

Federated Research Data Management in LEXIS

J. Munke^{1}, M. Hayek^{1*}, M. Golasowski², C. Koch-Hofer³, R. J. García-Hernández¹, M. Levrier³, S. Hachinger¹, A. Scionti⁴, D. Magarielli⁵, T. Goubier⁶, A. Parodi⁷, P. Harsh⁸, S. Murphy⁹, F. Apopei¹⁰, O. Terzo⁴, J. Martinovic²*

¹Leibniz Supercomputing Centre (LRZ) - Bavarian Academy of Sciences & Humanities, Garching, Germany

²IT4Innovations, VŠB - Technical University of Ostrava, Ostrava-Poruba, Czech Republic

³ATOS, Paris, France

⁴LINKS Foundation - Advanced Computing and Applications, Torino, Italy

⁵Avio Aero, Torino, Italy

⁶CEA - List, Gif-sur-Yvette, France

⁷CIMA Research Foundation, Savona, Italy

⁸Cyclops Labs GmbH, Winterthur, Switzerland

⁹ZHAW School of Engineering, Winterthur, Switzerland

¹⁰TESEO, Torino, Italy

johannes.munke@lrz.de, mohamad.hayek@lrz.de

ABSTRACT

Full Presentation: short – 15 mins

Within the LEXIS project (Large-scale EXecution for Industry & Society), a platform for optimized execution of distributed Cloud-HPC workflows for simulation and analysis of Big Data is developed. A user-friendly web portal, as a unique entry point, will provide access to data as well as workflow-handling and remote visualization functionality. The LEXIS platform federates European computing and data centers. It uses advanced orchestration solutions (Bull Ystia Orchestrator, based on TOSCA and Alien4Cloud), the High-End Application Execution Middleware (HEAppE), and new hardware systems (e.g. Burst Buffers with fast GPU- and FPGA-based data reprocessing). LEXIS heavily relies on its data-storage backend – an EUDAT-based “Distributed Data Infrastructure” for flexible, federated data and metadata management within workflows. All LEXIS systems are co-developed with three application Pilots, which represent demanding HPC/Cloud-Computing use cases in Industry (SMEs) and Science: i) Simulations of complex turbo- machinery and gearbox systems in Aeronautics, ii) Earthquake and Tsunami simulations which are accelerated to enable accurate real-time analysis, and iii) Weather and Climate simulations where massive amounts of in situ data are assimilated to improve forecasts.

This contribution focuses on the LEXIS Distributed Data Infrastructure (DDI). With EUDAT-B2SAFE and thus iRODS (the Integrated Rule-Oriented Data System) as a basis, a data grid with a unified view on LEXIS datasets has been realized. The core of our iRODS federation comprises the supercomputing centers IT4I (CZE) and LRZ (DEU). It can be extended to include further partners at any time. Leveraging EUDAT-B2SHARE, data in the DDI can be accessed via GridFTP. This follows our policy of adapting European federated computing and data-handling concepts for building a specialized, but open simulation-workflow environment, with focus on Cloud-HPC-Big Data convergence. This may be extended on the computing side by federating LEXIS computing resources e.g. with the EGI federated cloud. In general, we aim at seamlessly immersing LEXIS in the European computing and data landscape, with a focus on EOSC partners and the Big Data Value Association (BDVA).

On a technical level, the DDI features significant adaptations to the LEXIS ecosystem with respect to a pure iRODS or EUDAT-B2SAFE system. With an iRODS OpenID plugin extended to handle large tokens, it connects to the Keycloak-based LEXIS AAI (or any AAI compatible with OpenID-Connect and SAML). Using a redundant setup and appropriate policies (iRODS rules), data safety and quick data availability will be ensured. As a further feature in LEXIS, innovative REST APIs on top of the DDI ensure a smooth interaction with the LEXIS orchestration layer and portal. They provide, for example, the (meta-)data catalogue of the DDI, and offer an endpoint to trigger asynchronously-executed data transfers to/from the HPC and Cloud systems involved in LEXIS.

Besides performance-oriented functionality, the DDI features fundamental capabilities for Research Data Management (RDM) following the FAIR principles (“Findable, Accessible, Interoperable, Reusable”). Metadata is kept with the data, and PIDs will be acquired via EUDAT-B2HANDLE. This serves to disseminate open LEXIS (meta-)data to search facilities (B2FIND, BASE, web search engines, etc.), and thus to contribute to general data sharing and re-use.