Data Interoperability and Semantics

Erdem Önal

1 Portal and Data Topic

This document is a report on the recommended, enforced, and commonly used metadata formats and schemas in Sweden's national open data portal, Sveriges Dataportal. The portal is managed by the Swedish Digital Government Agency (Digg). The portal URL is: `https://www.dataportal.se`.

The selected data topic is Riksdagen, the Swedish Parliament. I chose 11 datasets from three different provider types to analyze:

- From Riksdagsförvaltningen (Parliament Admin):
  - Riksdagens dokument (Parliamentary Documents): `https://www.riksdagen.se/sv/dokument-och-lagar/riksdagens-oppna-data/`
    - Demo: `https://data.riksdagen.se/dokumentlista/`
  - Riksdagens ledamöter (Members of Parliament): `https://www.riksdagen.se/sv/dokument-och-lagar/riksdagens-oppna-data/ledamoter/`
    - Demo: `https://data.riksdagen.se/personlista/`
  - Riksdagens kalenderhändelser (Calendar Events): `https://www.riksdagen.se/sv/dokument-och-lagar/riksdagens-oppna-data/kalender/`
    - Demo: `https://data.riksdagen.se/kalender`
  - Riksdagens anföranden (Speeches): `https://www.riksdagen.se/sv/dokument-och-lagar/riksdagens-oppna-data/anforanden/`
    - Demo: `https://data.riksdagen.se/anforandelista/`
  - Riksdagens voteringar (Votings): `https://www.riksdagen.se/sv/dokument-och-lagar/riksdagens-oppna-data/voteringar/`
    - Demo: `https://data.riksdagen.se/voteringlista/`
- From Göteborgs universitet:
  - Riksdagens öppna data: Motion (Parliament Open Data: Motions): `https://doi.org/10.23695/9HD8-5T52`

  - Riksdagens öppna data: Proposition (Parliament Open Data: Propositions): `https://doi.org/10.23695/9RTB-TX57`
  - The Swedish PoliGraph: `https://doi.org/10.23695/WMCR-7C32`
    - Demo: `https://spraakbanken.gu.se/poligraph/`
  - Svenska partiprogram och valmanifest (Party Programs): `https://doi.org/10.5878/kcsf-k293`

- From Statistical Agencies (SCB and Kolada):
  - Tabell 1.4: Ledamöter i riksdagens utskott (Members of Committees): `https://api.scb.se/OV0104/v1/doris/sv/ssd/LE/LE0201/LE0201Makt/Tema14`
  - Förtroende för riksdagens politiker (Trust in Parliament's Politicians): `http://api.kolada.se/v3/kpi/N00665`

2 Used Data Formats

Dataset	Formats Used	Data Specific and Standard Schema Fields
Riksdagens dokument	JSON, XML, CSV, SQL, HTML, TXT (ZIP archives)	`dok_id, doktyp, organ, datum, titel, status`
Riksdagens ledamöter	JSON, XML, CSV, SQL, HTML (ZIP archives)	`intressent_id, förnamn, efternamn, parti, kön, valkrets`
Riksdagens kalender	JSON, XML, CSV, HTML, TXT (via API)	`CATEGORIES, LOCATION, UID, DTSTART, DTEND, SUMMARY, X-RD-DOKID`
Riksdagens anföranden	JSON, XML, CSV, SQL, HTML, TXT (ZIP archives)	`dok_id, anforande_nummer, talare, parti, intressent_id`
Riksdagens voteringar	JSON, XML, CSV, SQL, HTML (ZIP archives)	`votering_id, intressent_id, parti, valkrets, rost, avser`
Riksdagens Motion	XML (corpus file: `rd-mot.xml.bz2`)	Unstructured text data. No formal data schema fields.
Riksdagens Proposition	XML (corpus file: `rd-prop.xml.bz2`)	Unstructured text data. No formal data schema fields.
The Swedish PoliGraph	Prolog (file: `poligraph.tar.bz2`)	Prolog fact database with predicates for querying parliamentary data.
Svenska partiprogram	PDF, TXT, XLSX, CSV (all in ZIP)	Data is unstructured PDF/TXT. Doc file schema: `type, source, titles`.
Tabell 1.4	JSON (via API)	`Utskott, Kon, ContentsCode, Tid`
Förtroende	JSON (via API)	`id, title, description, is_divided_by_gender`

Table 1: Overview of Formats and Data Schemas.

API Access Example

The dataset Förtroende för riksdagens politiker can be accessed directly through a REST API. The snippet below shows a `curl` request:

```
$ curl "http://api.kolada.se/v3/kpi/N00665"
```

The API returns the following JSON output, containing both the metadata and values for this entry:

```
{
  "values": [
    {
      "id": "N00665",
      "title": "Medborgarundersökningen - Förtroende...",
      "description": "Andel som har svarat ...",
      "is_divided_by_gender": true,
      "municipality_type": "A",
      "publication_date": "2025-12-16",
      "publ_period": "2025"
    }
  ],
  "count": 1
}
```

3 Metadata Fields

The metadata fields used in the datasets follow Sweden's national metadata specification DCAT-AP-SE 3.0.1, which is the Swedish application profile of DCAT-AP.[1] This document defines which metadata properties are mandatory, recommended, or optional for each class.

Enforced Metadata Fields

According to DCAT-AP-SE 3.0.1, the following properties are mandatory for every dataset:

- dct:title, Dataset name
- dct:description, Summary of the dataset
- dct:publisher, A `foaf:Agent` responsible for publishing
- dcat:distribution, At least one distribution with an `accessURL`.

Recommended Metadata Fields

DCAT-AP-SE also defines several recommended properties for datasets.

- dcat:keyword, Free-text keywords
- dcat:theme, Theme classification using controlled vocabularies, such as the EU data themes at:
  `http://publications.europa.eu/resource/authority/data-theme`

[1] `https://docs.dataportal.se/dcat/en/`

- dct:language, Language(s) of the dataset
- dcat:contactPoint, A contact person or organisation, modelled as a `vcard:Organization`
- dct:accessRights, Public/restricted access using the controlled vocabulary at `http://publications.europa.eu/resource/authority/access-right`
- dct:issued and dct:modified, Publication and update date

Parliament datasets include most of the recommended fields, whereas research datasets do not provide several of them.

Used Metadata from the Datasets

The usage of DCAT-AP-SE metadata fields across the datasets is the following:

- Title, Description, Publisher: 11/11, 100%
- Keyword: 5/11 45% Only the Parliament datasets provide keywords.
- Theme: 5/11 45% Only datasets published by the Parliament include a DCAT theme classification.
- Contact point: 5/11 45% Present in the Parliament datasets.
- Access rights: 11/11 100% All datasets declare public access.
- Landing page: 9/11 82% Parliament and DOI datasets have landing pages, datasets from SCB and Kolada do not.

4 Used Schemas and Standards

Sweden's national open data portal does not use plain DCAT. Instead, it implements the Swedish metadata profile DCAT-AP-SE 3.0.1, which adapts DCAT-AP to national requirements. This profile defines the structure and semantics of dataset metadata. It also adds Swedish terminology, controlled vocabularies, and additional constraints to DCAT.

DCAT-AP-SE 3.0.1

DCAT-AP-SE is based on the W3C DCAT (Data Catalog Vocabulary), the EU application profile DCAT-AP, and a set of additional Swedish constraints, clarifications, and codelists.

The specification defines seven main classes: `Catalog`, `Dataset`, `DatasetSeries`, `Distribution`, `DataService`, `Agent`, and `ContactPoint`.

Each class has mandatory and recommended properties. For example, `dcat:Dataset` must include at least `dct:title`, `dct:description`, `dct:publisher`, and `dcat:distribution` with an `accessURL`.

Datasets can export their metadata in multiple formats:

- RDF/XML
- Turtle
- N-Triples
- JSON-LD

All Riksdagen datasets support these formats.

PROF-SE: Metadata for Specifications

In addition to dataset metadata, Sweden also has a metadata profile for describing specifications themselves, called PROF-SE 1.1.0.[2] This profile is based on the W3C Profiles Vocabulary and ADMS, and is used to describe the DCAT-AP-SE specification, its versions, its dependencies and its publishing organisation.

PROF-SE introduces structured metadata for:

- Specification: title, version, description
- SpecificationResource: documents, schemas
- Publisher: organisation
- ContactPoint

DCAT-AP-SE itself is described using PROF-SE.

DCAT-AP-SE Example

Below is an example in Turtle showing how DCAT-AP-SE metadata is represented for the dataset Riksdagens dokument:

```
@prefix dcat: <http://www.w3.org/ns/dcat#> .
@prefix dct:  <http://purl.org/dc/terms/> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .

<https://www.dataportal.se/datasets/98_3014>
```

[2] `https://docs.dataportal.se/prof/en/`

```
    a dcat:Dataset ;
    dct:title "Riksdagens dokument"@sv ;
    dct:description "Samling av riksdagens handlingar..."@sv ;
    dct:publisher [
        a foaf:Organization ;
        foaf:name "Riksdagsförvaltningen"
    ] ;
    dcat:keyword "riksdagen", "dokument" ;
    dcat:theme <http://publications.europa.eu/resource/authority/data-theme/GOVE> ;
    dcat:distribution <https://data.riksdagen.se/dokumentlista/json> .

<https://data.riksdagen.se/dokumentlista/json>
    a dcat:Distribution ;
    dct:format <http://publications.europa.eu/resource/authority/file-type/JSON> ;
    dcat:accessURL <https://data.riksdagen.se/dokumentlista/> .
```

5 Comparison and Recommendations

Common Patterns

Format	Dataset Count	Percentage
JSON	7/11	64%
XML	7/11	64%
CSV	6/11	55%
HTML	5/11	45%
SQL	4/11	36%
TXT	4/11	36%
PDF	1/11	9%
Prolog	1/11	9%

Table 2: Format distribution across all datasets.

As shown in Table 2, JSON and XML are the most common formats (64% each), followed by CSV (55%).

- Metadata: All 11 datasets share the mandatory metadata fields (Title, Publisher, Description).

- Encoding: The Parliament documentation uses UTF-8 encoding for all its datasets.

Differences and Interoperability

- Specific Provider Formats: Interoperability is limited by formats that are not shared. The Parliament offers SQL, while the university uses research formats like Prolog and PDF. University datasets miss recommended DCAT-AP-SE properties such as dcat:theme and dct:publisher identifiers.

- Data Schema Inconsistency: This is the biggest problem. As shown in Table 1, there is no shared data schema. Even when providers use the same JSON format, the field names differ.

  This inconsistency is visible even within the same provider. The Parliament Administration uses:

  - `dok_id` as the ID for documents.
  - `intressent_id` as the ID for members.

  This is different from Kolada, which uses `id`.

Recommendations

- Portal Metadata Schema: Strong. The portal uses the DCAT-AP-SE profile, enabling standardized discovery.

- Data Format Standardization: Partial. JSON/XML are common, but the University's use of Prolog/PDF differs.

- Data Content Schema: Limited. There is no standardization. This limits interoperability.

These issues could be improved with the following recommendations:

- Create a Common URI Namespace: The data schema inconsistency is the main problem. A central URI system should be created. Ideally, each person would be assigned a single persistent URI like `https://data.riksdagen.se/person/c643fdd9-746d-40da-9ee9-8aa25123e221`. All datasets should use this single, persistent identifier.

- Develop a Shared Riksdagen Ontology: A formal ontology using RDFS/OWL should be created for all parliamentary data. This ontology would define classes like `riks:Ledamot`, `riks:Votering` and properties like `riks:harParti` or `riks:röstade`. This would solve the schema problem semantically, by giving shared meanings to entities and field names.

Semantic Solution

To validate the Shared Ontology recommendation, I constructed an RDF/Turtle model. This shows how to link the heterogeneous IDs semantically, without changing the legacy systems.

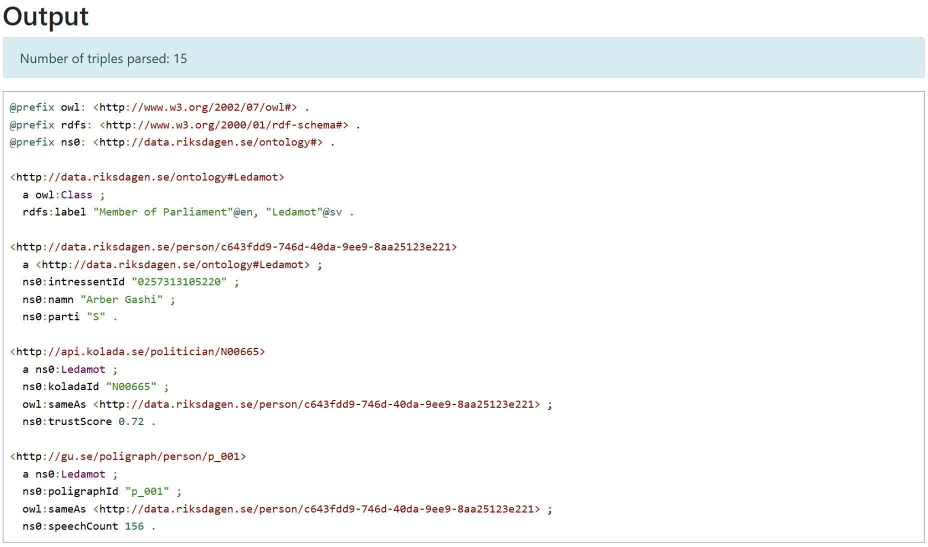
```
@prefix riks: <http://data.riksdagen.se/ontology#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .

# Define the Unified Class
riks:Ledamot a owl:Class ;
    rdfs:label "Member of Parliament"@en, "Ledamot"@sv .

# Parliament Data
<http://data.riksdagen.se/person/c643fdd9-746d-40da-9ee9-8aa25123e2⌋
↪  21> a riks:Ledamot ;
    riks:intressentId "0257313105220" ;
    riks:namn "Arber Gashi" ;
    riks:parti "S" .

# Kolada Data
<http://api.kolada.se/politician/N00665> a riks:Ledamot ;
    riks:koladaId "N00665" ;
    owl:sameAs <http://data.riksdagen.se/person/c643fdd9-746d-40da-⌋
    ↪  9ee9-8aa25123e221> ;
    riks:trustScore "0.72"^^xsd:decimal .

# University Data
<http://gu.se/poligraph/person/p_001> a riks:Ledamot ;
    riks:poligraphId "p_001" ;
    owl:sameAs <http://data.riksdagen.se/person/c643fdd9-746d-40da-⌋
    ↪  9ee9-8aa25123e221> ;
    riks:speechCount "156"^^xsd:integer .
```



Output

Number of triples parsed: 15

```
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix ns0: <http://data.riksdagen.se/ontology#> .

<http://data.riksdagen.se/ontology#Ledamot>
  a owl:Class ;
  rdfs:label "Member of Parliament"@en, "Ledamot"@sv .

<http://data.riksdagen.se/person/c643fdd9-746d-40da-9ee9-8aa25123e221>
  a <http://data.riksdagen.se/ontology#Ledamot> ;
  ns0:intressentId "0257313105220" ;
  ns0:namn "Arber Gashi" ;
  ns0:parti "S" .

<http://api.kolada.se/politician/N00665>
  a ns0:Ledamot ;
  ns0:koladaId "N00665" ;
  owl:sameAs <http://data.riksdagen.se/person/c643fdd9-746d-40da-9ee9-8aa25123e221> ;
  ns0:trustScore 0.72 .

<http://gu.se/poligraph/person/p_001>
  a ns0:Ledamot ;
  ns0:poligraphId "p_001" ;
  owl:sameAs <http://data.riksdagen.se/person/c643fdd9-746d-40da-9ee9-8aa25123e221> ;
  ns0:speechCount 156 .
```

Figure 1: RDF Turtle model validation

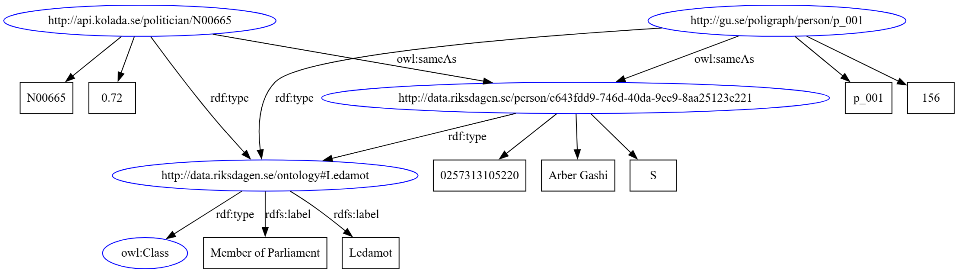

http://api.kolada.se/politician/N00665
http://gu.se/poligraph/person/p_001
owl:sameAs
owl:sameAs
N00665
0.72
rdf:type
rdf:type
http://data.riksdagen.se/person/c643fdd9-746d-40da-9ee9-8aa25123e221
p_001
156
rdf:type
http://data.riksdagen.se/ontology#Ledamot
0257313105220
Arber Gashi
S
rdf:type
rdfs:label
rdfs:label
owl:Class
Member of Parliament
Ledamot

Figure 2: Knowledge graph visualization