Controlled Vocabularies, Crosswalks, and Integration with European Infrastructure

Approximate duration: ~40 min. The text is intended to be read aloud together with the presentation; interactions and demonstrations are included in the timing.

Key idea
Interoperability begins with a shared machine-recognisable language. Controlled vocabularies and a semantic crosswalk make meanings explicit, but a mapping table is only the first step: genuine machine semantics requires persistent IRIs, formal representation, and measurable FAIR quality.

After this lecture, participants will be able to

- explain the role of controlled vocabularies;
- interpret a semantic-crosswalk row and mapping modes;
- distinguish CSV mapping from formal machine semantics;
- describe a realistic repository → discovery → EOSC path;
- interpret F-UJI appropriately and explain its limits.

Updated as of 2 November 2026
The EOSC Federation has been in production since 8 October 2026, with 14 production nodes. F-UJI v4.0.0 was released on 29.08.2026; the public web UI still carries a disclaimer that the service is rapidly evolving and is not a productive environment. Current indexing of specific DataverseUA records in OpenAIRE should be checked using the concrete PID immediately before teaching; DataverseUA → OpenAIRE → EOSC should not be presented as a guaranteed automatic pipeline.

1. “The same” for a person, different strings for a machine [0–4 min]

On screen
Examples: PBE / pbe / GGA-PBE; NVT / nvt; QE / Quantum ESPRESSO.

Let us begin with something simple. A person reads PBE, pbe, and GGA-PBE and will often understand that they refer to a related exchange-correlation concept. A machine sees different strings. The same applies to NVT/nvt, eV/A/eV Å−1, and QE/Quantum ESPRESSO.

If metadata are used only by a local author, this may not matter. If the metadata are expected to pass validation, aggregation, a knowledge graph, or machine search, their meaning needs to be stable.

A controlled vocabulary reduces variation in labels. An external IRI gives a concept an identity independent of its label. This is why Interoperability in FAIR is not simply “JSON instead of Word”. JSON can still contain ambiguous text.

Machine interoperability begins with agreement on which values are allowed, what they mean, which identifiers are used, and how a local profile relates to international standards.

Transition to the next block

Science changes, however. A vocabulary must be strict enough to be useful without blocking new methods. This is why we use different field patterns.

2. Three field patterns: closed enum, enum + Other, term/IRI [4–10 min]

On screen

Three columns: CLOSED ENUM / ENUM + OTHER / TERM + IRI.

A closed enum works well for stable operational categories such as role, status, or relation type. The validator rejects an unknown value, and that is useful: `Complete`, `complete`, and `done` do not spread through the dataset as separate forms.

Enum + Other is useful in an evolving domain. We provide common values but allow a new one with mandatory detail. `Other` without an explanation should not become an “easy exit”, otherwise the vocabulary stops doing its job.

Term/IRI is appropriate where an authoritative vocabulary already exists. There is no point in copying every QUDT unit locally. It is better to preserve the external IRI and, where useful, a local label.

These patterns can be combined. A stage type can use a common enum + Other; a unit can use an external IRI; a status can use a closed enum. Guidelines should explain which mode is used.

Vocabulary governance is more than the current list of terms. We need definitions, versioning, a procedure for adding and deprecating terms, and ownership. Otherwise even a good first release gradually degrades.

Audience interaction

Ask the audience: functionals, ensembles, codes, model types - which should be a closed enum, which enum + Other, and which an external IRI?

3. Semantic crosswalk: documenting mapping decisions [10–16 min]

On screen

Mini-table: MATSCI field | external scheme | term/IRI | mapping mode | unit.

A crosswalk is a table that shows relationships between local fields or concepts and external standards. It is needed because the schema by itself does not explain why a particular field corresponds to PROV Activity or DataCite RelatedIdentifier.

A crosswalk row contains the local path, external scheme, external term or IRI, mapping mode, unit IRI, and notes. For example, `provenance.activities[]` may reuse PROV-O Activity. `dft.ecutwfc` may profile an energy-cutoff concept. A transition selection may be a local extension.

The crosswalk makes semantics reviewable. Another team can inspect the mapping and propose an authoritative term. Later, if an external ontology changes, the mapping can be updated without redesigning the whole package.

A crosswalk is also important for transparency. If no suitable external term has been found, this becomes a visible local extension rather than a hidden assumption. If a mapping is only approximate, the MAP mode does not pretend it is exact reuse.

In MATSCI-NASU, the working crosswalk is a governance artefact and a basis for future machine semantics. It should not remain hidden in a developer's internal spreadsheet.

Transition to the next block
We will now examine five mapping modes, from maximum reuse to an explicitly local field.

4. REUSE → PROFILE → MAP → LOCAL EXTENSION → LOCAL [16–20 min]

On screen
Five mapping-mode blocks.

REUSE means that the external concept matches the exact intended meaning. PROFILE means that the semantics are reused, while the local profile adds constraints such as requiredness, cardinality, or domain usage.

MAP means a close correspondence, but the concepts are not identical. LOCAL EXTENSION means that the domain concept is absent or insufficiently represented externally, so we create a local term with a definition and relation. LOCAL is an operational field with no useful semantic analogue.

The advantage of this scale is that it does not force developers to pretend that everything has an external equivalent. Semantic honesty is more important than claiming "100% mapped".

The reuse-before-invent principle reduces duplication. But a wrong external IRI is worse than a correct local extension. Mapping therefore needs domain review, not only keyword searching.

For a future scientific article, this is also a measurable design result: we can show what proportion of concepts were reused, profiled, mapped, or kept local, and which gaps in community standards were revealed by the profile.

Transition to the next block
Rather than creating one giant ontology, the profile combines several authoritative semantic layers.

5. Several authoritative layers instead of a "super-ontology" [20–24 min]

On screen
Six cards: DataCite; PROV-O; QUDT; ASMO/PMDco; IUCr CIF; OPTIMADE.

DataCite provides citation and discovery metadata: creators, identifiers, relations, and funding. PROV-O provides a provenance graph. QUDT provides units and quantity kinds. ASMO/PMDco provide

atomistic simulation and process concepts. IUCr CIF covers crystallography. OPTIMADE provides semantics for structural APIs.

Each standard solves a different problem. It is perfectly normal for one package to use several vocabularies. Interoperability does not mean “one vocabulary for everything”; it means coherent use of authoritative semantics.

A local MATSCI vocabulary is needed for genuine gaps: transition-specific operational concepts, package statuses, or details that are not adequately represented externally. But the local vocabulary should be minimal and documented.

This approach also simplifies integration. The citation layer can map to DataCite/Dataverse; provenance can remain PROV-compatible; material structures can use domain standards; units can use QUDT. A downstream system can use the part of the semantics it understands.

A practical rule for a data steward: before introducing a new term, check an authoritative source; record the definition and provenance of the mapping decision; do not use Wikipedia as the formal vocabulary URL.

Transition to the next block
The next boundary is important: a CSV crosswalk does not by itself make metadata semantically machine-readable for arbitrary systems.

6. CSV crosswalk ≠ machine semantics [24–28 min]

On screen
CSV/Excel mapping → persistent IRI + SKOS vocabulary → JSON-LD / formal export.

A CSV crosswalk documents correspondence. It is excellent for review and development. But an external system does not automatically know that this file should be interpreted as an ontology mapping.

For stronger machine semantics, local terms should have persistent IRIs. A vocabulary can be published in SKOS or a similar formal representation. A JSON-LD context can connect fields and values to IRIs directly in the data representation.

For workflow or run exchange, RO-Crate/WRROC can be considered. But until an exporter and conformance tests are ready, it is accurate to say “conceptually aligned with RO-Crate/PROV” or “crosswalk prepared”, rather than “fully compatible”.

This is a general principle for scientific claims. The level of documentation, the level of machine semantics, and the level of tested interoperability are different. They should have different readiness labels.

The semantic roadmap is therefore: 1) crosswalk; 2) stable local IRIs; 3) publication of a controlled vocabulary; 4) machine-readable context or export; 5) conformance and interoperability tests.

Transition to the next block
Why do this? So that a repository object can be understood by downstream systems and federation services.

7. From DataverseUA to discovery and EOSC: a realistic path [28–32 min]

On screen

Research package → DataverseUA DOI/metadata → aggregators/discovery → EOSC/services; caption: “not an automatic pipeline”.

DataverseUA provides the publication and preservation layer: a landing page, dataset metadata, files, access, DOI/DataCite integration, and APIs or OAI endpoints depending on configuration. This is already an important part of FAIR infrastructure.

Visibility can then grow through aggregators and discovery services. But harvesting requires registration or an endpoint, a supported metadata format, and successful indexing. A concrete record should therefore be checked using its concrete PID.

For this reason, the course should not claim that every DataverseUA record automatically appears in OpenAIRE or EOSC. The accurate model is: DataverseUA can expose machine-readable metadata; downstream integration requires configuration and testing.

For the materials domain, NOMAD can be useful in parallel. Raw calculations can be parsed and normalised for domain search, while a complete package with transition and provenance documentation is preserved in DataverseUA. Related identifiers can connect the same object across systems.

Interoperability is the ability of these layers to connect identity and meaning, not necessarily to copy all files physically.

Transition to the next block

We will now update the EOSC context: in October 2026, the Federation moved from build-up to production.

8. EOSC Federation in production [32–35 min]

On screen

08.10.2026 → 14 nodes in production; National / Thematic / e-Infrastructure + EOSC EU Node.

On 8 October 2026, the EOSC Association announced the start of the production phase. The EOSC EU Node and 13 pioneering nodes are in production. These include national nodes from Czechia, Finland, Italy, Poland, and Slovakia; thematic nodes; and e-infrastructure nodes such as EGI, EUDAT, and GÉANT.

At the same time, the Federation website presents a broader landscape of 28 nodes with different statuses and waves. Some are still working towards production readiness. The Federation continues to develop policies, federating capabilities, and onboarding.

This matters for the course: EOSC is no longer only a governance plan. But it is also not a single portal into which “all data” should be uploaded. It is a federation of nodes and services in which repositories and domain infrastructures remain autonomous.

For MATSCI-NASU, the appropriate ambition is to become EOSC-ready through identifiers, metadata, AAI and service policies, and semantic interoperability - not merely to use the word EOSC in a description.

Lecture 4 needs only this current operational fact and should not overload participants with governance details. Those are covered in the separate EOSC course.

Transition to the next block
The final practical topic is FAIR assessment. A tool can measure a great deal, but not everything.

9. F-UJI: automated assessment and limits of interpretation [35–38 min]

On screen
Two columns: SEES / DOES NOT SEE; F-UJI v4.0.0 - 29.08.2026.

F-UJI is an automated FAIR data assessment service. It analyses machine-accessible signals such as a PID, landing-page metadata, metadata standards, licence and access information, data links, and other generic metrics.

F-UJI v4.0.0 was released on 29 August 2026. The public web UI still carries a disclaimer that the service is rapidly evolving and is not a productive environment. This does not prevent its use as a teaching or diagnostic tool, but its output should not be called a “certificate”.

F-UJI does not have enough domain knowledge to answer questions such as: Is the exact split correct? Was a human selection justified? Was the final MLIP really used in MD? Is the validity domain scientifically adequate? Those questions belong to domain package review.

An automated score should therefore be read as a set of signals and gaps. If the tool cannot see a licence, the issue may lie in author metadata or repository exposure. If it cannot detect a domain vocabulary, the issue may be the semantic profile. The score is not a ranking of the researcher.

A good teaching demo is to show one published object and examine which metrics depend on the repository and which depend on the author. Then ask: what could F-UJI never see in our DFT → MLIP transition?

Transition to the next block
We will finish with the division of responsibilities and six rules that summarise the entire course.

10. Researcher, data steward, repository. Course summary [38–40 min]

On screen
Three roles + six concluding rules.

The researcher creates scientific evidence: accurate descriptions, artefacts, licences and relations, decisions, and limitations. The data steward helps select a schema or profile, controlled vocabularies and a crosswalk, and checks completeness and consistency. The repository provides the PID, protocol, metadata exposure, access, and preservation.

When FAIR assessment shows a gap, the useful question is not “Who is to blame?” but “Who can fix it?” A missing licence is a researcher or curator issue; a broken metadata export is a repository issue; an ambiguous local term is a profile or data-steward issue.

Return to Lecture 1. The list of what should be preserved is now broader: final outputs + exact environment + stages and transitions + workflow/provenance/manifest + human decisions + controlled identifiers + validation evidence.

Six takeaways from the course: 1) preserve more than outputs; 2) separate stages and transitions; 3) connect plan/fact/inventory; 4) record the environment and human decisions; 5) use controlled terms and IRIs; 6) check schema + consistency + sufficiency.

The next step for MATSCI-NASU is testing on other materials and nonlinear active-learning workflows. The next step for each participant is to find one transition in their own work where context is currently preserved only in memory, and make it explicit.

Audience interaction
Ask: “What is one thing you will preserve from now on that you did not preserve before?”

Sources and materials

1. DataCite Metadata Schema 4.7. https://schema.datacite.org/meta/kernel-4/
2. W3C PROV-O. https://www.w3.org/TR/prov-o/
3. QUDT. https://qudt.org/
4. ASMO. https://ocdo.github.io/asmo/
5. PMD Core Ontology. https://materialdigital.github.io/core-ontology/
6. IUCr CIF dictionaries. https://www.iucr.org/resources/cif/dictionaries
7. RO-Crate specification. https://www.researchobject.org/ro-crate/
8. F-UJI v4.0.0 (2026). Zenodo DOI: 10.5281/zenodo.22159621; public UI: https://www.f-uji.net/
9. EOSC Association. EOSC Federation moves into production, 8 October 2026.
10. MATSCI-NASU Semantic Crosswalk v1.0 draft / Controlled Vocabulary draft.