Dataset Description

EHDS Linked Health Data Portal

https://mcp.linkeddata.es

Ontology Engineering Group, Universidad Politécnica de Madrid

Overview

The EHDS Linked Health Data Portal provides a synthetic, openly accessible health data resource designed for benchmarking web AI agents in the context of the European Health Data Space (EHDS), established by Regulation (EU) 2025/327. The resource comprises three components accessible through a single public endpoint:

- **Clinical knowledge graph** — 21.2 million RDF triples encoding synthetic patient records for 573 unique patients across 30 clinical cohorts, serialised as FHIR R4 on RDF using HL7 FHIR namespaces and coded with SNOMED CT, LOINC, and RxNorm.
- **HealthDCAT-AP catalogue** — 944 triples describing dataset metadata, temporal coverage, health categories, and ODRL usage policies in accordance with HealthDCAT-AP Release 5.
- **MCP connector** — seven typed tool interfaces implementing the Model Context Protocol over the SPARQL endpoint, enabling AI agents to query the knowledge graph without generating SPARQL directly.

All data is synthetic. Patient records were generated with Synthea and converted to RDF; no real patient data is included. All resources are released under CC-BY 4.0.

Endpoint and Access

Resource	URL
Landing page	https://mcp.linkeddata.es
SPARQL endpoint	https://mcp.linkeddata.es/sparql
SPARQL browser UI	https://mcp.linkeddata.es/sparql/
MCP connector	https://mcp.linkeddata.es/connector
RAG vector store	https://mcp.linkeddata.es/rag
Graph visualisation	https://mcp.linkeddata.es/visualization
Supplements	https://mcp.linkeddata.es/supplements

Named Graphs

The triple store (Apache Jena Fuseki 6, TDB2) hosts four named graph namespaces:

- `https://ehds-prototype.example.org/graph/catalogue` — HealthDCAT-AP catalogue
- `https://ehds-prototype.example.org/graph/{cohort-id}` — one named graph per clinical cohort (30 graphs)

Knowledge Graph

Generation Pipeline

Synthetic patient data was generated using Synthea (MITRE Corporation), which produces clinically realistic FHIR R4 JSON bundles. More than fifty thousand patients were generated

from a random population model and filtered by SNOMED CT-coded primary condition to obtain the 30 cohorts listed in Section 4. FHIR R4 JSON bundles were converted to RDF Turtle using a custom pipeline preserving five FHIR resource types:

- `fhir:Patient` — demographics, gender, race, birth year, consent flags
- `fhir:Condition` — SNOMED CT-coded diagnoses
- `fhir:Observation` — LOINC-coded clinical measurements
- `fhir:MedicationRequest` — RxNorm-coded prescriptions
- `fhir:Encounter` — healthcare visit records

Terminological Coverage

Resource type	Total	Coded	Coverage
Conditions (SNOMED CT)	60,794	60,794	100.0%
Observations (LOINC)	1,077,431	1,077,431	100.0%
Medications (RxNorm)	158,900	107,477	67.6%

Medication coverage of 67.6% reflects real-world coding patterns where proprietary drug identifiers do not always align with RxNorm controlled terminology.

Scale

Metric	Value
Total RDF triples	21,200,000
Unique patients	573
Clinical cohorts	30
Conditions (with overlap)	60,893
Healthcare encounters	94,181
Named clinical graphs	30

Patient, condition, and encounter counts are by design overlapping across cohorts. The figure of 573 unique patients is independently verifiable via:

```
SELECT (COUNT(DISTINCT ?p) AS ?n) WHERE {
  GRAPH ?g { ?p a <http://hl7.org/fhir/Patient> . }
  FILTER(?g != <https://ehds-prototype.example.org/graph/catalogue>)
}
```

Clinical Cohorts

The table below lists all 30 cohorts. Policy identifiers ($\mathcal{P}1$–$\mathcal{P}8$) refer to the ODRL policy sets described in Section 6. The dementia cohort contains zero patients and serves as a null-result robustness probe.

Cohort	Policy	Patients	Conditions	Encounters
Type 2 Diabetes	$\mathcal{P}1$	40	3,277	4,117
Essential Hypertension	$\mathcal{P}1$	40	2,979	4,802
Metabolic Syndrome	$\mathcal{P}1$	40	3,383	5,180
Hyperlipidemia	$\mathcal{P}1$	40	2,282	3,247
Hypothyroidism	$\mathcal{P}1$	40	2,166	3,196
Obesity	$\mathcal{P}1$	40	2,031	3,393

(continued)

Cohort	Policy	Patients	Conditions	Encounters
Prediabetes	$\mathcal{P}1$	40	1,893	2,981
Anemia	$\mathcal{P}1$	40	2,188	3,397
Heart Failure	$\mathcal{P}2$	15	999	913
Stroke	$\mathcal{P}2$	40	1,703	3,133
Myocardial Infarction	$\mathcal{P}2$	40	2,705	4,005
Ischaemic Heart Disease	$\mathcal{P}2$	40	3,292	5,340
Atrial Fibrillation	$\mathcal{P}2$	40	3,156	4,825
Anxiety Disorder	$\mathcal{P}3$	10	960	1,011
PTSD	$\mathcal{P}3$	10	807	1,240
Dementia	$\mathcal{P}3$	0	0	0
Alzheimer's Disease	$\mathcal{P}4$	40	4,258	5,408
Osteoporosis	$\mathcal{P}4$	40	2,390	3,770
Rheumatoid Arthritis	$\mathcal{P}4$	38	2,789	3,931
Chronic Kidney Disease	$\mathcal{P}4$	40	4,502	8,471
Asthma	$\mathcal{P}5$	28	1,487	2,568
COPD	$\mathcal{P}5$	40	1,884	2,634
Obstructive Sleep Apnea	$\mathcal{P}5$	40	2,024	3,563
Urinary Tract Infection	$\mathcal{P}5$	40	2,379	4,264
Breast Cancer	$\mathcal{P}6$	10	543	1,403
Prostate Cancer	$\mathcal{P}6$	10	1,144	1,938
Colorectal Cancer	$\mathcal{P}6$	10	642	1,158
Osteoarthritis	$\mathcal{P}7$	40	1,406	1,820
Substance Use Disorder	$\mathcal{P}8$	10	1,009	1,641
Chronic Pain	$\mathcal{P}8$	10	615	832
Total (with overlap)		**911**	**60,893**	**94,181**
Unique (via SPARQL)		**573**	**33,645**	**52,104**

HealthDCAT-AP Catalogue

The catalogue is serialised as 944 RDF triples and loaded into the catalogue named graph. Each dataset entry includes the following HealthDCAT-AP Release 5 properties:

- `dct:title`, `dct:description`
- `hdcat:populationSize`
- `hdcat:healthCategory` (SNOMED CT URI)
- `hdcat:dataStandard` (value: `"FHIR R4"`)
- `dct:temporal` with `dcat:startDate` and `dcat:endDate`
- `dct:license` (https://creativecommons.org/licenses/by/4.0/)
- `odrl:hasPolicy` linking to the applicable ODRL policy

Temporal Coverage by Cohort Group

Cohort group	Start	End
$\mathcal{P}1$ metabolic (Diabetes, Hypertension, Obesity, etc.)	2000-01-01	2025-12-31
$\mathcal{P}2$ cardiovascular (Heart Failure, Stroke, etc.)	2005-01-01	2025-12-31
$\mathcal{P}3$ mental health (Anxiety, PTSD, Dementia)	2010-01-01	2025-12-31
$\mathcal{P}4$ degenerative (Alzheimer's, Osteoporosis, etc.)	2003-01-01	2025-12-31
$\mathcal{P}5$ respiratory/UTI (Asthma, COPD, Sleep Apnea, UTI)	2000-01-01	2023-12-31
$\mathcal{P}6$ oncology (Breast, Prostate, Colorectal Cancer)	2010-01-01	2025-12-31
$\mathcal{P}7$ musculoskeletal (Osteoarthritis)	2008-01-01	2025-12-31
$\mathcal{P}8$ substance/pain (Substance Use Disorder, Chronic Pain)	2015-01-01	2025-12-31

ODRL Usage Policies

Eight ODRL 2.2 policies ($\mathcal{P}1$–$\mathcal{P}8$) govern data use. Each policy is a named individual in the catalogue graph and is linked from datasets via `odrl:hasPolicy`.

ID	Permitted	Prohibited	Obligations
$\mathcal{P}1$	Reproduce, distribute (research)	Commercial; re-identify	CC-BY 4.0 attribution
$\mathcal{P}2$	Reproduce, distribute (research; clinical care)	Commercial; re-identify	Attribution; data minimisation
$\mathcal{P}3$	Reproduce (research; ethics approval granted)	Commercial; re-identify; distribute without approval	Attribution; notify data controller
$\mathcal{P}4$	Reproduce (academic research)	Commercial; governmental; re-identify	Attribution; publish open access
$\mathcal{P}5$	Reproduce, distribute, derive (non-commercial)	Commercial	CC-BY 4.0 attribution
$\mathcal{P}6$	Reproduce (cancer research; IRB documented)	Commercial; non-cancer use; re-identify	Attribution; delete after study
$\mathcal{P}7$	Reproduce, distribute (research; $\leq$2 years)	Commercial; retain >2 years; re-identify	Attribution; delete after 2 years
$\mathcal{P}8$	Reproduce (research; anonymisation verified)	Commercial; law enforcement; insurance; re-identify	Attribution; ethics board approval; enhanced anonymisation

Patient-Level Consent Flag

In addition to dataset-level ODRL policies, 168 individual patient records carry an `odrl:hasPolicy` triple linking to `ehds-prototype.example.org/policy-patient-no-ai-training`, indicating withdrawal of consent for AI training purposes. This flag is distributed across cohorts and is queryable via SPARQL against the clinical named graphs.

MCP Connector

The connector at `/connector` exposes seven typed tools over SSE transport, compatible with Claude (claude.ai), DeepSeek, mcphost with Ollama, and the Python MCP SDK.

Tool	Description
`ehds_list_datasets`	List all datasets with HealthDCAT-AP metadata
`ehds_describe_dataset`	Full metadata for one dataset (input: `dataset_uri`)
`ehds_check_policy`	ODRL permissions, prohibitions, obligations (input: `dataset_uri`)
`ehds_search_datasets`	Keyword search over catalogue (input: `keyword`)
`ehds_get_patients`	Patient demographics (input: `dataset_uri`, `limit`)
`ehds_get_condition_stats`	Aggregate condition statistics (input: `dataset_uri`, `snomed_code`)
`ehds_query_clinical`	Free SPARQL over clinical data (input: `dataset_uri`, `sparql`)

To connect from Claude: add `https://mcp.linkeddata.es/connector` as an MCP connector in claude.ai settings. No installation is required.

Infrastructure

Component	Technology
Triple store	Apache Jena Fuseki 6, TDB2 persistent store
MCP server	Python, `mcp` SDK 1.27, SSE transport
Vector store	ChromaDB, `all-MiniLM-L6-v2` embeddings
Host	Ubuntu 24.04, 16 cores, 15 GB RAM (Horizon Europe funded VM)
Licence	CC-BY 4.0

FAIR Compliance

- **Findable** — HealthDCAT-AP catalogue with persistent URIs for all datasets and policies.
- **Accessible** — Open SPARQL 1.1 endpoint; downloadable ChromaDB vector store; no authentication required.
- **Interoperable** — Standard vocabularies throughout: FHIR R4, ODRL 2.2, SNOMED CT, LOINC, RxNorm, DCAT-AP 3.0.
- **Reusable** — CC-BY 4.0 licence; machine-readable ODRL policies; full provenance via HealthDCAT-AP metadata.

This document describes the resource as deployed at the time of ISWC 2025 submission. The portal is a living benchmark; cohorts and queries may be extended by the community following the instructions at `https://mcp.linkeddata.es/supplements`.