CHIST-ERA Projects Seminar 2019

Open Science Session

Supporting Data Sharing and Experiment Reproducibility

Wednesday, April 3rd, 2019 – 8:30-12:30

Intercontinental Hotel, Bucharest, Romania

Context and objectives: Open Science refers to practices of open sharing of information (publications, data, software…) contributing to the quality and economic efficiency of scientific research. These practices are at the heart of scientific methodology but can be impeded in practice by various constraints and need to be specifically supported. Over the last few years, this need has been increasingly acknowledged. CHIST-ERA, as many actors supporting scientific research, is developing a proactive support to Open Science.

The present session is focused on two interrelated aspects of Open Science that are particularly relevant to data-intensive research, which is a focus of CHIST-ERA:

- **Open Research Data & Data Sharing**: Properly annotated and curated data can require significant efforts. When this is the case, it is important for economic reasons to mutualise these efforts and make the data widely available. However, this requires an extra effort from the data producer.

- **Open Evaluation Methodology & Experiment Reproducibility**: Objective evaluation of models learned from data requires, in order to ensure that the results are comparable and unbiased, common experimental protocols and joint experiment campaigns on unseen data. This in turn requires a specific coordination effort and a trusted third party.

In both cases, the extra efforts are not done so often in practice even though they would be highly beneficial for the community. Finding a trusted third party is also not obvious. Specific support is thus needed.

Relying on dedicated organisations implementing such mutualised or joint activities for the benefit of the community can significantly enhance the efficiency and impact of research projects. When such organisations exist, CHIST-ERA projects can benefit from them. They can be associated to projects in various ways, ranging from informal exchange of information to direct participation with funding. One objective of the session is to raise awareness about the interest of such collaborations.

Furthermore, these dedicated organisations need sustainable funding in order to ensure their mission at the service of the community at a satisfactory level. However, despite their importance, their institutional role and economic model are understudied and not often reproduced, thus resulting in the lack of such services in many sectors of research. A second objective of the session is thus to raise awareness about the need to develop dedicated entities with the public mission to support Open Science.

Agenda

Time	Duration	Item
08:30	0h30	Registration & welcome coffee
		Introduction
09:00	0h20	Introduction to the Workshop & Presentation of the CHIST-ERA Support to Open Science *Edouard Geoffrois, Agence nationale de la recherche (ANR), CHIST-ERA Open Science Task Leader*
		Keynotes – Open Evaluation Methodology
09:20	0h20	Open Evaluation of Human Language Technologies in US Research Programs *Mark Liberman, Linguistic Data Consortium (LDC), Director*
09:40	0h20	LNE Activities for Open Evaluation of Artificial Intelligence *Guillaume Avrin, Laboratoire national de métrologie et d'essais (LNE), AI evaluation activity coordinator*
		Keynotes – Open Research Data
10:00	0h20	Centralised Production, Curation and Distribution of Open Research Data: The LDC experience *Mark Liberman, Linguistic Data Consortium (LDC), Director*
10:20	0h20	Open Science in the Climate Sciences and the role of DKRZ *Karsten Peters, Deutsche Klimarechenzentrum (DKRZ), Senior policy advisor*
10:40	0h20	Developing the European Open Science Cloud: how DANS supports FAIR and Open Science *Elly Dijk, Data Archiving and Networked Services (DANS), Senior policy advisor*
11:00	0h30	Coffee break
		Short Presentations from CHIST-ERA projects
11:30	0h05	Open Benchmarking Activities in the CHIST-ERA ALLIES Project *Anthony Larcher, Laboratoire d'informatique de l'Université du Mans (LIUM), ALLIES project coordinator* *Olivier Galibert, Laboratoire national de métrologie et d'essais (LNE), ALLIES project PI*
11:35	0h05	Open Benchmarking Activities in the CHIST-ERA CORSMAL Project *Andrea Cavallaro, Queen Mary University of London, CORSMAL project coordinator*
11:40	0h05	The UK Reproducibility Network *Etienne Roesch, University of Reading, Cocoon project coordinator*
11:45	0h05	Negotiating Interdisciplinarity in and through the PROVIDE DH Open Science Strategy Planning Process *Jennifer Edmond, Trinity College Dublin (TCD), PROVIDE DH Dissemination WP leader* *Roberto Therón, University of Salamanca, PROVIDE DH project coordinator*
		Round Table Discussion
11:50	0h30	Questions & Answers *All speakers at round table*
		Conclusion
12:20	0h10	Conclusions and Perspectives *Edouard Geoffrois, Agence nationale de la recherche (ANR), CHIST-ERA Open Science Task Leader*
12:30	1h00	End of session and lunch

Keynote Abstracts and Speaker Biographies

Open Evaluation of Human Language Technologies in US Research Programs
Mark Liberman, Linguistic Data Consortium (LDC), Director

Mark Liberman began his career at AT&T Bell Labs in 1975, and moved in 1990 to the University of Pennsylvania, where he is a professor in the departments of Linguistics and Computer and Information Science. He has participated in DARPA's Human Language Technology programs since the mid 1980s. His involvement in the open data movement began with the ACL Data Collection Initiative in the 1980s, and continued with the provision of shared data for DARPA and other HLT programs, and the founding of the Linguistic Data Consortium in 1992. His current research activities include methods for easier development of resources for languages that lack them, and the application of data-intensive linguistic analysis to clinical, educational, and legal issues.

Abstract

"Human Language Technology" (HLT) is a term coined to cover capabilities such as speech recognition, machine translation, and document understanding – systems that deal with inputs and outputs of human speech and language, or with representations of their information content. Systems based on these technologies, an impractical dream 50 years ago, are in wide and spreading use today. There are two obvious reasons that these capabilities are now an everyday reality: powerful cheap computing devices and ubiquitous digital networking. But a third reason deserves mention: the "Common Task" method, a research management technique developed in the 1980s and applied increasingly widely since then. Begun as a way to support long-term research and development in HLT, this approach is arguably responsible for the success of all modern "artificial intelligence" research. This talk will tell the story of this innovation, and suggest some lessons for the Open Science movement as well.

LNE Activities for Open Evaluation of Artificial Intelligence
Guillaume Avrin, Laboratoire national de métrologie et d'essais (LNE), Artificial Intelligence evaluation activity coordinator

Guillaume AVRIN has been a Technical Manager at LNE since 2017. Along with running evaluations and competitions in AI and robotics, he conducts research on evaluation protocol (metrics, corpora, etc.), develops evaluation tools and participates in international standardization activities (UNM81 and ISO/IEC JTC1/SC42). He is currently in charge of the organization of the ROSE Challenge in agricultural robotics. He also supervises the evaluations carried out by LNE on autonomous vehicles, companion robots, lifelong learning translation and diarization systems and on voice comparison in the forensic field.

Abstract

In recent decades, public and private funding for the development of intelligent systems has grown exponentially. However, the technological maturity and market of these systems are not evolving as quickly as expected. The insufficient performance of intelligent solutions and our difficulty in measuring these performances in a rigorous and objective manner contribute to explain this gap. In response to this evaluation need, the National Laboratory of Metrology and Testing (LNE) set up a team in 2008 to evaluate artificial intelligence (AI) systems, initially for automatic information processing and more recently for robotic systems.

LNE thus acts as a trusted third party for the evaluation of AI systems in research projects that aim to advance the state of the art in an AI field, often referred to as evaluation campaigns or "challenges". These multi-year projects offer a common evaluation framework for teams developing competing approaches. They enable systems' behaviour to be openly benchmarked in a comparable and reproducible way, their performance to be objectively measured, their technological maturity to be transparently assessed and the state of the art in the field to be updated. For technology developers, these campaigns are an essential means of motivation, a factor of collective emulation and a catalyst for innovation, allowing an increase in TRL. For funding agencies, the results of these campaigns are reliable indicators of the quality of research and technological progress, and enable to assess the impact of the investments made. In addition, for the higher TRL range, LNE performs qualification tests and certification actions to demonstrate that an intelligent system complies with a certain set of specifications, and thus contributes to a fair functioning of the AI market.

Centralised Production, Curation and Distribution of Open Research Data: The LDC experience

Mark Liberman, Linguistic Data Consortium (LDC), Director

Mark Liberman began his career at AT&T Bell Labs in 1975, and moved in 1990 to the University of Pennsylvania, where he is a professor in the departments of Linguistics and Computer and Information Science. He has participated in DARPA's Human Language Technology programs since the mid 1980s. His involvement in the open data movement began with the ACL Data Collection Initiative in the 1980s, and continued with the provision of shared data for DARPA and other HLT programs, and the founding of the Linguistic Data Consortium in 1992. His current research activities include methods for easier development of resources for languages that lack them, and the application of data-intensive linguistic analysis to clinical, educational, and legal issues.

Abstract

The modern model for AI research depends on effective integration of task design, shared datasets, and quantitative evaluation methods. In particular, large amounts of appropriate training material must be collected or created before research on a new task can begin. The provision of this research data must be responsive – because the dataset design must be well matched to the needs of the task; agile – because the dataset must be delivered to researchers on time, even if the research plan evolves; and efficient -- because dataset costs can be a substantial fraction of the overall program budget. The resulting datasets continue to be used for benchmarking purposes after the end of the original projects, and are often also re-used in new and unexpected ways, so reliable long-term public availability is important. The benefits of publication also often hold for datasets produced by research outside of such "common task" programs.

The Linguistic Data Consortium, an activity of the University of Pennsylvania, founded in 1992 with seed money from DARPA, now has 881 datasets in its catalog, and has distributed over 180,000 dataset copies to nearly 6000 companies, universities, and government research laboratories in 92 countries. This talk will sketch its history, draw some lessons from its experience, and discuss challenges for the future.

Centralised Production, Curation and Distribution of Open Research Data: The DKRZ Experience

Karsten Peters, Deutsche Klimarechenzentrum (DKRZ), Senior policy advisor

Karsten Peters works at DKRZ in the data management department, where he caters for the data management needs of Earth System Scientists and performs trainings and workshops in the area of data management. He is also responsible for the communication, outreach and development of DKRZ data management services. Before joining DKRZ, he worked at the Max Planck Institute for Meteorology in Hamburg and at Australian Research Council's Centre of Excellence for Climate System Science in Melbourne, where he used and actively developed climate and atmospheric models.

Abstract

Over the past three decades, global sharing, comparison and evaluation of climate research data has formed the backbone of the progress in Earth System Sciences. With its compute and data services, the German Climate Computing Center (Deutsches Klimarechenzentrum, DKRZ) has since its inauguration in 1987 shaped the German and international landscape of an entire community. Among others, DKRZ hosts the World Data Center for Climate (WDCC), represents a core node of the Earth System Grid Federation (ESGF, an international infrastructure of federated data nodes enabling the global distribution of Petascale climate data) and shapes the development of data standardisation and publication practices. The presentation will discuss the evolution of DKRZ, its sustainable source of funding as well as the data sharing culture of the Earth System Science community as a whole.

Developing the European Open Science Cloud: how DANS supports FAIR and Open Science

Elly Dijk, Data Archiving and Networked Services (DANS), Senior policy advisor

A human geographer by training, Elly Dijk is senior policy advisor at Data Archiving and Networked Services (DANS). DANS is the Netherlands institute for permanent access to digital research resources. At DANS she is coordinator of the national scholarly open access portal NARCIS (narcis.nl). She is currently project leader of the Horizon 2020 projects EOSCpilot, OpenAIRE Advance and EOSC-hub at DANS. She is, together with a colleague of the library of the Technical University Delft, the National Open Access Desk (NOAD) for OpenAIRE. DANS is regional coordinator Research Data Management in this project. In the Netherlands, she is member of the National Platform Open Science, and she is part of the editorial team of the website Openaccess.nl.

Abstract

Not only the European Commission (EC) promotes Open Science, but also international organisations like Science Europe and ALLEA (ALL European Academies) are in favour of Open Science. The EC is developing the European Open Science Cloud (EOSC) that will facilitate the sharing and re-use of scientific data and services, across disciplinary and state boundaries. It will be made mandatory for researchers to make data from EU-funded research available under the FAIR data principles, i.e. data must be Findable, Accessible, Interoperable and Reusable.

In the Netherlands, the institute DANS (Data Archiving and Networked Services) is involved in a number of Horizon 2020 projects to support the development of the EOSC: EOSCpilot, EOSC-hub, OpenAIRE Advance, FREYA and FAIRsFAIR. With her services (DataverseNL, EASY, NARCIS, and Research Data Management Training) DANS supports researchers and data librarians to become part of the European Open Science Cloud.