National Czech Programme | eosc

Building DANTE^{c} : Lessons from Community Engagement

A Hybrid Approach to Designing a Domain-Specific Research Data Repository

Anastasiia Prytuliak

DANTE^{c} Repository Manager

Heyrovský Institute of the Czech Academy of Sciences

Spolufinancováno Evropskou unií

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

Introducing MATECH EOSC WG

- **PS MATECH** (Working Group for Materials Science and Engineering) under the EOSC CZ National Repository Platform (NRP).

A vibrant network **of 50+ active members** representing 19 different research institutions and universities. Members span computational physics (DFT), experimental chemistry, and high-end characterization (TEM/SEM/NMR). The WG isn't just an audience; it is the primary architect of the **DANTE^{c}** repository's requirements.

The **DANTE^{c}** Repository is a thematic data repository for materials science, engineering, chemistry, and related disciplines.

- **Technical Foundation:** Built on the Invenio platform, integrated into the National Repository Platform (NRP).
- **Core Principle:** "***As open as possible, as closed as necessary***" — designed to handle both academic and sensitive industry data.

VYSOKÁ ŠKOLA CHEMICKO-TECHNOLOGICKÁ V PRAZE

TUL

ÚJF ÚSTAV JADERNÉ FYZIKY AV ČR

ZÁPADOČESKÁ UNIVERZITA V PLZNI

National Czech Programme | eosc

Understanding Our Community

J. Heyrovský Institute of Physical Chemistry

Disciplines

Computational physicists (DFT, modeling)

Experimental chemists (synthesis, analysis)

Materials design & development (microscopy, characterization)

Data stewards (few)

Engineers – technology development, products

Career Stages

Principal Investigators

Postdocs

PhD students

Technical specialists

Mix of seniority levels

Institutional Diversity

15+ different institutions

Different IT infrastructures

Different data practices

Different policies

Varied repository experience

Industry Partners (Broader Stakeholder Community)

Softli, Škoda, Crytur, ČVUT Electroforming, voda.cz, Entry Engineering, and others

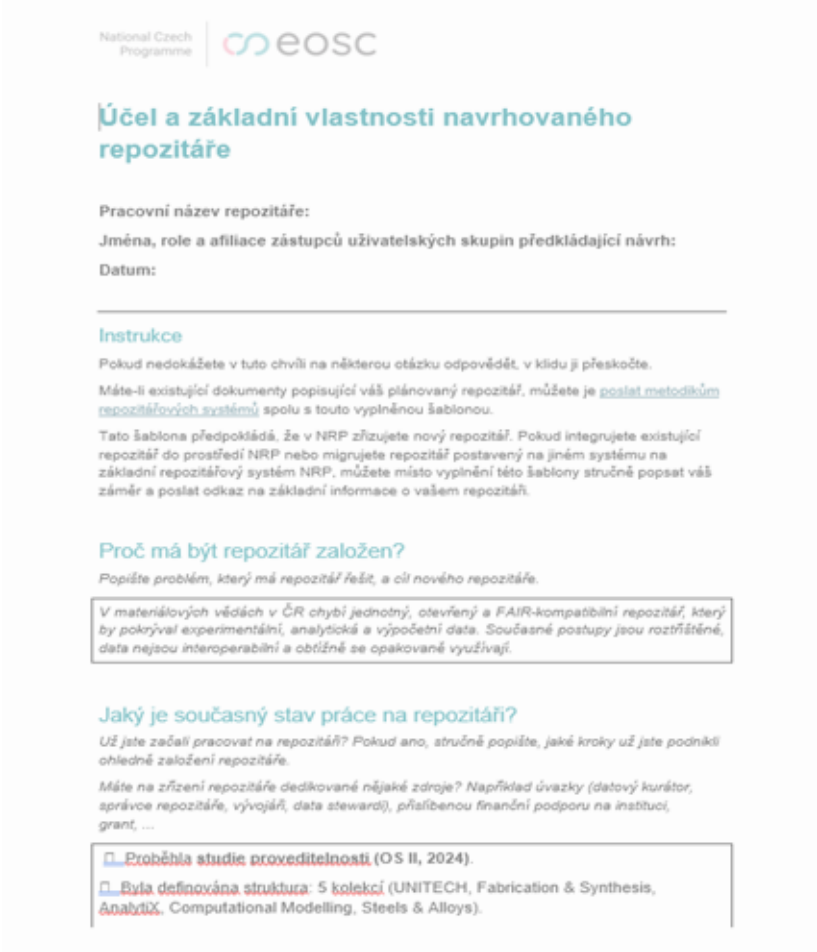
Defining the DANTE^{C} Blueprint

Building a repository requires understanding what researchers actually need — not what we think they need. This is the story of how we learned that lesson. We need community to define:

Motivation & Goals: What specific problems are we solving?

Data Scope & Scale: What types of data belong here? What are the realistic volumes (GB vs. TB) and entry counts?

Standards: Which domain-specific formats are actually used in the group? Any metadata?



National Czech Programme | eosc

Účel a základní vlastnosti navrhovaného repozitáře

Pracovní název repozitáře:

Jména, role a afiliace zástupců uživatelských skupin předkládající návrh:

Datum:

Instrukce

Pokud nedokážete v tuto chvíli na některou otázku odpovědět, v klidu ji přeskočte.

Máte-li existující dokumenty popisující váš plánovaný repozitář, můžete je poslat metodikům repozitářových systémů spolu s touto vyplněnou šablonou.

Tato šablona předpokládá, že v NRP zřizujete nový repozitář. Pokud integrujete existující repozitář do prostředí NRP nebo migrujete repozitář postavený na jiném systému na základní repozitářový systém NRP, můžete místo vyplnění této šablony stručně popsat váš záměr a poslat odkaz na základní informace o vašem repozitáři.

Proč má být repozitář založen?

Popište problém, který má repozitář řešit, a cíl nového repozitáře.

V materiálových vědách v ČR chybí jednotný, otevřený a FAIR-kompatibilní repozitář, který by pokrýval experimentální, analytická a výpočetní data. Současné postupy jsou roztříštěné, data nejsou interoperabilní a obtížně se opakovaně využívají.

Jaký je současný stav práce na repozitáři?

Už jste začali pracovat na repozitáři? Pokud ano, stručně popište, jaké kroky už jste podnikli ohledně založení repozitáře.

Máte na zřízení repozitáře dedikované nějaké zdroje? Například úvazky (datový kurátor, správce repozitáře, vývojáři, data stewardi), přislíbenou finanční podporu na instituci, grant, ...

☐ Proběhla studie proveditelnosti (OS II, 2024).

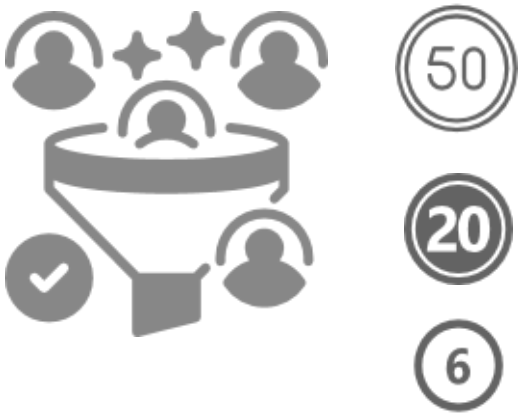
☐ Byla definována struktura: 5 kolekcí (UNITECH, Fabrication & Synthesis, AnalytiX, Computational Modelling, Steels & Alloys).

Access Levels: Who are the user groups? Who needs "Open" vs. "Restricted" or "Authorized" access?

Functional Priorities: What are the "Must-Have" features for their research workflows?

Ecosystem Fit: How should DANTE^{C} connect with existing tools and national infrastructures?

Written Templates — Uneven/Limited Response Depth



High-barrier tasks (e.g., long forms) lead to low engagement from busy researchers.

Response Quality	Count	What They Provided
✅ Comprehensive	2	Detailed technical specs, volume estimates, roles, requirements
⚠️ Focused	1	Infrastructure/roles detailed, but no volume forecasts
⚡ Minimal/Partial	3	Basic data types, brief or missing details

- Written forms work well for highly engaged respondents with time and a mandate
- Most researchers provided just enough to start a conversation — not design-ready specifications

Why Written Templates Failed to Engage

Abstract questions for concrete thinkers
→ "Why should the repository exist?" is too philosophical for bench scientists who think in work pipelines, not strategy
Vocabulary mismatch
→ "Use cases," "MoSCoW prioritization," "ecosystem fit" — business/IT jargon unfamiliar to MATECH researchers
Scope ambiguity
→ "Am I answering for myself? My lab? My field? All of materials science?"

Time investment unclear
→ 11 complex questions = 30-45 minutes, but no clear benefit stated upfront
No accountability
→ Easy to postpone indefinitely when there's no consequence

Question order backwards
→ Started abstract ("Why does this exist?"), should have started concrete ("What data do you generate?")
Audience mismatch
→ Questions designed for project leads (budgets, strategy), sent to bench researchers
No examples or context
→ "What does a good answer look like?" "Why does this question matter?"

The Questions That Stumped Even Engaged Respondents:
✗ "Estimate data volume/entries per year" → Requires tracking they don't do + forecasting future behavior
✗ "How does the repository fit into the existing ecosystem?" → Requires knowledge of other repositories that most don't have
✗ "Describe user groups and their use cases" → Requires UX thinking that researchers aren't trained in

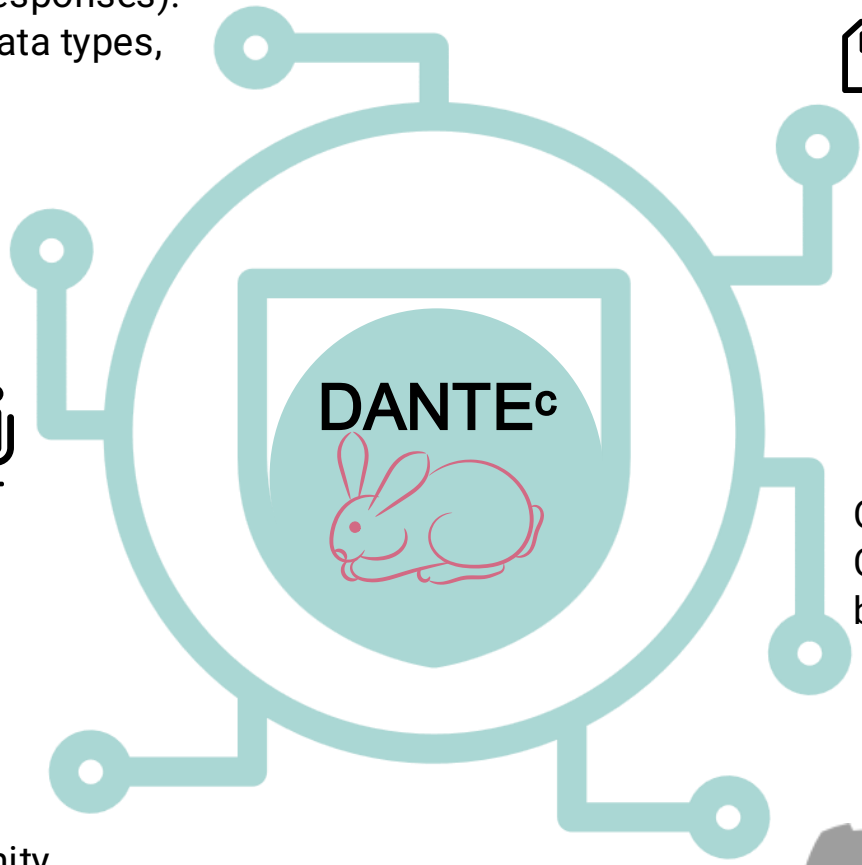
The template assumed strategic thinking, ecosystem knowledge, and UX expertise that most bench researchers don't have. It asked them to design a system, not describe their needs.

Our Solution - The Hybrid Approach: User-Centric

NRP questionnaire (6 responses): Baseline data: scope, data types, initial needs

1 Industry Round Tables (and more to come): Industry-specific needs (IP, access, data market, sandbox, schools)

6 face-to-face interviews (and more to come): Detailed workflows, pain points, "why" behind needs



Ongoing hallway/email exchanges: Clarifications, follow-ups, trust-building

Regular WG discussions: Consensus-building, priorities, community validation, community building …

Discord

Data types & formats:
Detailed lists: NMR, SEM, TEM, XRD, HPLC + file formats (.RAW, .SER, .csv)
✓ This was valuable inventory

Volume estimates:
Only 2 detailed, others vague: "Few kB to few TB" or blank

Why DANTEc is needed:
Generic: "Field-specific metadata, better than Zenodo"

Preferred system:
"Invenio" (surface-level)

Functional priorities:
Generic lists: "Search, API, user-friendly"

Semantic search needs:
Specific: "Find materials with U parameter between 0.41 and 3 eV"
Workflow critique:
"Community review before FAIR check doesn't make sense"
API requirements:
"We need CLI for automated deposition, not just web forms"
Domain-specific metadata:
"Generic DataCite fields aren't enough for NMR parameters"
Real fears:
"I'm worried about data theft and commercial misuse", “afraid of foreign spies”, ”afraid to be judged in case someone will find a mistake.”
Industry needs:
IP protection, controlled access, "data market" concept

Templates provided baseline inventory (what, who, basic preferences). Conversations revealed actionable requirements (why, how, concerns, critiques).

How We Conduct Verbal Engagement

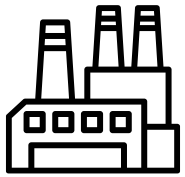
Duration: ~1 hour each

Free-flowing conversation

Prepared questionnaire as backup

Key techniques:

- "Show me your workflow" → they demonstrated on their laptops
- "Tell me about the last time you tried to publish data"
- "What frustrates you most about current systems?"
- Immediate follow-up: "What do you mean by that?" "Can you give an example?"

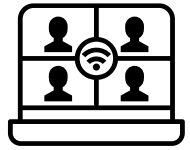


Participants:

Softli, Škoda, Crytur, ČVUT Electroforming, voda.cz, Entry Engineering, HIPCH

What made it work:

- Open dialogue format (only 1 short intro presentation)
- Mixed audience → industry and academia heard each other's concerns
- Real-time debate: IP protection, and semantic search.



MATEChat WG meetings:

Every 2 weeks
55+ members
Consensus-building, priority validation, trust-building over time

Informal channels:

- Hallway conversations
- Email exchanges
- Task-oriented WG -regular meetings
- Discord (planning more active use)

Actionable Lessons for Repository Builders

Actionable Lessons:

- Hybrid Approach Works → Combine multiple engagement channels for a comprehensive understanding
- Verbal > Written → Prioritize interviews and conversations over questionnaires alone
- Start Concrete, Not Abstract → Build conversations and design questionnaires using the "funnel approach"
- Iterate Templates → Refine written tools based on feedback and response rates
- Leverage Existing Communities → Use established groups (like MATEChat PS with 55 members) as engagement platforms






National Czech Programme | eosc

Thank you!

Contacts:

Anastasiia Prytuliak

DANTE[c] Repository Manager

anastasiia.prytuliak@jh-inst.cas.cz

Marek Cebecauer

DANTE[c] Repository Guarantor

marek.cebecauer@jh-inst.cas.cz

Spolufinancováno Evropskou unií

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY