



connect·share·advance

Making Data Ready for AI in Clinical

or: how fast is a jaguar?

Rewant, Hexaware

PHUSE Single Day Event

phuse.global



connect·share·advance

How fast is a jaguar?

Shout a number.



Three right answers

connect·share·advance



50 mph

the animal



200 mph

the car



1,000+ mph

the fighter jet

All three are correct.

The word did not change. The context did. An AI only sees the word.



DAY 1

connect·share·advance

Clinical is full of jaguars



Hands up. Day 1 is:

First dose

Randomisation

First visit

Is there a Day 0?

Four sponsors. Four answers. Same therapeutic area.

And every dataset downstream inherits the choice.



PHASE

Same word, same protocol



Phase III

the trial



Accelerated phase

the disease (CML)



Treatment phase

the epoch in the schedule of activities

Three meanings on one page.

A human reads the position. A model reads the letters.



How many diseases is this?

Real strings from AACT, the open ClinicalTrials.gov database

Type 2 Diabetes

T2DM

Diabetes Mellitus, Type 2

NIDDM

diabetes type II

Type II DM

DM2

Type-2 diabetes mellitus

500,000+

registered studies.
Free. Downloadable.

Also public: ClinicalTrials.gov, EU CTIS, WHO ICTRP, openFDA, PubMed, CDISC pilot data.
On request: Vivli, YODA, Project Data Sphere.

Shout a number.



Why AI can't just use it

One disease. Eight spellings.

01 The schema tells you shape, not meaning

design_outcomes, param_type, dispersion_type: a name is not a definition

02 Free text where structure should be

eligibility criteria: a paragraph, not a query

03 Same concept, forty spellings. And vice versa.

one spelling, three concepts: entity resolution both ways

The jaguar problem, at database scale.



Enter the knowledge graph

connect·share·advance



An ontology is context, written down.

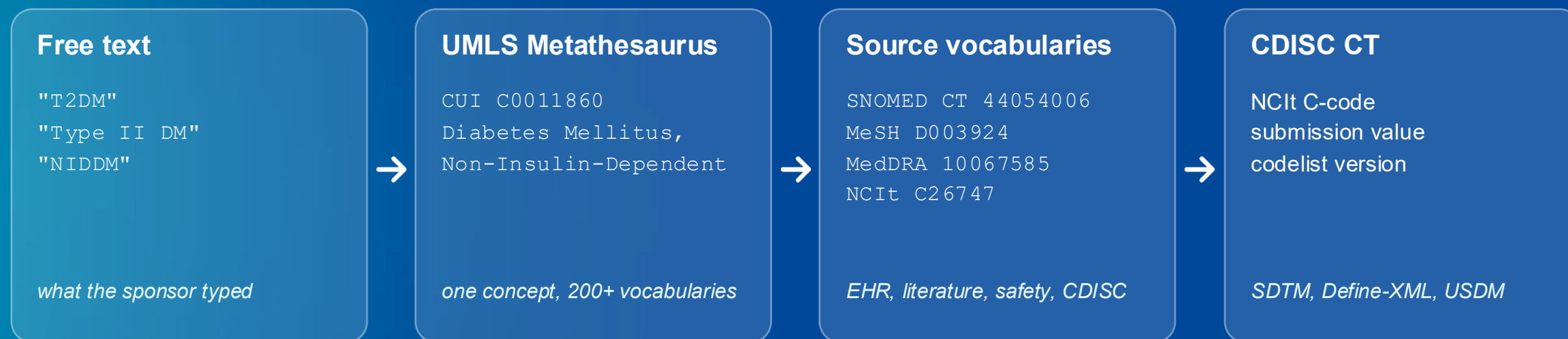
Strings in. Meaning out. Relationships explicit.



Coding is key: UMLS and CDISC CT

connect·share·advance

How one free-text string becomes something every system agrees on



- ✓ UMLS is the crosswalk: one CUI ties SNOMED, MeSH, MedDRA, LOINC, RxNorm and NCIt together.
- ✓ MedDRA for safety, SNOMED and LOINC for EHR and labs, NCIt C-codes for everything CDISC: SDTM, Define-XML, USDM.
- ✓ Code at the source, once, with the codelist version. Never re-code at submission.



The Jaguar Test, with examples

connect·share·advance

NOT AI-READY

AI-READY



Structured



The schedule of activities as a table in a PDF



ScheduleTimeline: instances, timings (P28D), windows (P3D), as JSON



Coded



"Phase 3", "Ph III", "phase-III" in three documents



C15602 Phase III Trial, CDISC CT 2025-03-28, one code everywhere



Contextual



"SBP 142" in a listing



142 mmHg, LOINC 8480-6, Baseline = last pre-dose, Treatment epoch, Arm A



Connected



Visit in Section 1.3, assessment in Section 8, joined by a reader



Activity to ScheduledActivityInstance to Encounter to Epoch, all linked by id



Described



A spreadsheet someone emailed



Define-XML and biomedical concepts, version, provenance, reviewer



Hands: how many of the five does your data pass today?



USDM 4.0: the protocol as data

connect·share·advance

60 years of Word documents, re-keyed into EDC, CTMS, SAP and registries, again at every amendment.

What one USDM study carries

Design, phase, blinding

Arms x epochs = study cells

Encounters and SoA timelines

Eligibility criteria

Objectives, endpoints, estimands

Interventions (IDMP aligned)

Amendments, governance dates

Narrative that renders as ICH M11

Where it stands



v4.0 released 3 June 2025 by CDISC with TransCelerate DDF



80 classes; categorical attributes are Code objects bound to CDISC CT



Built to the breadth of ICH M11: adopted Nov 2025, FDA guidance May 2026, EU June 2026



210 CORE conformance rules, executable since Dec 2025

For the first time, the protocol passes the Jaguar Test.



USDM 4.0 snippet: standard classes

connect·share·advance

Abridged: ids, CT versions and some required attributes trimmed to fit the slide

```
{
  "usdmVersion": "4.0",
  "systemName": "protocol-to-usdm pipeline",
  "study": {
    "instanceType": "Study", "name": "CDISC PILOT 01",
    "versions": [{
      "instanceType": "StudyVersion", "versionIdentifier": "2.0",
      "studyDesigns": [{
        "instanceType": "InterventionalStudyDesign",
        "studyPhase": { "standardCode": {
          "code": "C15602", "codeSystem": "http://www.cdisc.org",
          "codeSystemVersion": "2025-03-28", "decode": "Phase III Trial" }},
        "arms": [{ "id": "Arm_1", "name": "Drug A 10 mg",
          "type": { "code": "C174266", "decode": "Investigational Arm" } }],
        "epochs": [{ "id": "Epoch_2", "name": "Treatment",
          "type": { "code": "C101526", "decode": "Treatment Epoch" } }],
        "studyCells": [{ "armId": "Arm_1", "epochId": "Epoch_2",
          "elementIds": ["Element_1"] }],
        "scheduleTimelines": [{ "mainTimeline": true, "timings": [{
          "name": "Week 4", "type": { "code": "C201356", "decode": "After" },
          "value": "P28D", "valueLabel": "Day 29",
          "relativeToFrom": { "code": "C201355", "decode": "Start to Start" },
          "relativeFromScheduledInstanceId": "SAI_Week4",
          "relativeToScheduledInstanceId": "SAI_Day1",
          "windowLower": "P3D", "windowUpper": "P3D", "windowLabel": "+/- 3 days" } ] } ] } ] } ] }
```



Standard classes

Study, StudyVersion,
InterventionalStudyDesign, StudyArm,
StudyEpoch, StudyCell, ScheduleTimeline,
Timing.



Every category is a Code

code, codeSystem, codeSystemVersion,
decode. Phase III is C15602 in every study,
not a string.



Arm x Epoch = StudyCell

The design grid is explicit, so "each arm and
epoch combination occurs once" is a rule a
machine can check.



Provenance: built in, and extended

connect·share·advance

Standard classes carry the audit trail (abridged)

```
"versions": [{
  "instanceType": "StudyVersion",
  "versionIdentifier": "2.0",
  "rationale": "Amendment 2: SoA correction",
  "dateValues": [{
    "instanceType": "GovernanceDate",
    "name": "Protocol approval",
    "type": { "code": "C71476", "decode": "Approval Date" },
    "dateValue": "2026-05-14",
    "geographicScopes": [{ "type": { "code": "C68846",
                                   "decode": "Global" }}] }],
  "amendments": [{
    "instanceType": "StudyAmendment",
    "name": "Amendment 2", "number": "2",
    "summary": "Week 4 visit corrected to Day 29",
    "primaryReason": { "code": { "code": "C207603",
                                   "decode": "Inconsistency And/Or Error In The Protocol" }},
    "changes": [{ "summary": "Day 28 replaced by Day 29",
                  "changedSections": [{ "sectionNumber": "1.3",
                                       "sectionTitle": "Schedule of Activities" }] }],
    "documentVersionIds": ["Doc_v2"] }]
```

Extension attributes carry the lineage (abridged)

```
"epochs": [{
  "id": "Epoch_2", "name": "Treatment",
  "type": { "code": "C101526", "decode": "Treatment Epoch" },
  "extensionAttributes": [{
    "instanceType": "ExtensionAttribute",
    "url": "https://hexaware.com/prov",
    "valueExtensionClass": {
      "instanceType": "ExtensionClass",
      "url": "https://hexaware.com/prov/v1",
      "extensionAttributes": [
        { "url": "https://hexaware.com/prov#source",
          "valueString": "Protocol v2.0, Sec 1.3, Table 1" },
        { "url": "https://hexaware.com/prov#extractor",
          "valueString": "usdm-extractor 1.4, LLM + rules" },
        { "url": "https://hexaware.com/prov#confidence",
          "valueQuantity": { "value": 0.93 } },
        { "url": "https://hexaware.com/prov#reviewer",
          "valueString": "CDS lead, 2026-09-10" }
      ]
    }
  ]
}]
```

Every value can answer: which version, from where, produced by what, checked by whom.



QA and evals on USDM

connect·share·advance

Two different questions: is it valid USDM, and is it the right USDM?



Conformance: CDISC CORE rules

210 rules for USDM 4.0: 172 errors, 38 warnings

DDF00069	Each arm and epoch combination occurs at most once
DDF00009	Every schedule timeline has an anchor: a Fixed Reference timing
DDF00143	Amendment reasons come from codelist C207415
DDF00035	Within a code system version, one code has one decode (warning)



Evals: for AI-produced USDM

- ✓ Gold set: human-curated USDM for 20 to 30 real protocols
- ✓ Precision and recall per class: arms, epochs, criteria, endpoints
- ✓ SoA cell accuracy: every activity by visit, timing and window
- ✓ Code accuracy: right C-code, right codelist version
- ✓ Round trip: USDM to M11 render, diffed against the source
- ✓ Planted contradictions must be caught, not smoothed over

Schema-valid is not the same as right. Conformance checks the model; evals check the meaning.



Many protocols, one graph

connect·share·advance

Spot the contradiction: 10 seconds

Section 1.3 Schedule of Activities

Week 4 visit: Day 29 (+/- 3 days)

Section 8.1 Primary endpoint

Change from baseline at Week 4 (Day 28)

What the graph holds (USDM Timing)

value P28D, valueLabel "Day 29", Start to Start

Day 28 is not Day 29. The Day 1 problem, inside a single protocol.

200 protocols in one graph can:

- ✓ Flag contradictions before submission
- ✓ Answer "who defines Day 1 like us?"
- ✓ Re-use eligibility criteria that worked
- ✓ Auto-derive SDTM trial design domains
- ✓ Run feasibility across the whole portfolio



50



200



1,000+

connect·share·advance

Which jaguar is your data?

Give the machine the context, and every answer is right.

Pick one word in your organisation this week: Day 1, baseline, response. Define it in code.

connect·share·advance



THANK YOU!

Rewant | RewantR@hexaware.com

Making Data Ready for AI in Clinical | PHUSE Single Day Event