

Leveraging Content-based Metadata in Catalogs to Improve Discoverability in Data Spaces*

Adriana Morejón¹, Alberto Berenguer¹, Lucía de Espona¹, David Tomás¹, and Jose-Norberto Mazón¹

Department of Software and Computing Systems, University of Alicante, Carretera San Vicente del Raspeig s/n, 03690 San Vicente del Raspeig, Spain
adriana.morejon@ua.es (corresponding author), alberto.berenguer@ua.es, lucia.espona@ua.es, dtd@ua.es, jnmazon@ua.es

Abstract. Data spaces facilitate sovereignty-preserving data sharing within federated ecosystems; however, dataset discovery is often constrained by catalogs that rely primarily on high-level metadata standards such as DCAT-AP. These models typically lack structural and content-level details, which impedes fine-grained relevance assessment and complicates both internal dataset discovery and the federation of external open data into data spaces. This paper introduces cbmDCAT-AP, a backward-compatible extension of DCAT-AP that integrates content-based metadata as primary catalog elements. The proposed extension augments dataset descriptions with field-level semantics, and representative sample values, thereby enabling content-informed discovery without disclosing entire datasets. Furthermore, a clustering-based reduction technique is proposed to generate representative samples values through the use of word embeddings and K-means clustering. Finally, an end-to-end evaluation across two scenarios (internal dataset discovery and external open data selection) has been conducted to evaluate and validate our proposed extension of DCAT-AP.

Keywords: Data Space, Metadata, Catalog, DCAT-AP, Discoverability, Open Data

1. Introduction

Data spaces are emerging as federated infrastructures that integrate secure data exchange of data ecosystems with the economic incentives of data markets, while maintaining provider sovereignty [4, 41]. Grounded in the principles of interoperability, trust, and data sovereignty, data spaces represent a paradigm shift in business-to-business scenarios, where data can substantially enhance collaborative capabilities [12, 44].

Data spaces are commonly established using reference architectures such as the International Data Spaces Association (IDSA), Gaia-X, FIWARE, and iSHARE frameworks [24, 32]. These architectures are implemented through standardized connectors,

* This paper is an extended version of the conference paper “Improving Data Discovery Effectiveness: Experimental Evaluation of Content-Based Catalogs in Data Spaces” presented in the 7th Workshop on Modern Approaches in Data Engineering and Information System Design organized in conjunction with the conference ADBIS 2025.

including Eclipse Dataspace Components (EDC)¹ and the FIWARE Data Space Connector². Such connectors facilitate technical interoperability and enforce usage control and access policies, enabling providers to retain control while permitting consumers to discover and access data without compromising data sovereignty [25]. Also, the rollout of Common European Data Spaces is supported through EU-funded projects, including the Smart Open-source Middleware (SIMPL) programme, which provides an open-source, smart and secure middleware platform to foster interoperability and controlled data access across data spaces [18, 19].

In these reference architectures, connectors and middleware, data catalogs are a critical building block for two complementary discovery scenarios. First, consumers need to identify relevant datasets already offered within the data space by multiple sovereign providers, often under restrictive access policies and without the possibility of freely downloading or inspecting the full content. Second, data space participants increasingly aim to incorporate external open datasets (typically harvested from open data portals) in order to enrich the data space with contextual and baseline information, and unleashing the potential of open data within the same data space ecosystem. In this second scenario, the challenge is not only to locate open datasets, but to assess their fitness-for-use and to decide which ones should be onboarded and federated into the data space catalog with harmonized metadata.

Therefore, both scenarios require cataloguing datasets. Unfortunately, traditional data catalogs primarily depend on high-level descriptive metadata, such as title, license, keywords, and publisher, which are aligned with standards like the Data Catalog Vocabulary (DCAT) and its profiles (e.g., DCAT-AP) [20, 45]. While these metadata are effective for basic inventories, they consistently fail to communicate the practical utility of dataset content [35]. As a consequence, these dataset-centric metadata impedes consumers from comprehending internal structure and semantics, including field names, data types, column descriptions, and value ranges, as well as the specific semantics required to assess applicability for particular use cases [38]. Catalogs typically omit content evidence, forcing consumers to obtain full access (or to download large resources) merely to determine relevance. On one hand, in data spaces, this contradicts core principles of sovereignty and controlled disclosure [28]. This tension exemplifies Arrow's information paradox: exposing all data to demonstrate its value can compromise sovereignty and governance constraints, whereas concealing it prevents consumers from assessing practical utility [3]. On the other hand, in the open data scenario, it hinders dataset selection and onboarding decisions, discouraging the systematic federation of external datasets into the data space. Furthermore (and valid for both scenarios), empirical studies demonstrate that dataset discovery relying exclusively on high-level metadata yield low retrieval rates of relevant datasets and exhibit high abandonment rates due to insufficient semantic correspondence [6].

To address these limitations, we propose content-based catalogs through a DCAT-AP extension that incorporates structured content-level metadata describing dataset schema and semantics, as well as representative (sample) values, while remaining compatible with existing open data portals and data space catalogs. A central design principle is to treat dataset sampling as a first-class catalog artifact: compact, representative, and policy-aware samples provide evidence of content utility without requiring full disclosure or bulk

¹ <https://github.com/eclipse-edc>

² <https://github.com/FIWARE/data-space-connector>

transfer. Furthermore, we introduce an approach to extract representative samples suitable for inclusion in the catalog.

The proposed extension is integrated into the HELEADE data space³, a territorial data space deployed by the University of Alicante to collect, organise, and share environmental and socioeconomic data under interoperability and data sovereignty principles. Validation is performed by mapping the extension to the EDC data catalog, which underpins the HELEADE architecture, and by demonstrating how the same enriched metadata model supports both scenarios. First, discovery of sovereign datasets within the data space and, secondly, discovery and selection of external datasets from open data portals for subsequent cataloguing inside the data space. Furthermore, we introduce an approach for extracting representative samples to be included in the catalog.

While our previous work [39] introduced a dataset sampling methodology that utilizes word embeddings and clustering techniques to extract representative samples from large datasets within data spaces, the approach presented in this paper extends that foundation by incorporating the following contributions:

1. A novel DCAT-AP extension that introduces content-based metadata as first-class catalog elements, including new classes and properties for describing dataset structure and semantics (e.g., fields, types, descriptions) as well as representative sample values.
2. A practical mapping as adoption pathway for the proposed DCAT-AP extension in the Eclipse Dataspace Component (EDC) framework, validated through integration in the HELEADE data space catalog.
3. An end-to-end evaluation of discoverability improvements, by using our DCAT-AP extension, across two aforementioned scenarios: (i) dataset discovery within the data space, and (ii) discovery and selection of open datasets from external open data portals for cataloguing and reuse within the data space.

This paper is structured as follows. Section 2 reviews related work on data spaces, cataloguing standards, content-aware discovery, and sampling techniques. Section 3 introduces the DCAT-AP extension for content-level metadata, detailing new classes, properties, RDF implementation, and mapping to the EDC catalog. Section 4 presents the proposed solution for integrating the extension within federated data spaces. It examines the system architecture, content-based metadata integration, and the methodology for sample generation. Section 5 presents experiments that evaluate improvements in discoverability for sovereign datasets and open data onboarding. Section 6 concludes with a summary and suggestions for future research.

2. Related Work

This section reviews prior work relevant to our proposal from three complementary points of view: open data portals and their metadata publishing and harvesting practices, data space architectures and the role of connectors and catalog services in sovereignty-preserving discovery, as well as cataloguing standards and extensions, with emphasis on DCAT/DCAT-AP and existing profiles. We then summarize related work on content-aware dataset dis-

³ <https://dataspace.heleade.iuii.ua.es/>

covery and on sampling and preview generation for tabular data, highlighting the gap addressed by our DCAT-AP extension and sampling methodology.

2.1. Open Data Portals and Metadata Harvesting

Open data portals are online platforms where governments and organizations make public sector data available for access, use, and reuse. These portals serve as a central repository for datasets, providing a user-friendly way to discover, download, and analyze information. They play a crucial role in promoting transparency, innovation, and civic engagement by enabling citizens, researchers, and businesses to access a wide range of data [17].

Beyond human-facing search interfaces, open data portals increasingly support machine-to-machine interoperability through APIs and RDF-based metadata exports, enabling third parties to harvest and federate dataset descriptions across catalogs. In practice, however, cross-portal discovery and systematic reuse are hindered by heterogeneous modelling choices, inconsistent use of controlled vocabularies, and varying metadata completeness [2, 17]. As a result, dataset selection for downstream integration often requires manual inspection steps that do not scale, for example, when onboarding large numbers of external open datasets into a data space.

CKAN (Comprehensive Knowledge Archive Network)⁴ represents one of the most widely adopted open-source data portal platforms in the open data ecosystem [2]. Developed by the Open Knowledge Foundation and maintained by an active community, CKAN has been implemented by government agencies, research institutions, and organizations worldwide to publish and share their datasets. The platform implements a metadata schema that partially aligns with DCAT principles, offering a standardized approach to dataset description. In addition, CKAN's support for DCAT-AP is provided through the widely used extension `ckanext-dcat`, which enables CKAN-based portals to expose and consume metadata in RDF formats compliant with multiple DCAT-based application profiles. This extension offers pre-built CKAN schemas supporting DCAT-AP versions 1, 2, and 3, allowing data portals to align their internal metadata models with the European DCAT-AP standard while preserving domain-specific or national customizations via configuration-driven schemata (e.g., through `ckanext-scheming`)⁵. As a result, many national and regional open data portals built on CKAN already rely on DCAT-AP-aligned metadata exports to facilitate cross-portal discovery and interoperability within the European data ecosystem.

Many portals also provide lightweight content previews (e.g., showing the first rows of a table) or basic profiling indicators to assist users. While useful for quick inspection, these previews are typically neither standardized nor representative of the full dataset and may expose biased slices of the content (e.g., “head” rows only). Consequently, they provide limited support for automated relevance assessment and for onboarding decisions in federated environments (such as data spaces), motivating more principled and representative sampling mechanisms aligned with catalog metadata.

⁴ <https://ckan.org/>

⁵ <https://docs.ckan.org/projects/ckanext-dcat/en/latest/mapping/>

2.2. Data Space Architectures and Sovereignty-Preserving Discovery

A data space is a federated data infrastructure that supports trustworthy data sharing among data providers and data consumers [41]. To enable this, data spaces rely on connectors [25], which act as interfaces between participants and the ecosystem. Connectors ensure technical interoperability and enforce access policies, allowing providers to keep data sovereignty when sharing data while granting consumers seamless access without compromising control. Hence, data spaces enable new options for value creation where providers and consumers easily interact and work together to find, access, publish, consume, and reuse data, as well as to stimulate innovation [26].

In contrast to open data portals, discovery in data spaces is inherently constrained by policy: participants may advertise datasets and data services while restricting direct inspection of content until contractual, legal, or technical conditions are satisfied. Therefore, catalogs in data spaces must balance discoverability with controlled disclosure, exposing enough information to support relevance judgments while preserving sovereignty and minimizing unnecessary data transfer. This shifts the emphasis from purely descriptive metadata toward richer, yet governance-compatible, content evidence.

Numerous research initiatives have resulted in conceptual frameworks and reference architecture models to accelerate the development of data spaces [4].

In the European context, the deployment of Common European Data Spaces is a cornerstone of the EU data strategy, aiming to enable trustworthy cross-border and cross-sector data sharing and reuse [19]. To provide shared, reusable technical building blocks across these initiatives, the European Commission launched the Smart Open-source Middleware (SIMPL) programme, an open-source, smart and secure middleware intended to support interoperability and controlled data access across European data spaces and cloud-to-edge federations [18].

The i4Trust initiative emerges as a collaboration program, targeting the creation of data spaces by proposing an architecture based on commonly agreed building blocks [29]. They are a combination of components from FIWARE and iSHARE, two data space foundations. The iSHARE Foundation [34] maintains a trust framework for data spaces. The FIWARE Foundation [22] introduces an open source framework that promotes the development of interoperable, smart solutions.

Another innovative data space technical framework has been developed by Gaia-X, aiming to establish a trustworthy ecosystem where data is shared, maintaining the user's digital sovereignty of the data. This standard-based framework allows the implementation of distributed data systems in all European countries in a legally secure manner, enabling compliance with GDPR and other data regulations [24].

The International Data Spaces Association (IDSA) is a non-profit organization focusing on establishing and promoting standards for data spaces as trusted environments where organizations can share data while retaining full control over its use [33]. The IDS Reference Architecture Model (IDS-RAM) [32] materializes the standard developed from collecting requirements from various industries and the results gained from the model's implementation. Additionally, the IDS-RAM connectors fit the Gaia-X principles and architecture model [31].

The Eclipse Dataspace Components (EDC) [16] is a comprehensive framework (concept, architecture, code, samples) providing a basic set of features (functional and non-functional) that dataspace implementations can re-use and customize by leveraging the

framework's defined APIs and ensure interoperability by design. The EDC project is governed by the Eclipse Foundation and it is powered by the specifications of the Gaia-X AISBL Trust Framework and the IDSA Dataspace protocol.

A further recurring practice in deployed data spaces is the enrichment of the ecosystem with external open datasets (e.g., reference geodata, environmental baselines, administrative registries). In this onboarding scenario, operators and participants must identify candidate datasets in open portals and assess their fitness-for-use before federating them inside the data space with harmonized metadata. This reinforces the need for catalog mechanisms that support scalable, content-informed selection beyond high-level keywords.

One core element of both data space architectures and open data portals is the data catalog, which serves as a structured inventory of available datasets and associated metadata and access policies or usage licenses.

2.3. Cataloguing Standards for Data Discovery: DCAT and DCAT-AP

The Data Catalog Vocabulary (DCAT), a W3C recommendation, has emerged as the predominant metadata standard for data catalog implementation. Prior research has established DCAT's significant contribution to metadata interoperability across diverse data ecosystems, from open data portals utilizing CKAN to advanced data space architectures like those proposed by IDSA and Gaia-X. These architectures have operationalized DCAT through connector technologies such as EDC and the FIWARE Data Connector. Several publications [37] have highlighted DCAT's potential as a foundation for open data discovery in federated data spaces, given its widespread adoption and interoperability-focused design.

Unfortunately, current DCAT-AP-based catalogs are insufficient for effective data discovery, posing two key limitations:

- Limited metadata: existing data catalogs primarily rely on high-level metadata to describe datasets but lack detailed structural information of dataset content, such as field names or descriptions, which are essential for precise dataset discovery.
- Content blindness: current data catalogs exclude actual dataset content, relying solely on metadata. This prevents data spaces initiatives from gaining deeper insights into open data relevance before federating the full dataset.

As an attempt to enhance semantic interoperability across European data portals, the extension DCAT-AP⁶ emerges. Based on W3C's DCAT, it supports standardized dataset descriptions, enabling efficient data exchange and reuse. The basic use case for DCAT-AP is to enable cross-data portal search for datasets and make public sector data better searchable across borders and sectors. Although this extension improves data sharing among different sectors and includes additional support for access policies and legal aspects, the aforementioned two limitations regarding content description remain unresolved.

To overcome these challenges in data discovery, reference architectures for data spaces require enhancing data catalogs with content-based metadata. In response to these requirements, this work proposes an extension of DCAT-AP that integrates content-based metadata, while it is mapped to current EDC data catalog for the sake of compatibility.

⁶ <https://semiceu.github.io/DCAT-AP/releases/3.0.0/>.

2.4. Extending DCAT-AP: Sectoral and Domain Profiles

As this work involves the development of an extension to DCAT-AP, it is important to review previous extensions of this standard. DCAT-AP is recognized as the standard application profile for open data cataloguing in Europe, with multiple extensions developed to address national and sector-specific requirements. DCAT-AP serves as a foundational framework upon which multiple specialised extensions have been developed, establishing it as an extensible core for domain-specific and interoperability scenarios. One notable extension is StatDCAT-AP [13], a profile designed to represent statistical metadata within open data portals. StatDCAT-AP extends DCAT-AP to accommodate statistical concepts such as units of measurement, geographical and temporal coverage, and hierarchical series structures, while preserving compatibility with the European profile.

Within the health sector, HealthDCAT-AP [14] has been introduced as a targeted extension for describing health data and services in the context of the European Health Data Space (EHDS). This profile addresses the specific requirements of sensitive health data and health services, including coordination with access, consent, and anonymisation policies, while maintaining compatibility with DCAT-AP and preserving the integrity of the metadata ecosystem.

Alongside StatDCAT-AP and HealthDCAT-AP, GeoDCAT-AP [42] is among the most established extensions for describing geospatial datasets and services within open data environments. A technical report outlines its integration of geographic metadata vocabularies, such as INSPIRE, with DCAT-AP, thereby enhancing the description of spatial dimensions, resolution, projection, and geographic coverage.

Within data governance and administrative records, proposals such as BRegDCAT-AP⁷ have emerged as extensions of DCAT-AP for describing basic records and their associated services. This extension introduces specific properties to define record types, hierarchical structures, and relationships between entities. It maintains compatibility with the core DCAT-AP standard and enables the integration of record catalogs into open data infrastructures.

In scientific and research contexts, profiles such as EPOS-DCAT-AP have been developed to adapt DCAT-AP for describing Earth data infrastructure datasets and services. These profiles introduce properties related to data quality, traceability, and integration with disciplinary data models.

Overall, existing DCAT-AP extensions predominantly focus on aligning catalog metadata with sectoral concepts, services, governance constraints, or domain-specific semantics. While these profiles improve interoperability within their respective domains, they generally do not provide a generic, reusable model for content-level descriptions at the granularity required for schema-aware discovery (e.g., column-level semantics and representative values), nor do they treat representative samples as first-class catalog artefacts. This gap motivates our extension, which targets content evidence while preserving DCAT-AP compatibility.

⁷ <https://semiceu.github.io/uri.semic.eu-generated/BRegDCAT-AP/releases/3.0.0/>.

2.5. Content-Aware Dataset Discovery

A large body of research on dataset discovery in tabular corpora and data lakes explores how to retrieve relevant tables using signals derived from schema and content, including column naming patterns, statistical summaries, entity linking, and semantic representations (e.g., embedding-based encodings of columns or tables) [7, 8, 47]. These approaches demonstrate that content-derived signals can substantially improve relevance ranking compared to keyword-only metadata, especially when users search for datasets matching a particular structure or concept [15, 21].

However, most content-aware discovery pipelines implicitly assume that datasets (or substantial sketches of them) can be centrally accessed, indexed, and processed to build searchable representations [1]. This assumption is misaligned with the sovereignty and controlled-access constraints of data spaces [28, 41], where providers may be unwilling or unable to expose full content for indexing [5]. This motivates standardizing content-level evidence in catalogs through governance-aware metadata and representative samples that can be produced by providers and shared under explicit policies.

2.6. Sampling and Data Previews for Tabular Dataset Discovery

The initial theoretical task involves reconciling the extension of the DCAT-AP standard with updated information regarding dataset content. Furthermore, this paper addresses a significant limitation for data consumers: content opacity. This limitation compels consumers to access entire datasets for relevance assessment, which contradicts fundamental sovereignty principles. Accordingly, the research project examines practical methods for extracting data samples, enabling data consumers to evaluate a dataset using representative examples prior to determining its utility for their specific requirements.

Most existing approaches for data discovery assume direct access to dataset content for indexing and matching [1], an assumption incompatible with the sovereignty principles fundamental to data spaces. Furthermore, current techniques often struggle with semantic heterogeneity across domains, where the same concept may be expressed differently in various datasets. Scalability remains problematic in federated environments with thousands of distributed datasets, as content-based matching becomes computationally expensive. Additionally, most methods lack mechanisms to assess the trustworthiness and quality of discovered datasets, critical factors in a productive data space.

While using data samples for discovery rather than complete datasets improves computational efficiency, it introduces the critical challenge of generating representative samples that maintain both utility and privacy guarantees.

A variety of sample generation techniques have emerged to address this challenge. Random representative sampling remains a foundational method, providing broad coverage with minimal bias, though it may underrepresent rare patterns [30]. More advanced approaches include Virtual Sample Generation (VSG), which synthesizes data points through interpolation or neural methods to enrich sparse datasets [9]. Bootstrapping methods, popular in machine learning, generate multiple resampled datasets with replacement to support robust training and uncertainty estimation [27]. Meanwhile, synthetic data generation techniques using models such as GANs and variational autoencoders simulate realistic sample while preserving privacy, facilitating discovery in restricted-access

environments [10]. Other strategies include active sampling, where data is selectively retrieved based on query relevance or informativeness [43], and clustering-based sampling, which ensures semantic diversity by extracting representative elements from identified clusters [11]. Dimensionality reduction methods, such as PCA or autoencoder-based embeddings, also guide sampling by preserving the global structure of high-dimensional data [40]. Lastly, metadata-driven techniques leverage high-quality semantic annotations and quality metrics to inform targeted sample extraction [46].

However, these existing approaches exhibit critical limitations taking into account the context in which we want to use the information, i.e. a data discovery context. Methods like random, bootstrap, or synthetic sample generation, lack semantic awareness, often overlooking the underlying meaning or contextual relationships between data attributes. Moreover, methods based solely on metadata quality are ineffective in cases where metadata is sparse, inconsistent, or superficial.

2.7. Research Gap and Positioning

Prior work shows that DCAT/DCAT-AP and their extensions enable interoperable catalog publication, and that content-aware discovery methods can improve retrieval when rich metadata are available. Nevertheless, there remains a gap between (i) standards-based cataloguing practices that remain largely dataset-centric and (ii) discovery techniques that rely on direct access to content for indexing and matching. This gap is especially problematic in data spaces, where sovereignty constraints limit content exposure by design.

Our work addresses this gap by introducing a DCAT-AP extension that standardizes content-level metadata and representative samples as governed catalog artefacts, and by demonstrating practical adoption through mapping to the EDC catalog model. In doing so, we support both discovery scenarios considered in this paper: searching sovereign datasets within the data space and selecting external datasets from open data portals for systematic onboarding and federating.

3. DCAT-AP extension for content-level metadata

DCAT-AP extends DCAT to facilitate cross-portal discovery of public sector datasets throughout Europe by standardizing high-level descriptive, administrative, and structural metadata, including titles, publishers, licenses, and themes. Nevertheless, DCAT-AP retains DCAT’s dataset-level orientation and lacks native mechanisms for representing field-level structure or sample content, thereby constraining its effectiveness for fine-grained discovery within data spaces. To address this limitation, DCAT-AP is extended with content-based metadata, the cbmDCAT-AP extension.

Our cbmDCAT-AP extension preserves the semantics of both DCAT-AP and the underlying DCAT vocabulary. It introduces additional classes and properties that may be used in conjunction with standard DCAT-AP terms. This approach maintains backward compatibility with existing DCAT-AP catalogs and allows data providers to present more detailed structural, content-level information and data samples about their datasets. Comprehensive documentation has been established detailing the specifications of the cbmDCAT-AP extension⁸.

⁸ <https://amorejon6.github.io/cbmDCAT-AP/>

3.1. Scope and objectives

The cbmDCAT-AP extension is intended for catalog publishers and data space operators that already expose their metadata using DCAT-AP, while requiring support for dataset discovery beyond traditional metadata (i.e., content metadata, such as fields, types, sample values). Importantly, our cbmDCAT-AP extension keeps compatibility with existing DCAT-AP harvesting and validation workflows.

Specifically, cbmDCAT-AP extension is designed for structured datasets (such as tabular datasets) and supports the representation of both data dictionaries tailored to specific data distributions (which contain field-level metadata), as well as samples of datasets (which are small subsets of instances derived from the original dataset).

Support for additional data models, such as spatial, unstructured, or multimedia, is beyond the scope of this version and will be considered in future extensions.

3.2. New classes and properties

Figure 1 illustrates the cbmDCAT-AP meta-model, which integrates elements for DCAT/DCAT-AP, Dublin Core Terms, and Schema.org, along with new components introduced by this extension. The reused elements maintain their original semantics, and cbmDCAT-AP defines their application in representing the structure and representative content of tabular datasets.

The primary DCAT and DCAT-AP classes reused in this context are those related to dataset publication and access. The class *dcat:Catalog* denotes a collection of catalogued resources, which may include datasets and data services. The class *dcat:Dataset* refers to the conceptual collection of data described, regardless of serialization or access method. Each dataset can be linked to one or more *dcat:Distribution* instances, where each instance specifies a particular accessible form of the dataset, such as a CSV file, a JSON resource, or another technical format. The class *dcat:DataService* identifies a service or endpoint that enables access to one or more datasets.

The relationships adopted from DCAT maintain their original interpretations. *dcat:distribution* connects a dataset to one of its accessible representations. *dcat:accessService* links a distribution to the service that provides access, while *dcat:servesDataset* specifies the datasets made available by a data service. Additionally, catalog relationships associate *dcat:Catalog* with the datasets and services it encompasses.

The model incorporates descriptive properties from established vocabularies. *dcterms:title* and *dcterms:description* supply a human-readable name and explanation for a resource, whereas *dcterms:type* specifies its type. For field descriptions, *schema:name* designates the technical field name as it appears in the underlying distribution. All other reused properties maintain the semantics established by their source vocabularies.

On top of these standard elements, cbmDCAT-AP introduces the following classes:

- **cbm:Sample.** A limited set of representative records derived from a dataset and exposed to support preliminary relevance assessment without requiring access to the complete dataset.
- **cbm:DataDictionary.** A structured description of the fields contained in a particular dataset distribution.

- **cbm:Field**. A description of an individual variable or attribute, such as a column in a tabular distribution. A field may include its technical name, human-readable title, datatype, description, and representative values.

The proposed model differentiates between dataset-level and distribution-level information. A sample is linked to *dcat:Dataset* because it offers representative evidence regarding the conceptual dataset under evaluation. Conversely, a data dictionary is linked to *dcat:Distribution*, as various distributions of the same dataset may present distinct schemas, field names, datatypes, formats, or serializations. Corresponding properties of the CBM extension are described in Table 1.

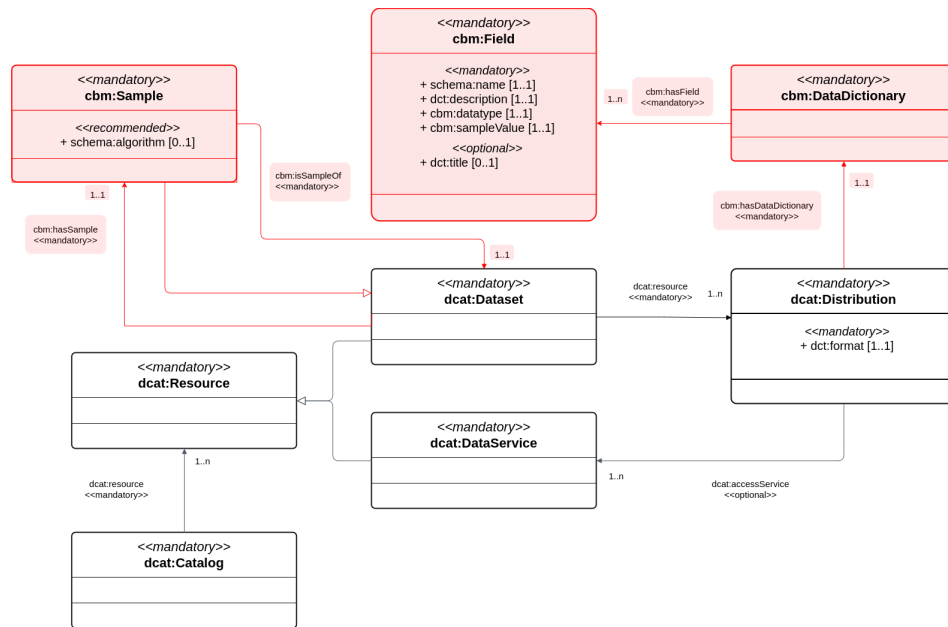


Fig. 1. Schema of the cbmDCAT-AP extension.

3.3. RDF implementation of cbmDCAT-AP extension

The cbmDCAT-AP extension is defined as an RDF schema using Turtle syntax, as well as a sample application to describe a dataset as described next.

```

@prefix dcat: <http://www.w3.org/ns/dcat#> .
@prefix cbm: <http://example.org/cbm#> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix schema: <http://schema.org/> .
@prefix ex: <http://example.org/resource/> .

ex:urban-biodiversity-dataset a dcat:Dataset ;

```

```

@prefix dcat: <http://www.w3.org/ns/dcat#> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix schema: <http://schema.org/> .
@prefix cbm: <http://example.org/cbm#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .

cbm:Sample a rdfs:Class ;
  rdfs:subClassOf rdf:Resource ;
  rdfs:label "Sample" ;
  rdfs:comment "Samples from a dataset." .

cbm:DataDictionary a rdfs:Class ;
  rdfs:subClassOf rdf:Resource ;
  rdfs:label "Data Dictionary" ;
  rdfs:comment "Describes fields in a dataset distribution." .

cbm:Field a rdfs:Class ;
  rdfs:subClassOf rdf:Resource ;
  rdfs:label "Field" ;
  rdfs:comment "A variable within a dataset (e.g., a column in
    a CSV file)." .

cbm:isSampleOf a rdf:Property ;
  rdfs:domain cbm:Sample ;
  rdfs:range dcat:Dataset ;
  rdfs:label "is sample of" .

cbm:hasSample a rdf:Property ;
  rdfs:domain dcat:Dataset ;
  rdfs:range cbm:Sample ;
  rdfs:label "has sample" .

cbm:hasDataDictionary a rdf:Property ;
  rdfs:domain dcat:Distribution ;
  rdfs:range cbm:DataDictionary ;
  rdfs:label "has data dictionary" .

cbm:hasField a rdf:Property ;
  rdfs:domain cbm:DataDictionary ;
  rdfs:range cbm:Field ;
  rdfs:label "has field" .

cbm:sampleValue a rdf:Property ;
  rdfs:domain cbm:Field ;
  rdfs:range rdf:Literal ;
  rdfs:label "sample value" .

```

Listing 1.1. Vocabulary for DCAT Extension

Table 1. Properties of the DCAT extension for content-based metadata.

Property	Domain	Range	Description
cbm:isSampleOf	cbm:Sample	dcat:Dataset	Relates example or sample data to a dataset.
cbm:hasSample	dcat:Dataset	cbm:Sample	Relates a dataset to example or sample data.
cbm:hasDataDictionary	dcat:Distribution	cbm:DataDictionary	Links a distribution to its structural metadata.
cbm:hasField	cbm:DataDictionary	cbm:Field	Lists the fields described by the data dictionary.
cbm:sampleValue	cbm:Field	Literal	Provides a representative value for the field.

```

dcterms:title "Urban Biodiversity Survey" ;
cbm:hasSample ex:biodiversity-sample ;
dcat:distribution ex:biodiversity-distribution .

ex:biodiversity-sample a cbm:Sample ;
schema:description "Example data from 3 surveyed gardens." ;
schema:algorithm "Hierarchical Clustering" ;
cbm:isSampleOf ex:urban-biodiversity-dataset.

ex:biodiversity-distribution a dcat:Distribution ;
dcterms:format "CSV" ;
cbm:hasDataDictionary ex:bio-data-dictionary .

ex:bio-data-dictionary a cbm:DataDictionary ;
cbm:hasField ex:field-species, ex:field-abundance .

ex:field-species a cbm:Field ;
schema:name "species" ;
dcterms:title "Species Name" ;
dcterms:type "string" ;
dcterms:description "Scientific name of observed species" ;
cbm:sampleValue "Bombus terrestris" .

ex:field-abundance a cbm:Field ;
schema:name "abundance" ;
dcterms:title "Abundance" ;
dcterms:type "integer" ;
dcterms:description "Count of individuals observed" ;
cbm:sampleValue "7" .

```

Listing 1.2. Example of Extended Dataset Metadata

3.4. Mapping DCAT-AP extension to Existing Catalog Architectures

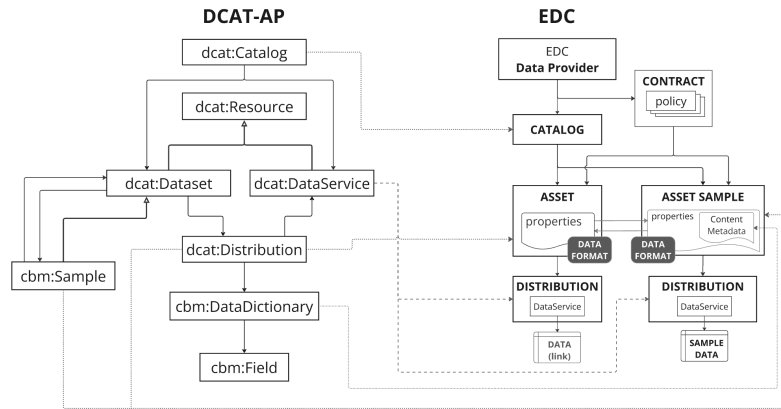
The proposed cbmDCAT-AP extension is independent of any specific data-space implementation. While this work uses Eclipse Dataspace Components (EDC) as its architectural basis, the extension can also be applied to other catalog and connector models. This section outlines its alignment with EDC, the International Data Spaces (IDSA), and CKAN. These mappings demonstrate how the proposed classes and properties can be integrated into diverse data ecosystems without significant architectural changes.

Figure 2 summarizes the correspondence between cbmDCAT-AP and the native entities of these three technologies. The mappings address the catalog, dataset, distribution, access service, representative samples, data dictionary, and field-level metadata concepts.

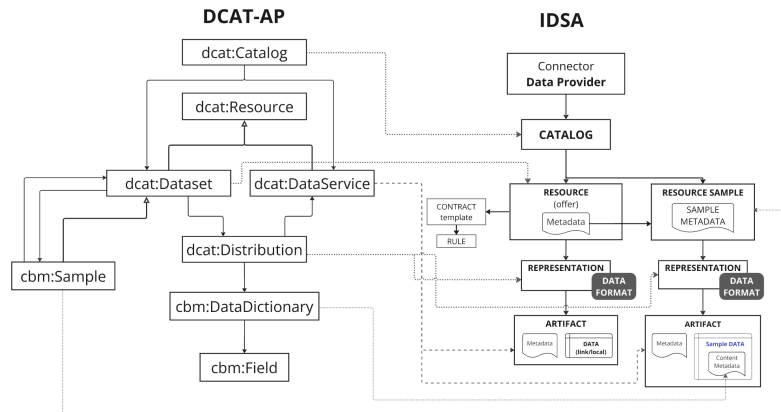
As shown in Figure 2a, the DCAT-AP extension entities can be mapped to the Eclipse Dataspace Components (EDC) [16] in a straightforward way. An EDC Data Provider has a "catalog" that corresponds to the *dcat:Catalog* class. This catalog contains multiple "assets" whose access is defined by the policies in the contract. Since this authorization is defined at the asset level, it is necessary to have the original and sample data as separate assets, corresponding to two different DCAT datasets: the original one with class *dcat:Dataset* and the sample one mapped to a *cbm:Sample* instance with open access policies. Since in EDC the data format is defined at the asset level, there is only one *dcat:Distribution* instance per dataset, and we can store the content metadata, equivalent to the *cbm:DataDictionary* within the asset metadata. The "distribution" entity in EDC is equivalent to the *dcat:DataService* and defines the data access methods.

The data model from the IDSA data provider and its correspondence to the DCAT-AP extension classes are depicted in Figure 2b. The "catalog" provided by an IDSA data provider aligns with *dcat:Catalog*, and the resources offered within the catalog correspond to *dcat:Dataset*. Access to these resources is regulated by contract templates and usage rules. Similar to EDC, access conditions are specified at the resource level. If the original data and the representative sample require distinct access policies, they may be modeled as separate resources. The protected resource corresponds to the original *dcat:Dataset*, whereas the resource providing access to the representative data corresponds to *cbm:Sample*. An IDSA resource may possess multiple representations, each associated with a specific data format. These representations correspond to *dcat:Distribution*. The artifacts linked to a representation specify the actual data object or access endpoint and are thus related to the function of *dcat:DataService*. The representative sample may be implemented as an artifact containing locally stored sample data or as an artifact referencing an external access endpoint. The associated content-level metadata, represented by *cbm:DataDictionary* and *cbm:Field*, can be stored alongside the artifact or resource metadata. This enables consumers to retrieve field descriptions and representative values when examining the sample resource, while the original data remain subject to the applicable contract rules.

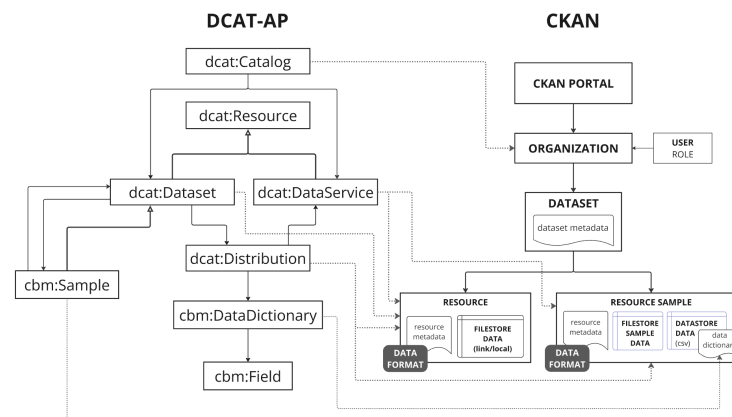
CKAN necessitates a distinct mapping, as it primarily functions as an open-data portal with datasets generally available for public read access. The mapping to CKAN can be seen in Figure 2c. The most analogous entity to *dcat:Catalog* is a CKAN organization, given that each dataset is linked to an organization responsible for its publication and metadata management. A CKAN dataset corresponds to the catalog record that describes the data resource. Both the original data and the representative sample may be published as distinct CKAN resources within the same dataset record. In contrast to the



(a) Mapping of the DCAT-AP extension to the EDC data model.



(b) Mapping of the DCAT-AP extension to the IDSA data model.



(c) Mapping of the DCAT-AP extension to the CKAN data model.

Fig. 2. Mapping of the DCAT-AP extension EDC, IDSA, and CKAN data models.

EDC and IDSA mappings, these resources do not require representation as separately governed datasets, since both may be publicly accessible under the license of the associated open dataset. A CKAN resource specifies its format and access location, and may either contain locally hosted data or reference an external URL. As such, it encompasses aspects corresponding to *dcat:Distribution* and *dcat:DataService*. A resource containing the original data serves as the accessible distribution of the dataset, while an additional resource containing representative rows can be linked to *cbm:Sample*. The sample resource may be stored in the CKAN filestore. For supported tabular formats, such as CSV, it can also be imported into the CKAN datastore, which offers structured storage, filtering, querying, and preview capabilities. The datastore's data-dictionary mechanism serves as an appropriate implementation location for *cbm:DataDictionary* and *cbm:Field*, as it already accommodates metadata including field names, types, and descriptions. Representative values may be provided through *cbm:sampleValue* or derived from the associated sample resource.

4. Including cbmDCAT-AP in data spaces for improving discoverability

This section introduces a solution to enhance dataset discovery in federated dataspace by integrating cbmDCAT-AP for enriched metadata and representative data sampling. The proposed approach combines architectural design, semantic enrichment, and content-aware reduction techniques to facilitate precise and efficient data search while preserving data sovereignty. The overall data space architecture and its integration with cbmDCAT-AP in a federated environment are first described. The incorporation of content-based metadata at both resource and data levels is then detailed, followed by the methodology for generating representative data samples using clustering-based reduction.

4.1. Data Space Architecture

Figure 3 presents the HELEADE dataspace architecture that implements the cbmDCAT-AP extension within a federated environment, adhering to the principles of the International Data Spaces Association and utilizing Eclipse Dataspace Components. This architecture aims to improve dataset discoverability while maintaining strict data sovereignty, enabling providers to share enriched, content-level metadata and representative samples without providing full access to entire datasets. The system incorporates catalog enrichment, semantic indexing, and policy-based access control into an integrated workflow that facilitates data discovery as an initial phase of federation.

On the provider side, each participant operates an EDC Connector that manages local datasets as assets. The original dataset is registered as a protected asset governed by contractual policies regulating access and usage. Simultaneously, a reduced version of the dataset, generated using the clustering-based sampling methodology described below, is registered as a separate asset. This reduced asset contains a limited number of representative rows and is generally associated with more permissive access policies. Both the original dataset and its sample are described using DCAT-AP metadata, further enriched with the cbmDCAT-AP extension, which provides field-level descriptions, data dictionaries, and representative sample values. This distinction between the full dataset and its sample allows for preliminary relevance evaluation while maintaining provider control.

The catalog layer is central to the architecture. Traditional dataset-level metadata are enhanced with structural and content-based descriptors, such as explicit links between datasets and their samples, structured data dictionaries for field descriptions, and representative values that provide semantic context. Despite these enhancements, full compatibility with DCAT-AP harvesting mechanisms is maintained, ensuring interoperability with existing portals and dataspace infrastructures. The extension is additive and backward compatible, enabling gradual adoption without disrupting established metadata ecosystems.

A semantic indexing component processes the available textual and structural information, including dataset content (either original or reduced), column names, and column descriptions when available. These elements are transformed into vector embeddings to enable similarity-based retrieval. To ensure scalability and performance, indexing is primarily performed over the reduced samples rather than the full datasets. This significantly decreases computational costs and indexing time while maintaining competitive retrieval quality, as demonstrated in the experimental evaluation. The architecture therefore balances representativeness and efficiency, which is essential in large-scale federated environments.

From the consumer perspective, interaction with the HELEADE dataspace begins with federated catalog queries or semantic search requests. The search engine retrieves candidate datasets based on embedding similarity, utilizing the enriched content-level metadata. Consumers can review descriptive metadata, field information, and representative sample values before determining dataset suitability. If a dataset is considered relevant, the consumer initiates a contract negotiation process through the EDC framework to request controlled access to the original asset. Access is granted only under agreed policies, thereby preserving sovereignty and compliance requirements.

Within the proposed architecture, dataset discovery, authorization, and data consumption are treated as distinct stages. The enriched catalog and representative sample enable consumers to evaluate the suitability of a dataset. However, these elements do not automatically provide access to the protected asset. After selecting a relevant dataset, the consumer initiates contract negotiation via the EDC connector. The resulting contract agreement delineates the conditions under which the consumer is authorized to access and utilize the asset, in accordance with the provider's defined policies.

The contract agreement operates independently of the technical delivery mechanism. Consequently, the proposed approach does not require datasets to be downloaded as files. Depending on the distribution and access services configured by the provider, data may be delivered as a downloadable resource, accessed via an API or query endpoint, or consumed through a streaming interface. These access mechanisms are represented by the standard DCAT/DCAT-AP concepts of *dcat:Distribution*, *dcat:DataService*, and *dcat:accessService*. In contrast, cbmDCAT-AP supplies the content-level information necessary to assess dataset relevance prior to requesting access. The representative sample is managed as a distinct asset and may be subject to a more permissive access policy than the original dataset. Therefore, access to the sample does not confer access to the complete dataset. Rather, the sample enables customers to make informed decisions prior to initiating contract negotiation for the protected asset.

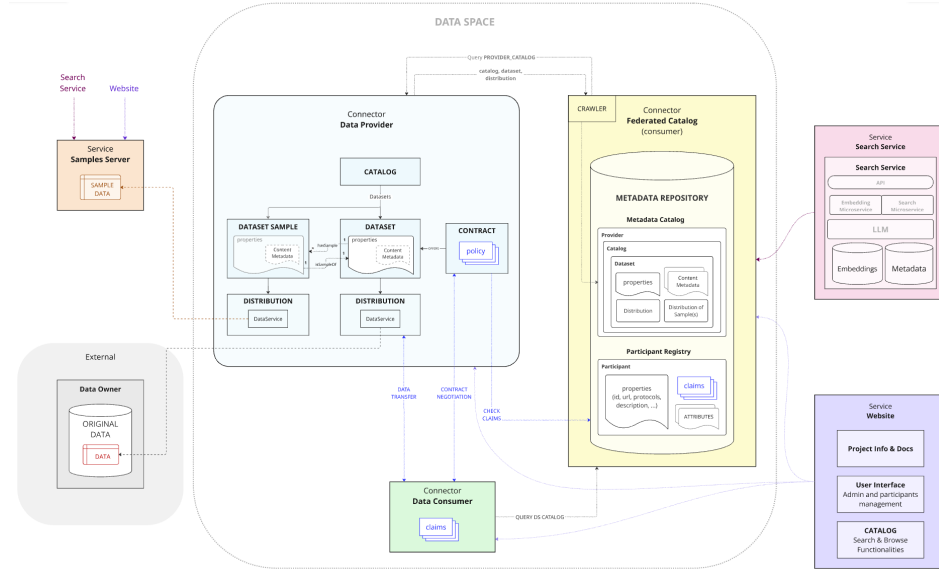


Fig. 3. HELEADE dataspace architecture.

4.2. Content-based Metadata

Our solution integrates content-based metadata at both resource and data content levels. At the resource level, metadata follows the IDS standard with fields like title, description, and license added by data providers. This is automatically enhanced with content-based metadata derived from the data itself, formatted using the Datapackage standard⁹ from the Frictionless Data project [23], and stored locally at the artifact level alongside a representative data sample. When consumers request the sample resource, they simultaneously receive this enriched metadata. A data sampling mechanism is also integrated in order to obtain the samples. This methodology is described below.

4.3. Clustering-based Reduction for Sample Generation

For tabular datasets, we propose a clustering-based reduction method for generating compact samples that retain semantically diverse and representative rows from the original table. The method is evaluated against two baselines: random row sampling and sequential selection of the first ten rows.

Given a target sample size k , each eligible table row is transformed into a numerical representation using word embeddings. K-means clustering is then applied to partition the resulting vector space into k groups of semantically similar rows. For each cluster, the original row whose vector is closest to the cluster centroid is selected. The selected rows approximate the central tendencies of their respective clusters while preserving the original tabular representation. The output is therefore a reduced table containing k actual rows from the source dataset, rather than synthetic centroid values.

⁹ <https://datapackage.org/standard/data-package/>

The use of actual rows facilitates interpretation by prospective data consumers, while the clustering process promotes coverage of different regions of the semantic space. However, selecting actual rows also means that the reduction method must not be interpreted as an anonymisation or privacy mechanism.

The clustering-based reduction method operates exclusively on data authorized by the provider for sample publication. The clustering algorithm does not assess whether a value is sensitive, restricted, or context-dependent. Instead, sample generation is preceded by a policy-aware preprocessing stage established by the data provider.

In this stage, fields are classified as publishable, conditionally publishable, or excluded for the sample. Excluded fields are removed prior to the generation of embeddings and clusters. Conditionally publishable fields may undergo transformation through masking, generalisation, aggregation, or suppression. For instance, exact geographic coordinates can be replaced with administrative areas, timestamps can be reduced to month-level granularity, and direct identifiers such as names, email addresses, or identification numbers can be removed. Clustering is then performed on the filtered and transformed data, ensuring that prohibited values neither influence nor appear in the generated sample.

Once representative rows are selected, the resulting sample is validated against the same publication policy before registration as a separate catalog asset. Consequently, the publication of a representative sample remains under the provider's control and is independent of the access policy governing the complete dataset. If a policy-compliant sample cannot be produced without incurring unacceptable disclosure risk, the provider may choose to publish only the data dictionary and omit sample values.

The current implementation assumes that the provider specifies the applicable field-level restrictions. Automatic identification of sensitive and context-dependent values constitutes a domain-specific challenge that may require semantic annotations, personal-data detection techniques, and risk assessment mechanisms. Automated inference and enforcement of such restrictions are beyond the scope of this work and will be addressed in future research.

To illustrate how the algorithm works, consider an environmental dataset containing observations of protected species with the fields *species_name*, *observation_date*, *latitude*, *longitude*, *municipality*, *habitat_type*, and *observer_email*. The provider's publication policy classifies *observer_email* as prohibited and exact coordinates as restricted because they may reveal the location of endangered species. Before sample generation, *observer_email* is removed, *latitude* and *longitude* are replaced by *municipality*, and *observation_date* is generalised from an exact date to year and month.

Embeddings are then generated from the remaining policy-compliant fields. K-means partitions the rows into k groups according to their semantic similarity, and the row closest to each cluster centroid is selected as a representative observation. For $k = 3$, the resulting sample may contain representative records corresponding to three combinations of species, habitat, municipality, and observation period. The published rows contain no observer identifier or exact location. The sample therefore illustrates the dataset's semantic coverage without exposing the restricted values used in the original resource.

4.4. Scenario-based architectural trade-off analysis

To complement the component-level experiments, we conducted a lightweight scenario-based architectural evaluation inspired by the Architecture Tradeoff Analysis Method

(ATAM) [36]. ATAM relates architectural decisions to concrete quality-attribute scenarios and uses this relationship to identify sensitivity points, trade-off points, and risks. The present analysis is deliberately scoped to the business drivers and empirical evidence available from the HELEADE data space architecture implementation. It was conducted by the authors and did not include the independent facilitation and multi-stakeholder scenario elicitation and prioritisation of a complete ATAM evaluation, thus it should be interpreted as an ATAM-inspired analysis rather than as a full ATAM analysis.

Two main business drivers guide the architecture: (i) enabling consumers to assess dataset relevance before obtaining access to the complete asset, while preserving provider sovereignty; and (ii) supporting scalable catalog ingestion and semantic index updates in data-intensive federated environments. Table 2 maps these drivers to quality-attribute scenarios, architectural decisions, available evidence, and their associated compromises.

The analysis identifies five principal sensitivity points: (i) the number of representative rows k , (ii) the availability and quality of field descriptions, (iii) the transformations applied during policy-aware preprocessing, (iv) the number and update rate of cataloged datasets, as well as (v) the embedding model used by the semantic index.

The first of these is also a major trade-off point: increasing k may improve semantic coverage and retrieval effectiveness, but also increases sample-generation and indexing costs and enlarges the disclosure surface. Similarly, richer field descriptions can improve retrieval but require additional curation and storage, while more restrictive masking or generalisation reduces disclosure risk at the possible expense of semantic fidelity.

Other residual risks include “sample staleness” after updating source datasets, the omission of rare but relevant values when rows close to cluster centroids are selected, and dependence on provider-supplied field classifications. Conversely, the additive nature of cbmDCAT-AP and the policy separation between sample and original assets constitute architectural mitigations for compatibility and unauthorised access, respectively. Overall, the architecture should be understood as a configurable compromise among retrieval effectiveness, indexing cost, interoperability, and controlled disclosure rather than as a configuration that maximises all quality attributes simultaneously.

5. Experimental evaluation

In this section, we describe the comprehensive evaluation of our proposed cluster through intrinsic, extrinsic, and temporal assessments. The intrinsic assessment compared original tables with their reduced counterparts to measure information preservation. The extrinsic assessment evaluated practical utility by employing reduced tables in information retrieval tasks within our proposed data space architecture, serving as a data discovery process. We benchmarked our cluster reduction technique against simpler methods such as random row selection and first rows selection. Additionally, we analyzed indexing performance differences between original and reduced datasets. All these tests are performed in the environment of the HELEADE data space architecture described in Section 4.1.

5.1. Evaluation Dataset

Two datasets have been selected for the experiments, in accordance with the two scenarios proposed: (i) dataset discovery within the data space, and (ii) discovery and selection of

Table 2. ATAM-inspired analysis of architectural quality attributes and trade-offs.

Quality attribute	Scenario and desired response	Architectural decision	Evidence and corresponding compromise
Discoverability and retrieval effectiveness	A consumer who cannot inspect the protected asset submits a content-oriented query, and relevant datasets should appear among the highest-ranked results.	Field-level metadata, representative samples, and embedding-based semantic indexing are incorporated into the catalog workflow.	Context 3 (see Section 5) generally provides the strongest retrieval effectiveness among the reduced configurations. Nevertheless, full-content indexing remains more effective, particularly for early-rank metrics, and field names without descriptions may introduce noise.
Indexing performance and scalability	A large number of datasets is added or updated, requiring the semantic index to be refreshed within an operationally acceptable time and computational budget.	The index is constructed from a fixed-size representative sample rather than from every row of the dataset.	End-to-end index construction is approximately $184\times$ faster for TData and $156\times$ faster for EnviDat when sample generation is included (see Section 5). This improvement is accompanied by a measurable loss of retrieval effectiveness and depends on the selected sample size.
Sovereignty and disclosure control	A consumer without a contract requests information for assessing dataset relevance but must not gain access to the protected asset.	The original dataset and its sample are registered as separately governed assets; sample generation is preceded and followed by policy-aware filtering and validation.	Access to the sample does not confer access to the original dataset. However, representative rows are actual records and do not constitute an anonymisation mechanism. Incorrect field classification or insufficient transformation may therefore create residual disclosure risks.
Interoperability and modifiability	Existing DCAT-AP harvesters and alternative catalog technologies must continue to process the catalog as the extension evolves.	cbmDCAT-AP is additive and backward compatible, with explicit mappings to EDC, IDSA, and CKAN.	Existing consumers can ignore the additional terms and continue processing standard DCAT-AP metadata. Conversely, exploiting the enriched metadata requires extension-aware consumers and platform-specific mappings, while separately governed sample assets increase lifecycle-management complexity.
Catalog freshness and availability	A provider updates a dataset or sample, and consumers should receive sufficiently current search results without accessing the provider's complete data.	Metadata and samples are harvested into a federated catalog and processed by a separate semantic indexing service.	Replication enables low-latency cross-provider search without centralising complete datasets. However, crawler and index-update intervals may produce stale descriptions, and the central search service can become a performance or availability bottleneck. These properties were not measured in the present evaluation.
Modifiability of semantic retrieval	A new domain, language, or embedding model must be supported without changing policy enforcement or the underlying asset model.	Semantic indexing is implemented as a service separated from the connector, contract, and metadata-representation components.	The separation localises changes to the retrieval subsystem. Nevertheless, replacing an embedding model requires re-indexing and may alter ranking behaviour, making model selection another source of performance variability.

open datasets from external open data portals for cataloguing and reuse within the data space.

For the first scenario, we selected a real-world dataset provided by a data space provider in the experiments conducted to evaluate the effectiveness of our table reduction methodology. The dataset comprises 100 tables extracted from a tourism open data portal from the Valencia Region in Spain, TData¹⁰. This collection was carefully selected to encompass a wide array of data structures and column types, providing a robust testing ground for assessing the adaptability and performance of the reduction approach.

For the second scenario, the real-world dataset comprises 242 tables extracted from an environmental research open data portal from Switzerland, EnviDat.¹¹ This collection was carefully selected to encompass a wide array of data structures and column types, providing a robust testing ground for assessing the adaptability and performance of the extension approach.

Table 3 offers an overview of the statistics of the evaluation datasets.

Characteristic	TData	EnviDat
Number of tables	100	242
Total number of rows	748,088	1,184,256
Total number of columns	1,926	5,946
Total number of text columns	1637	1,860
Avg. number of rows	7,480.88	4,893.62
Avg. number of columns	19.26	24.57
Avg. number of text columns	16.37	7.69
Max. number of rows	90,792	15,000
Max. number of columns	44	471
Max. number of text columns	43	469

Table 3. Statistics of the collected datasets.

5.2. Intrinsic Assessment

To evaluate the effectiveness of the proposed table reduction approach, a comparison was conducted between the original tables and their reduced versions, using descriptive statistics as the basis for similarity assessment. This was an indicator of how well the reduced tables retained the statistical properties of the original data.

Seven descriptive statistics were calculated along each column: the mean, median, standard deviation, minimum, maximum, as well as 25th and 75th percentiles. These seven metrics were calculated for each of the columns, and the resulting values are stored in a vector representing each table, thus allowing us to quantitatively compare tables. The similarity between the original and reduced tables was determined by calculating the cosine similarity between these vectors, providing a metric of alignment between the two tables.

¹⁰ <https://tdata.dlsi.ua.es/>

¹¹ <https://www.envidat.ch/>

We evaluated the table reduction approach across three distinct methods to benchmark its performance. In the first method, we applied the cluster reduction methodology. In the second method, we created a reduced table by randomly sampling rows from the original dataset. In the third method, we simply selected the first ten rows from the initial table, representing a sequential selection without methodological design.

For all three methods, we generated reduced tables with a standardized 10-row count. Cosine similarity analysis yielded matching scores of 0.933 (random reduction), 0.935 (first ten rows), and 0.934 (cluster reduction) for TData dataset, and 0.912 (random reduction), 0.911 (first ten rows), and 0.915 (cluster reduction). These nearly identical scores suggest our proposed reduction approach offers no significant advantage over simpler random or sequential selection methods when preserving statistical features from the original dataset. The minimal variation in similarity scores (0.003 maximum difference) indicates that further research is needed to develop methods that better preserve distinctive characteristics across original tables.

5.3. Extrinsic Assessment

To evaluate the table reduction utility for data discovery services, we conducted experiments comparing the performance of the tables against the original datasets using embeddings generated by the JinaAI¹² Spanish model, specifically the *jina-embeddings-v2-base-es* variant. This systematic comparison allowed us to thoroughly assess how well reduced tables perform relative to original datasets in information retrieval tasks.

The experimentation was carried out by first selecting randomly 10 rows from each original table to serve as a query in the experiments. Next, embeddings were created for both the original dataset and its three reduced versions as described in Section 4.3. Finally, search metrics were calculated using the queries against the original and sample tables.

We established three search contexts by generating indexes from distinct metadata components in our proposed content-based catalog. In *Context 1*, indexes incorporated only the table content. *Context 2* expanded this approach by including both table content and column names. *Context 3*, applicable to tables with available column descriptions, generated indexes using a comprehensive combination of column names, column descriptions and table content.

To evaluate retrieval quality, we employed two metrics: Mean Reciprocal Rank (MRR), which measures the position of the first relevant retrieved result, and Precision at k (P@k metric), which calculates the proportion of relevant results among the top k retrievals. These metrics provide an objective assessment of each approach's performance. In our results, we highlight statistically significant differences between the random and clustering reduction methods with asterisks, as these two approaches emerged as the most competitive. In the results tables, light shading indicates a statistically significant difference between the Cluster and Random reduction methods at $p < 0.05$, while dark shading indicates significance at $p < 0.01$.

Table 4 shows the search metrics results using reduction techniques in the three different contexts for the TData dataset. The results reveal interesting trends and variations across contexts and methods.

¹² <https://jina.ai/embeddings/>.

In Context 1, Cluster reduction and Random reduction produce comparable results, both slightly below those of the original dataset yet clearly outperforming First rows reduction. First rows reduction demonstrates the weakest performance, with an MRR of 0.6392 and a P@1 of 0.5426, underscoring its limitations in this setting. In Context 2, the Cluster approach yields the highest reduction scores, surpassing both Random reduction and First rows reduction. Notably, First rows reduction again performs the weakest, with an MRR of 0.6311 and a P@1 of 0.5362, revealing a distinct performance gap relative to Cluster reduction. In Context 3, Cluster reduction attains the highest MRR and P@1 scores, at 0.7856 and 0.7043, respectively, demonstrating strong effectiveness. Both Random reduction and First rows reduction also exhibit substantial improvements in Context 3 compared to the previous contexts. Although their performance remains slightly below that of Cluster reduction, both methods achieve high results, with MRR scores of 0.7508 and 0.6720, respectively.

Overall, the Cluster reduction achieves the highest scores among the reduction methods in most evaluated settings. However, the improvements over Random reduction are generally modest and are not statistically significant for every metric or context. The results therefore indicate a context-dependent advantage rather than a systematic superiority of clustering over random sampling.

Table. 4. Search metrics combining different contexts and reduction techniques for the TData dataset.

	Reduction	MRR	P@1	P@5	P@10
Context 1	Original data	0.8875	0.8340	0.9919	0.9979
	Cluster	0.7327	0.6521	0.8415	0.9511
	Random	0.7100	0.6351	0.8043	0.9191
	First rows	0.6392	0.5426	0.7723	0.8766
Context 2	Original data	0.7304	0.6479	0.8777	0.9447
	Cluster	0.6668	0.5713	0.7979	0.9266
	Random	0.6537	0.5606	0.7840	0.9160
	First rows	0.6311	0.5362	0.7755	0.8745
Context 3	Original data	0.8974	0.8473	0.9839	0.9979
	Cluster	0.7856	0.7043	0.9161	0.9613
	Random	0.7508	0.6720	0.8624	0.9613
	First rows	0.7279	0.6376	0.8710	0.9290

For the second scenario, the search metrics in different contexts can be seen in Table 5. Similar to the TData dataset, these metrics demonstrate notable outcomes across various contexts and methodologies.

In Context 1 the best performing reduction is Cluster reduction. However, Random reduction and First rows reduction perform significantly worse in this context, with First rows reduction being the weakest, with its MRR dropping to 0.7648 and its P@1 dropping to 0.6963, indicating its limitations in this context. In Context 2, the situation is similar: Cluster approach also provides the best reduction scores, with Random reduction and First

rows reduction performing the worst. One striking fact about this context is that, despite including column names in the index, results are worse for all methods, so we tested adding more metadata in Context 3. In this context, Cluster Random has the best metrics once again. Both Random and First rows reduction also show a substantial improvement in Context 3 compared to the other contexts. Although their performance is slightly lower than that of the Cluster reduction, they achieve high results, with MRR and P@1 scores of 0.9130 and 0.8595, respectively.

These findings also highlight the variability in method effectiveness depending on the specific context. Having analysed the contexts separately, we turn to a general observation of the table. As we can see, the best performance is offered by the original dataset in Context 1, where only the content of the table is taken into account. The original dataset in Context 3 offers almost similar metrics to Context 1 but slightly worse than Context 1, except for P@10.

Across both datasets, Cluster reduction obtains the highest scores among the reduction methods in most contexts, especially when content is combined with richer metadata. Nevertheless, its advantage over Random reduction is usually limited and is not statistically significant across all metrics and settings. The results therefore identify clustering as a competitive and comparatively robust reduction strategy, rather than as a universally superior alternative to random sampling. First rows reduction generally provides the weakest performance.

Table. 5. Search metrics combining different contexts and reduction techniques for the EnviDat dataset.

Ctx	Reduction	MRR	P@1	P@5	P@10
1	Original data	0.9285	0.8817	0.9871	0.9975
	Cluster	0.8929	0.8357	0.9639	0.9892
	Random	0.8726	0.8195	0.9369	0.9718
	First rows	0.7648	0.6963	0.8548	0.9220
2	Original data	0.8476	0.7712	0.9452	0.9826
	Cluster	0.8196	0.7411	0.9220	0.9685
	Random	0.8067	0.7290	0.9050	0.9585
	First rows	0.7182	0.6373	0.8241	0.8963
3	Original data	0.9262	0.8779	0.9869	0.9977
	Cluster	0.9130	0.8595	0.9815	0.9964
	Random	0.8997	0.8491	0.9685	0.9887
	First rows	0.8345	0.7649	0.9234	0.9671

To ensure a realistic evaluation of the data discovery service in our data space architecture, we considered varying data quality levels, particularly regarding sparsity. Therefore, next experiments introduced different percentages of null values into the indexed tables, reflecting the common scenario where tables often contain incomplete rows or missing data. This approach allowed us to assess the service’s performance under conditions that mirror the imperfect data typically encountered in real-world scenarios. This experiment considers only the TData dataset, due to significant differences between the two scenarios. The Envidat dataset represents open data onboarding, characterized by unrestricted

access and the absence of data space policies; therefore, artificial simulation of null values is unnecessary. In practice, external datasets are federated in their entirety after selection, and sparsity tests are reserved for datasets already governed within the data space, as exemplified by TData.

Table 6 presents different levels of null values that were introduced into the original tables for Context 1, Context 2, and Context 3, respectively. In these three contexts, four levels of null value percentages were established: 10%, 25%, 50%, and 75% of missing values. In all these experiments, it is evident that the method based on the First rows consistently yields the worst results, just as it did when working without null values. Therefore, we will focus on discussing the results of comparing the Cluster reduction with the Random reduction.

Table 6 presents the impact of introducing varying levels of null values (10%, 25%, 50%, and 75%) into the original tables across all three contexts. Consistent with our previous findings, First rows reduction consistently produces the poorest results, mirroring its performance in datasets without null values. Given this consistent underperformance, our analysis will focus primarily on comparing Cluster reduction with Random reduction methods.

In *Context 1*, the comparison between Cluster and Random reduction shows similar performance trends across all levels of null values. At 10% null values, Cluster reduction slightly outperforms Random reduction in all the metrics. At 50% and 75% null values, Cluster reduction maintains a narrow advantage over Random reduction. The differences are more clear at 25% null values, where Cluster reduction surpasses Random reduction more clearly in MRR (0.7807 vs. 0.7288) and P@1 (0.6840 vs. 0.6351). This suggests that Cluster reduction has a good behavior in medium sparse datasets. This trend is consistent across the other two contexts. In *Context 2*, it is evident that in situations with a medium number of null values (25%), the Cluster reduction method clearly outperforms the Random reduction across all metrics. The same pattern is observed in *Context 3*, where the cluster-based method shows a significant advantage over the random method. This suggests that in real-world scenarios, where tables often contain a medium proportion of null values, the proposed Cluster reduction can be significantly more effective than the Random reduction.

5.4. Indexing Performance

Indexing speed is a critical performance metric for evaluating search system efficiency, particularly in dynamic environments managing large datasets like data spaces. This paper compares the indexing times of two data catalogs employing distinct approaches: Catalog A indexes complete data tables, calculating embeddings for each row to ensure comprehensive dataset representation, while Catalog B utilizes our proposed Cluster reduction method (detailed in Section 4.3) to generate embeddings from reduced table versions. This latter approach was specifically developed to address the performance limitations observed when processing full datasets.

In our experimental setup for the TData dataset, Catalog A processed datasets averaging 7,480 rows and 19 columns, while Catalog B utilized the same tables reduced to just 10 rows per table through our clustering approach. For TData, Catalog B completed end-to-end index construction in 32 seconds, comprising 20 seconds for sample generation and 12 seconds for indexing, compared with 5,891 seconds for full-content indexing. This

Table 6. Search performance under different percentages of null values.**(a)** Context 1: table content.

N%	Reduction	MRR	P@1	P@5	P@10
10%	Original data	0.8803	0.8198	0.9896	1.0000
	Cluster	0.7172	0.6250	0.8510	0.9615
	Random	0.6990	0.6063	0.8323	0.9302
	First rows	0.6380	0.5333	0.7792	0.8927
25%	Original data	0.8656	0.7968	0.9684	0.9979
	Cluster	0.7329	0.6347	0.8663	0.9716
	Random	0.6847	0.5916	0.8095	0.9368
	First rows	0.6264	0.5200	0.7758	0.8979
50%	Original data	0.8054	0.7198	0.9344	0.9958
	Cluster	0.7043	0.6010	0.8521	0.9688
	Random	0.6543	0.5427	0.8031	0.9115
	First rows	0.6131	0.5146	0.7510	0.8656
75%	Original data	0.6900	0.5688	0.8646	0.9667
	Cluster	0.5919	0.4677	0.7750	0.9052
	Random	0.5506	0.4146	0.7490	0.8677
	First rows	0.4907	0.3740	0.6479	0.8063

(b) Context 2: table content and column names.

N%	Reduction	MRR	P@1	P@5	P@10
10%	Original data	0.7403	0.6438	0.8865	0.9865
	Cluster	0.6633	0.5594	0.8073	0.9458
	Random	0.6414	0.5427	0.7885	0.9104
	First rows	0.6047	0.4833	0.7646	0.8938
25%	Original data	0.7052	0.5874	0.8716	0.9937
	Cluster	0.6542	0.5337	0.8232	0.9463
	Random	0.6158	0.5032	0.7726	0.9032
	First rows	0.6031	0.4853	0.7674	0.9053
50%	Original data	0.6410	0.5125	0.8156	0.9531
	Cluster	0.6033	0.4688	0.7938	0.9281
	Random	0.5540	0.4292	0.7208	0.8542
	First rows	0.5461	0.4302	0.6979	0.8583
75%	Original data	0.5123	0.3698	0.7052	0.9000
	Cluster	0.4932	0.3594	0.6625	0.8333
	Random	0.4238	0.2854	0.6021	0.7729
	First rows	0.4229	0.2896	0.6042	0.7438

(c) Context 3: table content, column names, and column descriptions.

N%	Reduction	MRR	P@1	P@5	P@10
10%	Original data	0.8932	0.8368	0.9905	1.0000
	Cluster	0.7688	0.6726	0.9137	0.9842
	Random	0.7429	0.6537	0.8811	0.9695
	First rows	0.7286	0.6379	0.8463	0.9526
25%	Original data	0.8811	0.8180	0.9723	0.9979
	Cluster	0.7807	0.6840	0.9298	0.9851
	Random	0.7288	0.6351	0.8745	0.9691
	First rows	0.7257	0.6287	0.8713	0.9532
50%	Original data	0.8282	0.7526	0.9421	0.9958
	Cluster	0.7527	0.6484	0.9032	0.9853
	Random	0.7083	0.6116	0.8379	0.9526
	First rows	0.7203	0.6211	0.8495	0.9495
75%	Original data	0.7319	0.6211	0.8916	0.9800
	Cluster	0.6668	0.5568	0.8168	0.9495
	Random	0.6523	0.5389	0.8116	0.9316
	First rows	0.6594	0.5432	0.8211	0.9442



Fig. 4. Search metrics for different null percentages (N%) with Context 1, 2, and 3. A consistent colour scheme has been used to facilitate the interpretation of the results: black represents Original data, red corresponds to reduction by Clustering, blue indicates Random reduction and yellow represents First rows of the dataset.

corresponds to an approximately 184-fold reduction in processing time. For EnviDat, the corresponding average times were 0.327 seconds per table (0.122 seconds for reduction and 0.205 seconds for indexing) compared with 51 seconds for the full-content configuration, corresponding to an approximately 156-fold improvement when both stages are included.

These efficiency gains are not cost-free. In Context 3, the clustered ten-row samples retained 87.5% of the full-content MRR and 83.1% of P@1 for TData, whereas the corresponding values for EnviDat were 98.6% and 97.9%. The reported P@10 values were retained at 96.3% and 99.9%, respectively. Full-content indexing therefore remains preferable when maximum early-rank retrieval effectiveness is the principal requirement and the complete content can be processed under the applicable policies. In contrast, sample-based indexing provides a more appropriate operating point when index-update latency, computational cost, or controlled disclosure is the dominant concern.

The sample size $k=10$ is consequently an architectural sensitivity point rather than a universally optimal setting. Increasing k may improve content coverage and retrieval effectiveness, but it also increases sample-generation cost, indexing time, storage requirements, and the amount of content disclosed. Determining k adaptively according to dataset size, heterogeneity, and disclosure policy remains an important direction for future work.

5.5. Ablation-style analysis of architectural mechanisms

The experiments in Tables 4 and 5 permit an incremental ablation-style analysis along two dimensions. The first concerns catalog representation: Context 1 uses table content only, Context 2 adds column names, and Context 3 further adds column descriptions. The second concerns sample generation: the full-content index is compared with clustered, random, and first-row samples. Table 7 reports the absolute changes in MRR and P@1 for the principal contrasts. For the metadata comparisons, the clustered sample is held constant and positive values indicate an improvement attributable to the added mechanism.

The results indicate that the most influential enrichment mechanism is not the mere presence of additional metadata, but the availability of meaningful field descriptions. Adding names without descriptions reduces retrieval effectiveness in both datasets, whereas adding descriptions more than compensates for this loss. Metadata completeness and descriptive quality should therefore be treated as sensitivity points of the architecture.

The sampling results lead to a different interpretation. Reducing each table to ten rows is responsible for the large improvement in indexing efficiency, while clustering contributes a smaller but consistent retrieval advantage over random sampling. Thus, representative sampling is an essential scalability and controlled-disclosure mechanism, whereas clustering is a supportive optimisation that improves the quality of the reduced representation. Sequential first-row selection is the least suitable alternative because it combines the information loss associated with reduction with a potentially biased sample.

Not every layer shown in Figure 3 can be meaningfully removed while retaining the same system objectives. The connector and policy layer, the separation of protected and sample assets, and policy-aware preprocessing are architectural prerequisites for sovereignty-preserving discovery rather than ranking mechanisms. The semantic indexing service is similarly required to execute the evaluated retrieval process, while the user-interface layer is supportive and does not affect the reported ranking metrics. These components were held fixed in the present experiments.

Table. 7. Incremental ablation-style analysis of catalog representation and sample-generation mechanisms. Absolute changes are reported; positive values favour the added mechanism.

Contrast	TData Δ MRR / Δ P@1	EnviDat Δ MRR / Δ P@1	Interpretation
Add field names: Context 2 – Context 1	−0.0659/ − 0.0808	−0.0733/ − 0.0946	Field names alone do not improve retrieval and may introduce ambiguous or noisy tokens.
Add field descriptions: Context 3 – Context 2	+0.1188/ + 0.1330	+0.0934/ + 0.1184	Descriptions provide the largest positive marginal contribution among the tested metadata elements.
Add names and descriptions: Context 3 – Context 1	+0.0529/ + 0.0522	+0.0201/ + 0.0238	Complete field-level enrichment produces a positive net effect over content-only indexing.
Clustered rather than random sample in Context 3	+0.0348/ + 0.0323	+0.0133/ + 0.0104	Clustering contributes consistently, although its marginal effect is smaller than that of descriptive metadata.
Clustered sample rather than full content in Context 3	−0.1118/ − 0.1430	−0.0132/ − 0.0184	Reduction substantially lowers processing cost but sacrifices some early-rank effectiveness, particularly for TData.

The analysis should consequently be interpreted as an ablation of the representation and sampling mechanisms, rather than as a complete causal decomposition of every architectural layer. In particular, the evaluation does not include a standard index for only DCAT-AP, an only data-dictionary configuration, alternative embedding models, or an ablation of the policy transformations. These configurations will be considered in future work to quantify the independent and interacting contributions of catalog-level metadata, field-level metadata, representative content, and semantic indexing.

6. Conclusions

This paper addresses the persistent limitations of discovery mechanisms in federated data spaces (only based on high-level descriptive metadata) by introducing cbmDCAT-AP, a content-based extension of DCAT-AP designed to enhance dataset discoverability while maintaining data sovereignty. Traditional catalog approaches rely predominantly on high-level descriptive metadata, which is often inadequate for accurately assessing dataset relevance, especially in heterogeneous and large-scale environments. By incorporating structured field-level descriptions, data dictionaries, and representative sample values into the metadata layer, cbmDCAT-AP facilitates a more expressive and semantically rich representation of datasets without exposing full data content.

The proposed cbmDCAT-AP extension has been validated in both discovery scenarios in data spaces addressed in this paper: internal sovereign dataset discovery and external open data onboarding. These results demonstrate its versatility as a content-aware cataloguing mechanism for federated data spaces.

A central contribution of this paper is the integration of metadata enrichment with a clustering-based dataset reduction strategy to enable scalable semantic indexing. Experimental results indicate a substantial improvement in indexing efficiency, achieving speedups of several orders of magnitude when operating on reduced samples, while maintaining competitive retrieval performance across multiple search contexts and different scenarios. Intrinsic statistical similarity is comparable across the evaluated reduction strategies. In the information retrieval experiments, clustering generally outperforms sequential sampling and obtains modest, context-dependent improvements over random sampling, particularly when richer metadata or moderate levels of missing values are considered. However, these differences are not statistically significant in every setting. Therefore, the main benefit of the clustering-based approach lies in combining competitive retrieval quality with semantically informed sampling and substantial indexing-efficiency gains, rather than in providing universal superiority over simpler baselines.

Conceptually, the proposed approach helps mitigate the tension described by Arrow’s information paradox by enabling meaningful pre-access evaluation of dataset content without compromising sovereignty constraints of data spaces. By exposing controlled, representative information rather than raw datasets, providers can enhance transparency, trust, and interoperability while maintaining governance and compliance requirements.

The architectural and ablation analyses further show that no single configuration dominates across all quality attributes. Full-content indexing remains preferable when maximum early-rank retrieval effectiveness is required and processing of the complete dataset is permitted. In contrast, combining policy-compliant clustered samples with field descriptions provides a balanced configuration for high-throughput, sovereignty-constrained catalogs. This operating point is sensitive to sample size, metadata quality, disclosure transformations, dataset-update frequency, and embedding-model selection. Future work will therefore include stakeholder-led architectural evaluation and a factorial ablation of these sensitivity points.

Furthermore, future research directions include exploring adaptive sampling strategies tailored to various data types and quality conditions, conducting large-scale validation in operational data spaces, and refining reduction techniques to better preserve distinctive dataset characteristics. Further investigation into integrating multimodal embeddings and domain-specific ontologies may also enhance semantic alignment across heterogeneous providers. Overall, cbmDCAT-AP and its supporting architecture constitute a practical and scalable advancement toward content-aware, sovereignty-preserving data discovery in next-generation federated data ecosystems.

Acknowledgements. This research was funded by the Spanish Ministry for Digital Transformation and the European Union NextGenerationEU, under HELEADE (TSI-100121-2024-24) and CLIO (TSI-100130-2024-69) projects. Also, part of the research was partially funded by AFFAIR project (PID2025-174636OB-I00), funded by MICIU/AEI/10.13039/501100011033.

References

1. Abedjan, Z., Esmailoghli, M., Galhotra, S.: Data discovery in data lakes: Operations, indexes, systems. *Proceedings of the VLDB Endowment* 18(12), 5455–5459 (2025), <https://doi.org/10.14778/3750601.3750694>

2. Attard, J., Orlandi, F., Scerri, S., Auer, S.: A systematic review of open government data initiatives. *Government information quarterly* 32(4), 399–418 (2015), <https://doi.org/10.1016/j.giq.2015.07.006>
3. Azcoitia, S.A., Laoutaris, N.: A survey of data marketplaces and their business models. *ACM SIGMOD Record* 51(3), 18–29 (2022), <https://doi.org/10.1145/3572751.3572755>
4. Bacco, M., Kocian, A., Chessa, S., Crivello, A., Barsocchi, P.: What are data spaces? systematic survey and future outlook. *Data in Brief* 57, 110969 (2024), <https://doi.org/10.1016/j.dib.2024.110969>
5. Berenguer, A., Alcaraz, O., Tomás, D., Mazón, J.N.: From research on data-intensive software to innovation in data spaces: a search service for tabular data. *IEEE Software* 41(3), 59–66 (2024), <https://doi.org/10.1109/MS.2024.3359333>
6. Bernhauer, D., Nečaský, M., Škoda, P., Klímek, J., Skopal, T.: Open dataset discovery using context-enhanced similarity search. *Knowledge and Information Systems* 64(12), 3265–3291 (2022), <https://doi.org/10.1007/s10115-022-01751-z>
7. Bogatu, A., Fernandes, A.A.A., Paton, N.W., Konstantinou, N.: Dataset discovery in data lakes. In: 2020 IEEE 36th International Conference on Data Engineering (ICDE). pp. 709–720. IEEE (2020), <https://doi.org/10.1109/ICDE48307.2020.00067>
8. Chapman, A., Simperl, E., Koesten, L., Konstantinidis, G., Ibáñez, L., Kacprzak, E., Groth, P.: Dataset search: A survey. *The VLDB Journal* 29(1), 251–272 (2020), <https://doi.org/10.1007/s00778-019-00564-x>
9. Chu, M.T., Khosla, R.: Index evaluations and business strategies on communities of practice. *Expert Systems with Applications* 36(2, Part 1), 1549–1558 (2009), <https://doi.org/10.1016/j.eswa.2007.11.053>
10. Cinquini, M., Giannotti, F., Guidotti, R.: Boosting synthetic data generation with effective non-linear causal discovery. In: 2021 IEEE Third International Conference on Cognitive Machine Intelligence (CogMI). pp. 54–63 (2021), <https://doi.org/10.1109/CogMI52975.2021.00016>
11. Cui, Y., Jin, Z., Jiang, J.: A novel supervised feature extraction and classification fusion algorithm for land cover recognition of the off-land scenario. *Neurocomputing* 140, 77–83 (2014), <https://doi.org/10.1016/j.neucom.2014.03.034>
12. Curry, E., Tuikka, T.: An organizational maturity model for data spaces: A data sharing wheel approach. In: *Data Spaces: Design, Deployment and Future Directions*, pp. 21–42. Springer (2022), https://doi.org/10.1007/978-3-030-98636-0_2
13. Dekkers, M., Kotoglou, S., Nelson, C., Pellegrino, M., Hohn, N., Peristeras, V.: Statdcat-ap, a common layer for the exchange of statistical metadata in open data portals. In: *SemStats@ISWC* (2016), <https://ceur-ws.org/Vol-1654/article-05.pdf>
14. Derycke, P., Barros, B.J., Schutte, N.M., Vande Catsbyne, C.A., Fonseca, M.B.: Designing dcat-ap extensions for common european data spaces: The ehds healthdcat-ap case study. In: *NeXt-generation Data Governance workshop 2025* (2025), <https://ceur-ws.org/Vol-4064/NXDG25-paper5.pdf>
15. Dong, Y., Xiao, C., Nozawa, T., Enomoto, M., Oyamada, M.: Deepjoin: Joinable table discovery with pre-trained language models. *Proceedings of the VLDB Endowment* 16(10), 2458–2470 (2023), <https://doi.org/10.14778/3603581.3603587>
16. Eclipse Dataspace Components: *Eclipse Dataspace Components* (2025), <https://projects.eclipse.org/projects/technology.edc>
17. Enríquez-Reyes, R., Cadena-Vela, S., Fuster-Guilló, A., Mazón, J.N., Ibáñez, L.D., Simperl, E.: Systematic mapping of open data studies: Classification and trends from a technological perspective. *IEEE Access* 9, 12968–12988 (2021), <https://doi.org/10.1109/ACCESS.2021.3052025>

18. European Commission: Simpl: Cloud-to-edge federations empowering eu data spaces (2024), <https://digital-strategy.ec.europa.eu/en/policies/simpl>, last updated 11 October 2024. Accessed: 2026-02-28
19. European Commission: Common european data spaces (2025), <https://digital-strategy.ec.europa.eu/en/policies/data-spaces>, last updated 19 November 2025. Accessed: 2026-02-28
20. European Commission: Dcat application profile for data portals in europe (2025), <https://interoperable-europe.ec.europa.eu/collection/semic-support-centre/solution/dcat-application-profile-data-portals-europe/release/301>, last updated 19 November 2025. Accessed: 2026-02-28
21. Fan, G., Wang, J., Li, Y., Zhang, D., Miller, R.J.: Semantics-aware dataset discovery from data lakes with contextualized column-based representation learning. *Proceedings of the VLDB Endowment* 16(7), 1726–1739 (2023), <https://doi.org/10.14778/3587136.3587146>
22. FIWARE Foundation: FIWARE - Open APIs for Open Minds (2024), <https://www.fiware.org>
23. Fowler, D., Barratt, J., Walsh, P.: Frictionless data: making research data quality visible. *International Journal of Digital Curation* 12(2), 274–285 (2017), <https://doi.org/10.2218/ijdc.v12i2.577>
24. Gaia-X European Association for Data and Cloud AISBL: Gaia-X Architecture Document (2024), <https://docs.gaia-x.eu/technical-committee/architecture-document/latest/>
25. Gieß, A., Hupperz, M.J., Schoormann, T., Möller, F.: What does it take to connect? unveiling characteristics of data space connectors. In: Bui, T.X. (ed.) 57th Hawaii International Conference on System Sciences, HICSS 2024, Hilton Hawaiian Village Waikiki Beach Resort, Hawaii, USA, January 3-6, 2024. pp. 4238–4247. ScholarSpace (2024), <https://hdl.handle.net/10125/106895>
26. Gieß, A., Möller, F., Schoormann, T., Otto, B.: Design options for data spaces. In: Thirty-first European Conference on Information Systems (ECIS 2023) (06 2023), https://aisel.aisnet.org/ecis2023_rp/287/
27. Hewagamage, P., Mihiranga, A., Perera, D., Fernando, R., Thilakarathna, T., Kasthurirathna, D.: Computer-vision enabled waste management system for green environment. In: 2021 3rd International Conference on Advancements in Computing (ICAC). pp. 276–281 (2021), <https://doi.org/10.1109/ICAC54203.2021.9671222>
28. Hummel, P., Braun, M., Tretter, M., Dabrock, P.: Data sovereignty: A review. *Big Data & Society* 8(1), 2053951720982012 (2021), <https://doi.org/10.1177/2053951720982012>
29. i4Trust: B2B Data Sharing Playbook - the i4Trust approach to Data Sharing (2021), https://i4trust.org/wp-content/uploads/i4Trust_DataSharingPlaybook.pdf
30. Imielinski, T., Mannila, H.: A database perspective on knowledge discovery. *Commun. ACM* 39(11), 58–64 (Nov 1996), <https://doi.org/10.1145/240455.240472>
31. International Data Spaces Association: Position Paper - GAIA-X and IDS (2021), https://internationaldataspaces.org/wp-content/uploads/dlm_uploads/IDSA-Position-Paper-GAIA-X-and-IDS.pdf
32. International Data Spaces Association: IDS Reference Architecture Model, Version 4.0 (2022), https://docs.internationaldataspaces.org/ids-knowledgebase/ids-ram-4/introduction/1_1_goals_of_the_international_data_spaces
33. International Data Spaces Association: International Data Spaces Association (2022), <https://internationaldataspaces.org/>

34. iSHARE Foundation: iSHARE - Trust Framework for Data Spaces (2021), <https://ishare.eu/>
35. Jahnke, N., Otto, B.: Data catalogs in the enterprise: applications and integration. *Datenbank-Spektrum* 23(2), 89–96 (2023), <https://doi.org/10.1007/s13222-023-00445-2>
36. Kazman, R., Klein, M., Clements, P.: Atam: Method for architecture evaluation. Tech. Rep. CMU/SEI-2000-TR-004, Software Engineering Institute, Carnegie Mellon University (Aug 2000), <https://www.sei.cmu.edu/library/atam-method-for-architecture-evaluation/>, accessed: 2026-Jul-16
37. Kirstein, F., Dutkowski, S., Dittwald, B., Hauswirth, M.: The european data portal: Scalable harvesting and management of linked open data. In: ISWC (Satellites). pp. 321–322 (2019), <https://ceur-ws.org/Vol-2456/paper88.pdf>
38. Kitchin, R.: The data revolution: Big data, open data, data infrastructures and their consequences. Sage (2014), <https://doi.org/10.4135/9781473909472>
39. Morejón, A., Berenguer, A., Espona, L.d., Tomás, D., Mazón, J.N.: Improving data discovery effectiveness: Experimental evaluation of content-based catalogs in data spaces. In: Chrysanthis, P.K., Nørvåg, K., Stefanidis, K., Zhang, Z., Quintarelli, E., Zumpano, E. (eds.) *New Trends in Database and Information Systems*. pp. 280–295. Springer Nature Switzerland, Cham (2026)
40. Nargesian, F., Asudeh, A., Jagadish, H.V.: Responsible data integration: Next-generation challenges. In: *Proceedings of the 2022 International Conference on Management of Data*. p. 2458–2464. SIGMOD '22, Association for Computing Machinery, New York, NY, USA (2022), <https://doi.org/10.1145/3514221.3522567>
41. Otto, B.: A federated infrastructure for european data spaces. *Communications of the ACM* 65(4), 44–45 (2022), <https://doi.org/10.1145/3512341>
42. Perego, A., Cetl, V., Friis-Christensen, A., Lutz, M., et al.: Geodcat-ap: Representing geographic metadata by using the “dcat application profile for data portals in europe”. In: *Joint UNECE/UNGIM Europe Workshop on Integrating Geospatial and Statistical Standards*, Stockholm, Sweden (2017), <https://publications.jrc.ec.europa.eu/repository/handle/JRC107410>
43. Ramos Rojas, J.A., Beth Kery, M., Rosenthal, S., Dey, A.: Sampling techniques to improve big data exploration. In: *2017 IEEE 7th Symposium on Large Data Analysis and Visualization (LDAV)*. pp. 26–35 (2017), <https://doi.org/10.1109/LDAV.2017.8231848>
44. Scerri, S., Tuikka, T., de Vallejo, I.L., Curry, E.: Common european data spaces: Challenges and opportunities. *Data Spaces: Design, Deployment and Future Directions* pp. 337–357 (2022), https://doi.org/10.1007/978-3-030-98636-0_16
45. W3C: Data catalog vocabulary (dcat) - version 3 (2024), <https://www.w3.org/TR/vocab-dcat/>, last updated 19 November 2025. Accessed: 2026-02-28
46. Xu, P., Ji, X., Li, M., Lu, W.: Virtual sample generation in machine learning assisted materials design and discovery. *Journal of Materials Informatics* 3(3) (2023), <https://doi.org/10.20517/jmi.2023.18>
47. Zhang, S., Balog, K.: Web table extraction, retrieval, and augmentation: A survey. *ACM Transactions on Intelligent Systems and Technology* 11(2), 13:1–13:35 (2020), <https://doi.org/10.1145/3372117>

Adriana Morejón is a researcher in the field of artificial intelligence and open data management, affiliated with the University of Alicante. Her work focuses on the application of AI techniques for data discovery and reuse, as well as on improving reference architectures for data spaces and open data ecosystems in the contexts of smart cities and tourism. She regularly collaborates on research and development projects, contributing

to the development of solutions aimed at the integration, visualisation and generation of open data, with a strong focus on quality and added value for business scenarios and data intelligence.

Alberto Berenguer holds a Ph.D. in Computer Science from the University of Alicante. His work sits at the intersection of machine learning and information science. His research focuses on applying advanced machine learning techniques to information retrieval and extraction. Beyond academia, he takes part in innovative projects on smart cities and data spaces, where he contributes expertise in data retrieval and metadata enhancement.

Lucía de Espona Pernas is a Senior Data Spaces Researcher and Computer Scientist specializing in data engineering and open data frameworks. She holds a PhD in Computer Science from the Universidad de Santiago de Compostela. Her technical background includes molecular biology research at ETH Zurich and leading engineering tracks for Innosuisse projects. Currently, she researches with the WaKe group at the University of Alicante, focusing on data spaces architecture, scalable sharing ecosystems, semantic interoperability, and trusted data infrastructures.

David Tomás is a Full Professor in the Department of Software and Computing Systems at the University of Alicante. His primary research focus is the application of Artificial Intelligence techniques to Natural Language Processing and Information Retrieval, fields in which he has published more than one hundred research articles in international conferences and high-impact journals. In these areas, he has participated in over twenty publicly funded projects at both the national and European levels.

Jose-Norberto Mazón is currently Full Professor in the University of Alicante, Spain. He is the author of more than 100 scientific publications in international conferences and journals. His research interests include open data, dataset discovery, data spaces, design of data-intensive web applications, and smart tourism destinations.

Received: March 1, 2026; Accepted: August 25, 2026.

