Charter of Working Group Data Management for Artificial Intelligence and Machine learning

Version 1.0 of 28. 11. 2023, by Jiří Josef Šivic in collaboration with the members of the group.

1 Introduction

The Czech Republic has world-class AI research. Examples include computer vision, robotics, natural language, speech processing, and automatic reasoning. For instance, computer vision at CVUT is Top 5 in Europe according to CSRankigs; VUT is among the top 5 most influential organizations in speech processing; robotics at CVUT is top 6 in Europe. Three out of the top five (worldwide) machine reasoning systems are co-developed at CVUT. In all these areas, large-scale training data and benchmarking datasets are of utmost importance for scientific progress. The Czech AI community is involved in a number of such dataset collection and benchmarking activities.

There is an interest from several institutions: CAS (E. Pelikan), MFF UK (J. Hajic), CVUT CIIRC (J. Sivic, V. Marik) and FEL (J. Matas, T. Svoboda), VSB (J. Martinovic), VUT (P. Zemcik, J. Cernocky), ZCU (J. Psutka), and MU (S. Mazurenko) to build data management repositories for research in artificial intelligence. In particular, the interest is in the following "core AI areas": computer vision, natural language processing, automated reasoning, speech processing, robotics, and simulation as well as interdisciplinary areas such as analysis of complex systems, medical imaging, and machine learning for Earth, Geo-, and Life sciences. An important aspect is the proximity and fast access of the data to the appropriate computing power, e.g. Karolina computing cluster.

Data is the heart of artificial intelligence, serving as the foundation upon which AI models are built and trained. It includes vast amounts of information, from text and images to numbers and sensors, enabling AI systems to learn and make human-like decisions. However, all these data must meet certain levels of quality, normalisation, and fairness due to negative impacts on outcomes and decisions. In summary, fair and standardized data in AI is essential for ensuring that AI systems are ethical, unbiased, and trustworthy. It benefits not only individuals and society as a whole but also organizations that develop and deploy AI technologies.

Fair data in AI is crucial for several reasons:

- **Avoiding Bias and Discrimination**: Fair data helps prevent bias and discrimination in AI systems.
- **Ethical Considerations**: Ensuring fairness in AI aligns with ethical principles. Fair data practices help reduce the potential harm that AI systems can cause.
- **User Experience**: Fair data can lead to better user experiences.
- **Long-Term Viability**: Fair data practices are essential for the long-term viability of AI systems.
- **Reputation and Trust**: Fair data practices enhance the reputation of organizations and build trust with customers, users, and the public.

Co-funded by the European Union

MINISTRY OF EDUCATION, YOUTH AND SPORTS

European Open Science Cloud
Czech Republic (EOSC-CZ)
Masaryk University, Institute of Computer Science
Mailing address: Sumavská 33, 602 00 Brno
info@eosc.cz; www.eosc.cz

2 Objectives

Main task:

To create data infrastructure to support large-scale data together with supercomputing capabilities to develop a new generation of advanced machine learning models opening-up new data analysis applications in the exascale era.

Sub -tasks:

- Design and implementation of data standards in the field of artificial intelligence and machine learning (to meet European legal and ethical standards).
- FAIRification and sharing of own data in the NRP environment.
- Design of metadata structure for unified storage of AI/ML data.
- Maintaining the community.
- Estimate current and future needs in the field of data storage and data management for the AI/ML community in the Czech Republic.
- Connections to European repositories / initiatives: BDVA/DAIRO, CLAIRE, ELLIS, EUDAT and BigScience.
- Outputs and their applications.

3 Outputs and their applications

- FAIR standards for storing AI/ML data and their implementation.

4 Membership, expected members, mode of operation

The group includes institutional experts in AI and ML across the spectrum of AI/ML tasks. PS membership is voluntary and may include experts outside the AI domain. The primary target group is the academic and research community of the Czech Republic. Active users who contribute to the functioning of the group with their experience in AI/ML data management work on the basis of a "living" document, where all use cases are shared among all its members. The document is further structured into AI/ML categories that function as its subgroups (e.g. Computer Vision, NLP).

European Open Science Cloud
Czech Republic (EOSC-CZ)
Masaryk University, Institute of Computer Science
Mailing address: Sumavská 33, 602 00 Brno
info@eosc.cz; www.eosc.cz