in open dataset repositories: Insights from zenodo. *Procedia Computer Science*, 106, 2017.

Max Roser et al. *Our World in Data*. URL `https://ourworldindata.org/`.

NASA. *NASA's Earth Observing System Data and Information System (EOSDIS)*. URL `https://worldview.earthdata.nasa.gov/`.

NOAA. *NOAA data access*. URL `https://www.nodc.noaa.gov/access`.

OECD. *OECD Data*. URL `https://data.oecd.org`.

P. Cudré-Mauroux et al. Picshark: mitigating metadata scarcity through large-scale p2p collaboration. *The VLDB Journal*, 17, 2008.

R.C. Fernandez et al. Aurum: A data discovery system. *IEEE 34th International Conference on Data Engineering (ICDE)*, 2018.

D. Rolnick and P. L. Donti et al. Tackling climate change with machine learning. *arXiv*, pp. 1906.05433, 2019.

Special Issue. The essential variables for sustainability. *Int. J. Digital Earth*, 13, 2020.

UNEP. *Sustainable Development Goals Interface Ontology (SDGIO)*. URL https://github.com/SDG-InterfaceOntology/sdgio.

WMO. *Essential Climate Variables*. URL https://public.wmo.int/en/programmes/global-climate-observing-system/essential-climate-variables.

WRI. *Resourcewatch*. URL https://resourcewatch.org.