Shared Challenges, Different Architectures: Comparing EOSC-CZ and CLARIN Infrastructure Approaches

Ondřej Košarko
Charles University
Faculty of Mathematics and Physics
Prague, Czech Republic
kosarko@ufal.mff.cuni.cz

Pavel Straňák
Charles University
Faculty of Mathematics and Physics
Prague, Czech Republic
stranak@ufal.mff.cuni.cz

Abstract

In 2024 Czechia started national implementation of the European Open Science Cloud: EOSC-CZ project centred around data repositories. CLARIN and EOSC-CZ share foundational infrastructure challenges – repository onboarding, metadata interoperability, and federated authentication and authorisation – but have developed divergent solutions. This paper examines three themes where these divergences are most instructive for the CLARIN community. The National Repository Platform (NRP), the core part of EOSC-CZ, offers shared hardware and curated software stacks – including CLARIN-DSpace – as an alternative to fully independent CLARIN centres. The Czech Core Metadata Model takes a mandatory, linked-data baseline approach, contrasting with CMDI's community-driven flexibility; its national metadata aggregator is directly analogous to the VLO. On authentication and authorisation, participation in two infrastructures simultaneously introduces specific challenges: identifier continuity, entitlement propagation, and cross-community resource access. These challenges are not unique to the Czech context and will recur wherever CLARIN centres integrate with (national) EOSC nodes.

1 Introduction

This paper reports on our experience as a CLARIN centre that engages with the Czech infrastructure built under the EOSC umbrella. The Czech implementation of EOSC is coordinated through several national projects. The EOSC-CZ project[1] establishes and governs the National Data Infrastructure (NDI) and its shared components, including the National Metadata Directory (NMD). The CARDS project provides methodological support for metadata and persistent identifiers. The technical backbone is the National Repository Platform (NRP): a shared hardware, software, and policy environment designed to lower the barrier to setting up new repositories and running them.

We compare repository and platform architecture (Section 2), metadata interoperability (Section 3), and authentication and authorisation (Section 4). For each theme, we identify the CLARIN counterpart and examine where design choices diverge, what are the trade-offs, and what can we learn from this in CLARIN.

2 Repository and Platform Architecture

CLARIN is a distributed network of centres. Each centre builds and operates its own repository infrastructure independently, which preserves architectural autonomy, but places the full burden of setup, maintenance, and operation on the centre. The NRP[2] operates at national scale and takes a different approach. Rather than prescribing rules and leaving implementation free, like CLARIN, or providing a single full software stack in the other extreme, it provides shared hardware – S3-compatible object storage and a Kubernetes cluster to run containerised applications – as a common substrate and offers three supported repository platforms to start a new repository.

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/

[1] https://www.eosc.cz/en

[2] https://www.eosc.cz/projekty/narodni-repozitarova-platforma-pro-vyzkumna-data-nrp/nrp

The three supported software stacks – InvenioRDM, CLARIN-DSpace, and ASEP/ARL – reflect deliberate heterogeneity, but compared to CLARIN network, the most distinctive feature is what NRP repositories gain by sharing the underlying platform: geographically redundant storage, a dedicated expert team, and common policy defaults. Storage is distributed across multiple sites – currently two, with a third planned – each with several replicas. This level of geographic redundancy is almost impossible for a single-repository operation to sustain. Supporting only a small set of systems also yields economies of scale, and policy defaults developed at the platform level are inherited by new repositories, further accelerating setup.

A community can engage with the NRP in more than one mode (Matyska & EOSC-CZ, CARDS, and NRP Projects, 2026). In the first mode, a new repository is provisioned directly as a tenant on the shared NRP infrastructure, using one of the three supported software stacks as its basis. In the second mode, repositories built on alternative software gain access to the S3 and Kubernetes platform while retaining responsibility for their own deployment. In the third mode, an existing repository joins by exposing the CCMM metadata (see Section 3.1) to the NMD via an OAI-PMH endpoint and connecting to the EOSC-CZ AAI without using the shared hardware. Exposing metadata and joining the EOSC-CZ AAI are requirements in all three modes.

The LINDAT/CLARIAH-CZ repository is joining in the second mode: nothing changes in how we deploy the repository, and we additionally gain access to the NRP S3. Since we develop the CLARIN-DSpace stack, we also have experience with the first mode. Among other things, CoreTrustSeal certification covers user communities, storage and backup procedures, so joining the NRP touches those parts of our documentation.

For CLARIN centres running CLARIN-DSpace, the NRP's containerisation artefacts – Kubernetes descriptors and S3 backend integration – reduce the operational cost of modernising existing deployments, regardless of NRP membership.

3 Metadata Landscape

Metadata interoperability is a key challenge for research infrastructures spanning heterogeneous repositories. CLARIN addresses this through CMDI (Component Metadata Infrastructure (Windhouwer & Goosen, 2022)) and the CLARIN Component Registry, which allow communities to define, share, and reuse XML-based metadata components, composing them into profiles suited to specific resource types. EOSC-CZ approaches the problem differently, relying throughout on semantic web and linked-data technologies: the Czech Core Metadata Model (CCMM) serves as the shared metadata baseline for all NDI repositories, composed of a semantic vocabulary and an application profile,[3] with provision for further extensions by repositories via profiling.

3.1 Czech Core Metadata Model (CCMM)

CCMM (The Metadata Unit of the National Library of Technology, 2025) is the common metadata standard for all repositories within the NDI. Its development began with a systematic analysis of international metadata schemas (Černohlávková et al., 2023), which concluded that no single existing standard was fully adequate. Drawing on that analysis, a survey of community needs (Vyčítalová et al., 2024), and NMD interoperability targets, the EOSC-CZ Metadata Working Group derived CCMM as based primarily on DCAT-AP and DataCite, supplemented by DCAT-AP-CZ and Dublin Core.

CCMM requires IRI-based identification for all vocabulary terms, reflecting its linked-data orientation. Controlled vocabularies are maintained by the National Technical Library (NTK) via a SKOSMOS instance[4], or are inherited read-only from external authorities such as COAR and IANA. The modelling toolchain mirrors the CLARIN one in spirit: CCMM is modelled in DataSpecer[5], a tool for managing semantic data specifications, which generates consistent physical serialisations (JSON, XML, CSV schemas, human-readable documentation, and more) from a single conceptual model. Unlike the

[3] https://www.dublincore.org/resources/glossary/application_profile/

[4] https://vocabs.ccmm.cz

[5] https://dataspecer.com/

CLARIN Component Registry, which combines a design environment with an authoritative publishing registry, DataSpecer is not a publishing platform; profile governance is documented separately[6] as part of the NRP project.

3.2 Discovery and Aggregation: NMD and VLO

The National Metadata Directory (NMD) is analogous to CLARIN's Virtual Language Observatory (VLO), which aggregates CMDI metadata from CLARIN centres and provides cross-infrastructure discovery. The NMD[7] is the discovery layer for the Czech NDI, harvesting and indexing metadata from all participating repositories. Integration with the NMD is mandatory for NRP membership. From 2026, the NMD also plays a role in evaluation, funding decisions, and performance tracking of research institutions. A new type of research output – a digital data collection – is now officially recognised in the national database of scientific outputs RIV.[8] Each such collection requires an entry in the NMD, which means the dataset has to reside in a repository that provides CCMM metadata and is indexed by the NMD.

Taken together, CCMM and the NMD reflect a design philosophy different from CLARIN's: where CMDI was deliberately flexible (no schema fits all), CCMM encodes a minimum viable description – the fields required to register a research output or obtain a DOI. CLARIN provides recommendations for basic metadata categories (Fišer & Lenardič, 2023) and a shortlist of components and profiles, but these remain advisory rather than mandatory. The analysis that preceded CCMM (Černohlávková et al., 2023) notes that DCAT-AP is gaining adoption as part of the blueprint for building data spaces – an alignment pressure that CLARIN's metadata ecosystem will likely need to address.

4 Authentication and Authorisation

Research infrastructures serving users across institutional and national boundaries require an Authentication and Authorisation Infrastructure (AAI) to manage access, in line with the principle of "as open as possible, as closed as necessary." As a CLARIN centre preparing to integrate with the NRP, we find ourselves at the intersection of two distinct AAI ecosystems. Per Matyska and EOSC-CZ, CARDS, and NRP Projects (2026), a repository joining the NRP must connect to the EOSC-CZ AAI, built on Perun[9]. As a CLARIN centre, we are also a member of the CLARIN Service Providers Federation. CLARIN is modernising the SPF, introducing the CLARIN Single Sign-On (SSO) proxy (Moreira, 2026).

Both infrastructures implement the AARC Blueprint Architecture (AARC Community members and AppInt members, 2019) (AARC BPA), which organises AAI functionality around three capabilities: **Identity Management** (persistent user identifiers), **Collaboration Management** (group membership and entitlements), and **Service Integration** (a single controlled gateway per infrastructure).

4.1 Key Challenges for Dual-AAI Participation

The proxy architecture on its own represents progress: where persistent identifiers and attribute release could not previously be assumed, an SP-IdP-Proxy makes both possible. The proxy can inject an attribute value (such as email) even when the upstream IdP does not release it. The proxy setup also opens the way to more nuanced AAI-level authorisation. Participating in two distinct AAI ecosystems – the national one (EOSC-CZ) and the ERIC one (CLARIN) – nevertheless introduces challenges for service providers and users who hold identities in both simultaneously.

1. **Identifier Continuity**: Each SP-IdP-Proxy issues its own persistent identifiers; established service providers therefore see proxy-authenticated users as new accounts. Users experience this as loss of history (past submissions, preferences). Reconciling proxy-issued identifiers with existing accounts is non-trivial because identities are often pairwise – unique per IdP-SP combination – making reliable matching difficult.

[6] https://libcas.github.io/Metodika-Dataspecer/
[7] https://nma.eosc.cz/
[8] https://www.isvavai.cz/riv
[9] https://perun-aai.org

2. **Entitlement Propagation**: Which entitlements a user holds depends on which proxy brokered their session. The CLARIN proxy-IdP does not document forwarding upstream attributes beyond *eduPersonPrimaryAffiliation*; Perun, by contrast, automatically forwards *eduPersonEntitlement* from linked identities. A user whose group membership grants access to a resource may find those rights absent after authenticating through the CLARIN SSO, but present when logging in via EOSC-CZ.

3. **Cross-Community Resource Access**: Automated pipelines (a researcher combining restricted CLARIN language resources with EOSC-CZ compute, for instance) require token-based delegated access[10] that crosses AAI boundaries. Coordinating that authorisation where access rights are tied to different digital identities in each AAI remains an open challenge; OAuth 2.0 Proxied Token Introspection (Kanellopoulos et al., 2026) is a candidate approach, pending inter-AAI trust agreements.

These challenges are not merely theoretical: they shape how CLARIN service operators onboard users and plan for interoperability with other research infrastructures, and they will recur wherever CLARIN centres integrate with EOSC nodes.

5 Conclusion

The full scope of what to prepare for when engaging with another infrastructure, such as a national EOSC node, depends on the mode of integration. In the case of the NRP even the least demanding mode of engagement requires metadata and AAI alignment. It is reasonable to expect this will be the recurring minimum for other nodes and centres, which in practice entails preparing for:

- **Metadata alignment**: engaging with another metadata infrastructure built on different philosophies, technologies, and standards such as DCAT-AP.
- **AAI alignment**: connecting to another AAI alongside CLARIN's, and facing the dual-AAI challenges of Section 4 – identifier continuity, entitlement propagation, cross-community resource access.
- **Administrative alignment**: specifying the commitments of both sides and updating the materials (documentation, processes) that are impacted.

Interoperability, ease of access, and good governance are the core of data-sharing infrastructure. Engaging with different communities helps see these from new angles and aim for better services.

Acknowledgments

This output was developed with financial contributions from the EOSC CZ initiative through the National Repository Platform for Research Data project (CZ.02.01.01/00/23_014/0008787), funded by the Programme Johannes Amos Comenius (P JAC) of the Ministry of Education, Youth and Sports of the Czech Republic (MEYS).

It has also been supported by the Ministry of Education, Youth and Sports of the Czech Republic, Project No. LM2023062 LINDAT/CLARIAH-CZ.

References

AARC Community members and AppInt members. (2019). Aarc blueprint architecture 2019 [AARC-G045]. https://doi.org/10.5281/zenodo.3672785

Černohlávková, P., Vyčítalová, H., & Klímek, J. (2023). *Analýza mezinárodních metadatových schémat pro popis výzkumných dat*. https://doi.org/10.48813/6KJX-CR80

Fišer, D., & Lenardič, J. (2023). Qualitative recommendations for depositing lrts at clarin repositories. https://office.clarin.eu/v/CE-2022-2138-qualitative-depositing-recommendations.pdf

[10]The SP-IdP-Proxy can offer a protocol translation layer and offer OIDC even if the upstream IDPs do not; there is also a shift of focus from SAML towards OIDC and OAuth 2.0 in the upcoming AARC BPA update.

Kanellopoulos, C., Liampotis, N., Bučík, D. F., Hardt, M., Jensen, J., Kanakarakis, I., Sallé, M., Kuczyński, T., & Pavlíček, J. (2026, January). Oauth 2.0 proxied token introspection (aarc-g052). https://doi.org/10.5281/zenodo.10205863

Matyska, L., & EOSC-CZ, CARDS, and NRP Projects. (2026). Conditions for creating new and modifying existing repositories in the nrp, version 3.4. https://docs.nrp.eosc.cz/en/docs/repo_admins/operating-repositories-in-the-nrp/conditions-for-creating-repositories

Moreira, A. (2026). Introducing the clarin single sign-on proxy. https://forum.clarin.eu/t/introducing-the-clarin-single-sign-on-proxy/1216

The Metadata Unit of the National Library of Technology. (2025). Homepage - czech core metadata model for research data. https://www.ccmm.cz/en/

Vyčítalová, H., Kuchynka, A., Černohlávková, P., & Šlesaríková, J. (2024). Shrnutí výsledků dotazníku k potřebám popisu a vyhledávání datasetů. *National Library of Technology*. https://doi.org/10.48813/6YTW-BE20

Windhouwer, M., & Goosen, T. (2022). Component metadata infrastructure. In D. Fišer & A. Witt (Eds.), *The infrastructure for language resources* (pp. 191–222). De Gruyter. https://doi.org/doi:10.1515/9783110767377-008