



FAIR & AI Symposium Highlights: A Community Shaping the Future of Trustworthy Research Data and AI

Stefan Reichmann, Birgit Söser, Ilire Hasani-Mavriqi

The FAIR & AI Symposium, organized under the scope of Cluster Research and Data, successfully brought together a vibrant community of researchers, data stewards, infrastructure specialists, and policy experts to explore one of today's most pressing intersections: the evolving relationship between FAIR research data and artificial intelligence.

Held at Graz University of Technology, the event offered a rich program that sparked lively discussions on how data management and AI development can advance together in a responsible and sustainable way.

Key themes and insights

Throughout the symposium, participants reflected on the pivotal question:

Are the FAIR principles still sufficient in an era where AI increasingly guides how we create, manage, and reuse data?

The sessions made clear that while FAIR remains a strong foundation, AI brings both powerful opportunities and new responsibilities. Automated metadata generation, semantic enrichment, and improved data discoverability were highlighted as significant enablers for FAIRification. At the same time, speakers emphasized that challenges such as transparency, bias detection, accountability, and ethical decision-making cannot be delegated to machines alone.

A program that sparked dialogue and collaboration

Participants engaged in a dynamic mix of keynote addresses, expert presentations, lightning talks, and hands-on discussions. Breakout groups explored concrete use cases at the intersection of FAIR data and AI, while panel debates created space for critical reflection on Austria's and Europe's evolving research data and AI infrastructures.

In a warm welcome to the participants, Ilire Hasani-Mavriqi, head of the RDM Team at Graz University of Technology and the team behind the workshop, stressed the timeliness of the workshop, and the questions to be tackled: Are the FAIR principles enough to ground Open Science in the age of AI? How can AI support FAIR (e.g. automatic metadata creation, enhanced enrichment)? Can machines be ethical in this regard? This was followed by a welcome address from Andrea Höglinger, Graz University of Technology Vice Rector for Research, in which she emphasized that FAIR and AI mark the future of the research ecosystem, and pointed to the emerging close cooperation at the international level under the auspices of the European Open Science Cloud. Sabine Neff-Kolassa, the coordinator of Cluster Research and Data stressed that the Cluster works to tackle



challenges in research data management, to enable open science and findability. The workshop was moderated throughout by Suvini Lai and Livia Beck (both from TU Wien).

The event started with two keynote speeches, by Jana Lasser (University of Graz), and Daniel Garijo (Universidad Politecnica de Madrid), respectively.

Jana Lasser spoke to the challenges in working with sensitive data (from and about humans) from a researcher's perspective, presenting three case studies on privacy implications of FAIR data management. The first case study revolved around using Natural Language Processing methods to substitute (paid) actors in training therapists with (trained) large language Models simulating patient interactions. Training these models, poses severe privacy risks as the personal data involved are highly sensitive and can only be partially anonymized so as not to render them useless. Using secure data processing environments for training these models proved unviable, because this approach causes high friction and large delays. Further, with technical expertise scattered across different groups, users would need to set up their own processing environments.

Second, Jana Lasser reported on the STAIRCASE survey on researcher mental health, which is a very large dataset consisting of very sensitive data that is difficult to fully anonymize. International interest in these data would further require a scalable and secure remote data access solution, which needs to be planned for before data collection. In this case, the team brought in a data protection officer from the get-go, to determine the needed levels of anonymization from the start, to accommodate the complex interplay of anonymization, research requirements, and legal/technical constraints. Still, this process is costly, and institutional willingness to bear these costs is minimal, and decisions needed for GDPR compliance may, in fact, conflict with journal open data policies.

Finally, Jana Lasser spoke to the Digital Services Act, a new piece of European legislation pertaining to data access by vetted researchers to privately held social media data, with a view towards mitigating societal risks created by amplifying extreme viewpoints, by obliging platforms to implement risk mitigation measures. This has meant that translating regulations into concrete platform design elements is an open research field at the moment. As these platforms harbour legitimate concerns regarding data reuse, applying for data access under the DSA is an extremely complex process, mandating applicants (inter alia) to specify exactly what type of data is requested, describe the necessity of data access, as well as measures undertaken with a view to data protection. Together this has created an adversarial environment for researchers, in that being granted data access requires investment in infrastructure. Formulating the required data access requests requires the collaboration of experts who deeply understand all aspects of data, to navigate the complex landscape of research requirements, the DSA, and the GDPR.

In this landscape, the required skills are clearly beyond the capabilities of individual researchers, which means that data professionals such as Data Stewards are highly sought after. Among the key skills required are anonymisation techniques, esp. As they



pertain to unstructured data. In this landscape, professional, efficient and secure (R)DM practices would constitute a real competitive advantage for RPOs.

Next, Daniel Garijo (UPM) spoke to the topic of quality in heterogeneous digital objects. Starting from a roundabout introduction to the FAIR principles, Daniel introduced the motivations for FAIR. While FAIR is often framed as an end in itself, it should actually be understood as a means towards a higher end, mainly to improve scientific credit, to acknowledge datasets as important research outputs, and (thereby) to improve reproducibility. The FAIR principles were an initiative from the FORCE11 working group in 2016 and have become fundamental for the EOSC and related initiatives, as well as having been extended to other research outputs, such as software, workflows, etc., and by introducing FAIR Assessment tools to support the implementation of the FAIR principles by producing a FAIR score. However, most FAIR Assessment Tools test for metadata completeness, which does not affect data quality (nor vice versa). When it comes to FAIR for AI, finding relevant data is actually very difficult (even when data have metadata and API access) on account of the inability to do granular data search. Since only a small fraction of datasets contain the relevant data, data need still to be manually inspected (data may be multilingual, inconsistent, with duplicate values, incomplete values, inconsistent annotations). FAIR allows for the retrieval of lots of datasets (on account of rich metadata), but retrieval is different from relevance. Further, metadata of poor quality will yield incorrect results for training AI models (FAIRness does not affect data quality).

There are various initiatives that focus on these issues, both for data as well as for research software. These include standards like the the W3C DCAT, schema.org (a web-based vocabulary to annotate the web, including metadata for datasets, to be harvested by search engines); further, CroissantML is a schema.org extension for describing ML datasets which has been used to annotate datasets from Kaggle, Hugging Face and DataVerse; finally, the RDA I-ADOPT Working Group wants to use a dedicated vocabulary to capture variables included under different names across different datasets to create homogenous representations. The CodeMeta project wants to create a concept vocabulary to standardize the exchange of software metadata, creating crosswalks between different software packages. The EOSC EVERSE project develops indicators for research software assessment. The FAIR4ML initiative aims to enable understanding how an ML model was trained.

These initiatives are great to ensure that fine-grained information on datasets can be found, and that models can be compared with respect to their performance. Still, data quality remains an issue. Detailed metadata quality requires manual validation, which constitutes a practical restriction as it is important to balance additional requirements for researchers. Daniel concluded that while FAIR is key for AI (metadata, interoperability, search, provenance), data quality is (still) an open issue. Since FAIR is a means to an end (reproducibility), this means that high-quality datasets and models should be FAIR to allow training AI models.



The keynotes were followed by three lightning talks, by Markus Stöhr (ASC), Jeanette Gorzala (Act.AI.now), and Emily Kate/Michael Feichtinger (Uni Wien), on the topics of high-performance computing, AI governance and literacy, and data stewardship and RDM, respectively.

The ACA is backed by a small consortium that includes the University of Vienna and TU Wien, runs the Austrian Supercomputing Community (ASC) — a shared national infrastructure that opens the door to high-performance computing for researchers and, where appropriate, industry. At its core are systems like VSC-5 at TU Wien, a modern CPU/GPU machine, and MUSICA, a new AI-focused platform distributed across three sites. Together with Austria’s access to Italy’s LEONARDO supercomputer, they give users a flexible and powerful environment to experiment, scale up ideas, and tackle data-intensive challenges.

MUSICA in particular marks a shift toward AI-driven workloads, with large-memory nodes and enough capacity to support ambitious training runs and complex models — a clear step beyond earlier, CPU-only systems. VSC-5 complements this by offering a more energy-efficient setup for a wide range of established workloads. Looking ahead, Austria plans to formally join the EuroHPC ecosystem in 2026, becoming part of a Europe-wide network of AI Factories. The Austrian contribution, AI:AT, is coordinated by ACA and AIT and brings together ten partner institutions.

What makes this ecosystem especially attractive is not just the hardware, but how accessible it is. Through the EuroHPC portal, researchers and SMEs will be able to use the infrastructure free of charge up to a generous quota of GPU hours. Alongside this, EuroCC provides hands-on support: from project advice and consulting to training courses and educational programmes. Taken together, the infrastructure, expertise, and training opportunities create real room for experimentation, skill-building, and growth in both HPC and AI.

Jeannette Gonzala, a legal expert and founder of Act.AI.now highlighted the growing need for AI governance and literacy as organizations struggle to implement the new AI Act amid an “AI gold rush” that has seen investment grow more than thirtyfold since 2019. While public sentiment mixes enthusiasm with anxiety, even business leaders often hesitate to trust AI systems, which brings significant risks alongside opportunities: GDPR concerns, bias and discrimination, IP issues, hallucinations, explainability gaps, cybersecurity threats, misuse, sustainability, and global dependencies. Many AI companies still emerge from research environments without knowing how to navigate legal obligations—OpenAI’s trajectory and cases such as Samsung’s leakage of sensitive information to ChatGPT underscore what can go wrong. Poor data can lead to misleading models, chatbots hallucinate in ways that cost companies money and credibility, and academia faces growing problems with error-ridden AI-generated papers. Ensuring trustworthy AI therefore requires governance and legal guardrails: compliance with existing laws, alignment with EU ethical principles, and technical and social robustness, expressed through seven core principles. The EU AI Act, approved in 2021 and now in force, introduces a risk-based, use-centric regulatory framework in which only about 10–15% of applications fall into the high-risk category. Its core message is procedural: define what



you want to build, assess the risks, and implement measures to mitigate them—remembering that the most innovative applications usually come with the highest stakes.

Emily Kate reflected on two years of experience in building meaningful support for researchers, noting that while goals are clear, practical guidance for achieving them is often lacking. FAIR Data Austria initiated this work through an RDM working group and a pilot phase with a small team of coordinators and embedded data stewards, followed by a growth phase in 2024 and, since June 2025, a new structure integrating data stewards directly into the RDM team within Research and Publication Services. Along the way, several challenges surfaced: fragmented infrastructure and missing key tools, heterogeneous and often cumbersome workflows, and human factors such as lower-than-expected RDM knowledge and limited researcher attention—despite generally strong goodwill. These experiences raise two deceptively simple but fundamental questions: what does “good” research data management actually look like in practice, and how can institutions persuade researchers to accept help, secure the necessary resources, and build sustainable, well-structured support processes?

A vision for better RDM centres on a few concrete, pragmatic practices: using stable and shared storage instead of ad-hoc solutions, documenting decisions and workflows, organising data consistently, depositing published datasets in trusted repositories, strengthening PhD training, and deepening researchers’ engagement with available services. Making a real impact, however, requires more than guidance—it means removing structural barriers, offering rapid and targeted support, aiming for realistic incremental improvements, actively cultivating engagement, and embedding change-management principles. A useful rule of thumb is to treat anything resolved within two emails as a simple request and anything beyond that as a project requiring structured coordination. Persistent challenges include aligning solutions with limited resources, bridging the gap between immediate needs and slower institutional rollouts, juggling competing commitments, and winning attention in an already saturated environment. To sustain momentum, teams must clearly define goals and success criteria, secure adequate resources, pace their efforts like a marathon rather than a sprint, and continuously help data stewards upskill in a rapidly evolving open-science landscape.

The afternoon concluded with a networking session that encouraged participants to exchange experiences, identify shared challenges, and initiate new collaborations.

Take away: A community moving forward

One of the most resounding takeaways from the symposium was the shared recognition that **trustworthy AI and FAIR data must evolve together**. Ensuring that AI-driven research workflows remain transparent, reusable, and ethically grounded will require ongoing interdisciplinary cooperation — across research domains, institutions, and infrastructures.

We thank all speakers, contributors, and participants for their valuable input and strong engagement. The conversations sparked at this symposium mark an important step toward shaping a responsible, FAIR, and AI-ready research landscape.