Information Interoperability and Application Model

Stefano Nativi

THREDDS Project

UNIDATA/UCAR

Draft 1.2.1

Boulder November 2001

Main Concepts

Interoperability issue

Information Interoperability

Information interoperability is concerned with the possibility of utilising disparate (i.e. distributed and heterogeneous) information content to achieve a common task.

To achieve information interoperability, THREDDS must be able to implement dataset interchange.

Knowledge Sharing

To implement dataset interchange with a given data server, the server and the THREDDS systems must rely on 3 pieces of shared knowledge:

1. A common Application Schema;
2. A common encoding rules to produce/understand transferable dataset;
3. A common transfer protocol.

For example, the following logical steps must be taken in order to transfer a dataset from a given data server to the THREDDS system.

1. **Sender: Application Schema Level**
   Two possible solutions:
   A. The Data Server translates its internal data into a data structure according to a common THREDDS application schema.
   B. The Data Server does not translate its internal data according to a common THREDDS application schema; that is because it assumes that THREDDS application shares its application model knowledge.
   For the first case, the procedure consists in defining a mapping from the concepts of the data server schema to the concepts defined in the THREDDS common schema; an appropriate software, developed on the top of such conceptual mapping, maps instances from one data structure to another.

2. **Sender: Dataset Level**
   The Data Server utilises an encoding service (based on a set of encoding rules), which creates a dataset that is system independent and therefore suitable to be transferred. This dataset may be stored in a file system or transferred using a transfer service.

3. **Sender: Message Level**
   The Data Server invokes a transfer service (on the THREDDS system) to send the encoded dataset to the THREDDS system. The transfer service follows a transfer protocol which specifies packaging and transport rules. Transportation should be possible either over an on-line or off-line communication medium.

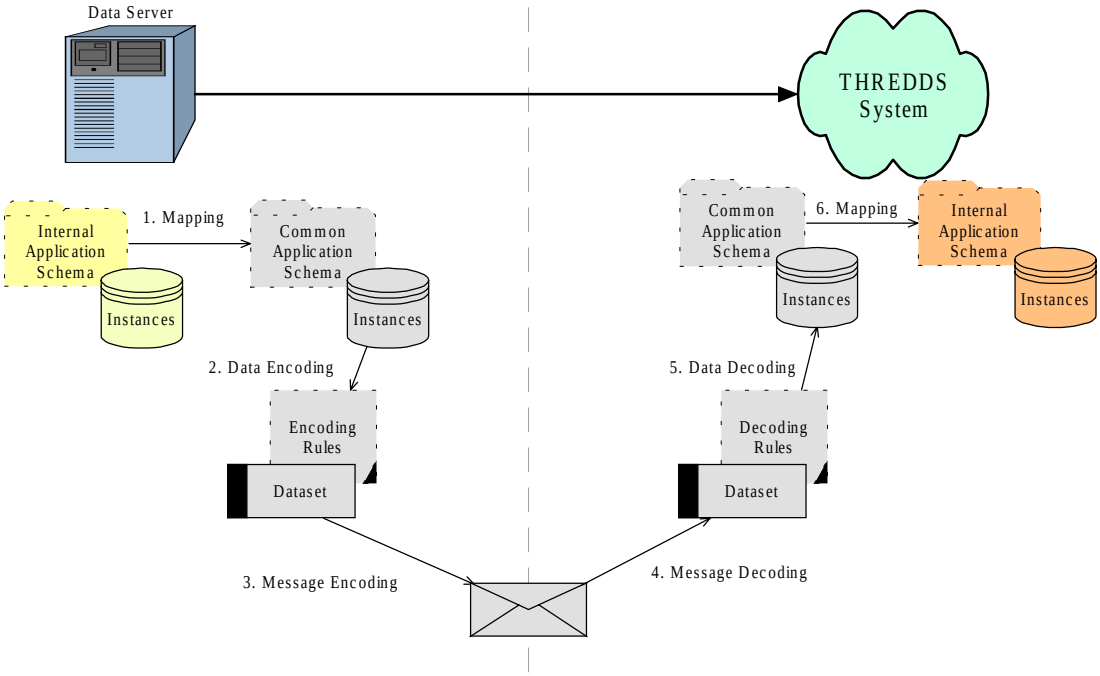
4. **Receiver: Message Level**
   The THREDDS system receives the packed message through the transfer service; according to the transfer protocol, it unpacks the message and stores the dataset (e.g. on an intermediate file).

5. **Receiver: Dataset Level**
   THREDDS applies the inverse encoding rules –utilised by the server- to decode the dataset and get a data structure related to a specific application schema.

6. **Receiver: Application Schema Level**
   In order to fully utilise the received dataset, THREDDS must translate the data structure –which is specific for the data server application schema- into its internal data structure.
   Two possible situations:
   A. The data structure is according to the common THREDDS model;
   B. The data structure is according to the internal application schema of the data server.

   For both the cases, the procedure consists in defining a mapping from the concepts of either the data server schema or the common THREDDS schema to the concepts defined in the internal schema of THREDDS; an appropriate software, developed on the top of such conceptual mapping, maps instances from one data structure to another.

This interaction is depicted in the following schema.



Information interoperability interaction

Sharing Application Models

The knowledge sharing about the provider application model is essential for understanding the semantics and the structure of transferred dataset.

Three main solutions are possible:

1. Adopting standardised Reference Models: several data structures defined by reference model can be adopted as an application common model (e.g. standard catalogue models).

2. Sharing custom structures of an application model in an off-line way: the client and server must share such information before starting data interchange session. It is possible for the server to *publish* its application schema, and therefore the client can get it; in this case they must share other meta-models such as: the formalisation language utilised for the application schema publishing.
3. Sharing custom structures of an application model in an on-line way: client can exploit ontology discovery technology (e.g. RDF, XMI technologies).

The Cataloguing Service case

A particular case of application schema sharing is represented by cataloguing service: specifications to describe and interchange dataset, in a given domain.

In fact, such service implementations specify a common interface as well as a common application schema; this model consists of concept such as: catalogue, dataset, catalogue metadata etc.

Data Interchange Implementations

Data interchange between SW systems may take place in two ways:

- **The data transfer model (traditional model)**
  the data supplier creates a dataset that is transferred to the user. The structure and the content of data are described in the application schema for the dataset. The dataset is sent in a transfer format according to the encoding model.
- **The interoperability model (transaction model)**
  the user application communicates with the supplier application through a common communication protocol. In this scenario, the user invokes services that result in data being passed from the service provider to the user application. The application schema describes not only the structure and content of the exchanged data but also the structure of the interfaces involved in the transaction.

There is a fundamental distinction between a data transfer and a data transaction:

- In data transfer, a dataset is predefined in an application schema. Users request and receive a copy of the dataset. The request regards the entire dataset as a unique opaque object.
- In data transaction, users first specifies selection criteria -such as *who*, *when*, *where*, *what*- for composing the dataset, and then the request can be completed. Data meeting the selection criteria is then retrieved from the data store and provided to the user. Both the request and the resulting dataset are defined according the application schema

Application Model

Introduction

For a SW application, the application model is the framework in which the application schema is achieved.

An application schema is a conceptual schema for data required by a given SW application.

The main purposes of an application schema are:

- To provide a computer-readable data description defining the data structure, which makes it possible to apply automated mechanisms for data management;

- To achieve a common and correct understanding of the data which characterise the application; that is obtained by documenting the data content of the application field, making it possible to unambiguously retrieve information from the data.

Application model Views

An application schema defines two main aspects of an application:

1. content and structure of dataset;
2. specifications of operations for manipulating and processing data by the given application.

Therefore it is possible to introduce two different views of an application model:

1. the Information view;
2. the Computational view.

The first view is essential for data interchange between systems, the second one is useful for describing data interfaces and therefore available services (e.g. Catalogue service).

Application models for Information Interoperability

In the context of information interoperability three different application models have a significant role (see the information interoperability interaction schema):

1. The data supplier application model;
2. The common application model;
3. The data consumer application model.

In the case of THREDDS, its internal model plays the role of the data consumer application model; meanwhile, data server model is the data supplier application model.

Naturally, it is possible to invert the roles.

Proposed Interoperability Solution

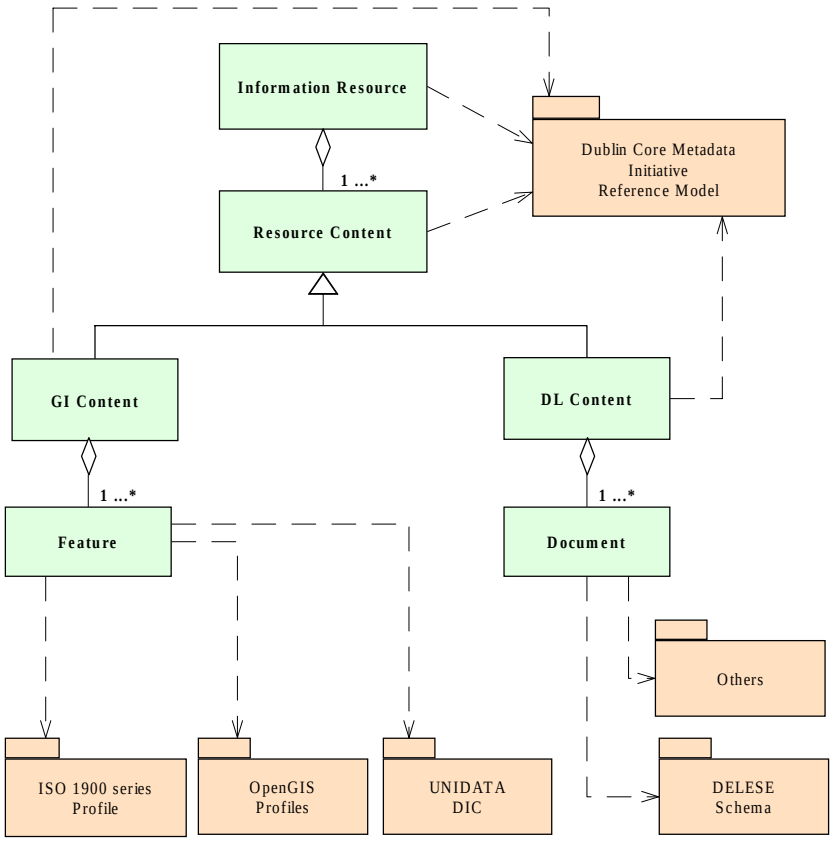
THREDDS Information Interoperability

In order to support information interoperability several solutions are possible; the proposed solution consists on the following choices.

Common Application Schema

The General Resource schema

As far as the common application schema is concerned, THREDDS will support a common application domain which is called General Resource schema (GR schema). The following figure reports the GR schema.



THREDDS common application schema: General Resource schema

The GR schema presents 2 main sub-domains:

1. Geospatial Information;
2. Digital Library Information.

The DCMI model specifies the schema for describing the information resource, meanwhile the others reference models describe the content data structure.

In order to work, the system must support at minimum the information resource reference model (i.e. the Dublin Core schema) and one of the others (e.g. the DIC schema or the DLESE schema).

The system can be extended by adopting other standardised reference models, as far as information content is concerned.

Digital Library Information

As far as the Digital Library Information application model, at least 2 initial standardised reference models as foundations:

1. The Dublin Core Metadata Iniziative model (DCMI);
2. The DLESE model.

Others may be considered (e.g. IMS, ADL, etc.);

Geospatial Information

As far as Geospatial Information is concerned, THREDDS will support a common application domain which has at least 2 initial standardised reference models (the first two in the list); but other 4 reference models are soon recognised as useful extensions, and can be be considered as part of the GR schema:

1. The Dublin Core Metadata Initiative model (DCMI);
2. The UNIDATA Dataset Inventory Catalogue model (DIC model);

3. A profile of the OpenGIS - Simple Feature specification;
4. A profile of the OpenGIS - GridCoverage specification;
5. A profile of the ISO 1900 standard series for Feature;
6. A profile of the ISO 1900 standard series for Coverage.

Others may be considered in future (e.g. FGDC - SDTS etc.).

The annex A reports an UML-based formalisation of the UNIDATA Dataset Inventory Catalogue model.

Sharing the Knowledge

The sharing of the application schema knowledge is obtained following two main approaches:

1. Adopting a quite simple and general application schema (i.e. the GR schema): the schema considers very broad –as far as semantics- and general concepts, such as: information resource, feature, document, etc.
2. Adopting standardised reference schema for structuring complex contents.

Common Encoding/Decoding Rules

As far as Geospatial Information is concerned, THREDDS will support the following encoding/decoding rules:

Common Application Schema	Encoding/Decoding Schema
DCMI Schema	RDF/XML
DCMI Schema	HTML 3.0

DCMI Schema	HTML 4.0
DIC Schema	DIV encoding rules/XML
OpenGIS - Grid Coverage Schema	ISO 19118/XML
OpenGIS - SimpleFeature Schema	OpenGIS GML 2.0/XML
ISO 1900 Profile for Feature	ISO 19118/XML
ISO 1900 Profile for Coverage	ISO 19118/XML

The DIV encoding rules/XML specification is available at the address:
http://www.unidata.ucar.edu/projects/THREDDS/tech/InvCatalog.html

Prioritisation

For implementation, the following table shows a possible criterion for content schema prioritisation:

Common Application Schema	Encoding/Decoding Schema	Priority
DIC Schema	DIC encoding rules/XML	1
DCMI Schema	RDF/XML	2
DCMI Schema	HTML 4.0	3
DCMI Schema	HTML 3.0	4
OpenGIS - Grid Coverage Schema	ISO 19118/XML	5
ISO 1900 Profile for Coverage	ISO 19118/XML	6
OpenGIS - SimpleFeature Schema	OpenGIS GML 2.0/XML	7
ISO 1900 Profile for Feature	ISO 19118/XML	8

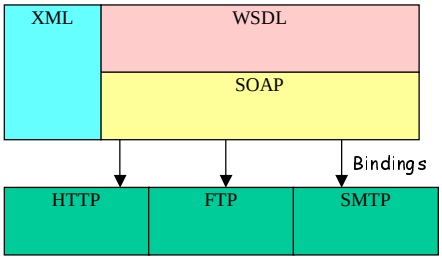
Sharing the Knowledge

The sharing of the encoding rules knowledge is obtained following two main approaches:

1. Realising the specification documentation about the DCV encoding rules (e.g. XML schema specification).
2. Adopting standardised encoding schema for encoding rules (e.g. RDF, HTML 4.0, GML 2.0/XML, ISO 19118/XML).

Common Transfer Protocol

The proposed common transfer protocol is the SOAP protocol along with WSDL. Both are W3C standards. SOAP protocol can be bind with several transport protocol such as: HTTP, FTP, SMTP, etc.



Sharing the Knowledge

The sharing of the transfer protocol knowledge is obtained Adopting a standard and neutral technology.

Data Interchange Implementations

As far as the data interchange paradigm is considered, THREDDS should implements the traditional **data transfer model.**

THREDDS data suppliers create dataset to be transferred where the structure and the data content are described according to the common application schema (see the GR schema). The dataset is sent in a transfer format according to the common encoding model.

Therefore, a dataset is predefined and THREDDS requests and receives a copy of the dataset. The request regards the entire dataset as a unique opaque object.

Extensible Interoperability Scenario

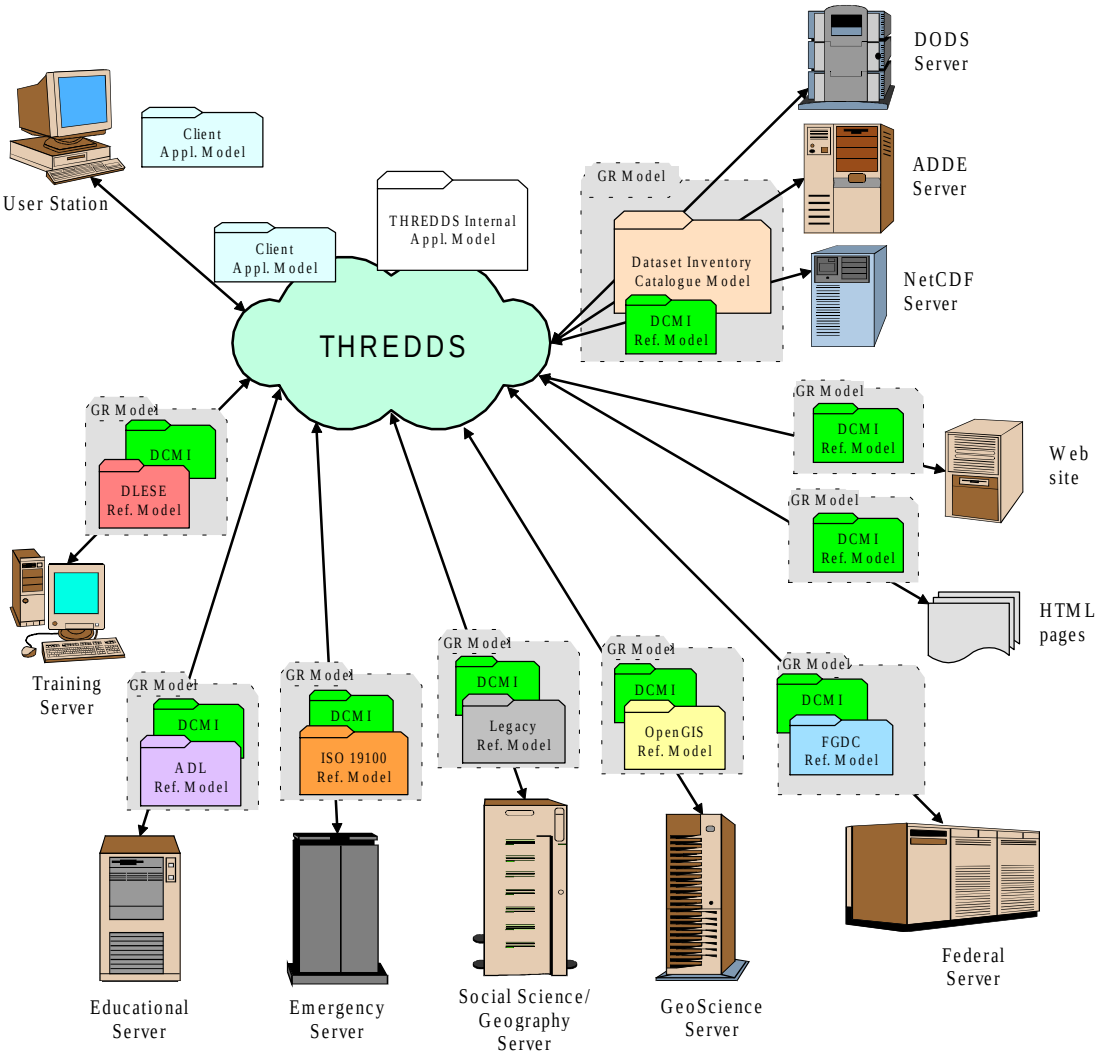
One of the most interesting asset of THREDDS technology is its potentiality to be a scalable and extensible solution to bring together research and educational Communities dealing with different scientific subjects.

This process can be seen as the possibility to easily extend THREDDS system, making it able to discover and access heterogeneous information resources.

The proposed interoperability solution lays the foundation to easily extend THREDDS system.

Extended Scenario

For example, it is possible to conceive a very extended scenario which considers a set of heterogeneous data source types, as depicted in the figure.



THREDDS Extended interoperability view for GI and Educational resources

Each server, depicted in the figure, is characterised by its own application schema; server is able to map from its application schema instances to GR schema instances.

The THREDDS system is able to map from its internal application schema to the Client SW application schema (e.g. MetApps application schema).

For the resource discovery domain, we assumed that every resource description is compliant with Dublin Core (DC) reference model (i.e. green package). Resource content description reports the reference model utilised to conceptualise the content application model.

In this scenario, we assumed that every content schema is based on a different standardised reference model (i.e. multicolour packages).

Moreover, the particular case of a legacy model is also represented; for special purposes or study cases it is also possible to implement a customised application domain interface.

The vision is that THREDDS should be able to discover and access any information resource content whose application domain is expressed according to the THREDDS common application model (i.e. the GR schema).

New Public Available Specifications can be added when necessary or when new international de-facto standards arise.

Requirements

In summary, the main interoperability requirements to extend the system are:

1. Data suppliers must map their content according to the common GR model; that is:
   A. Resources metadata must be compliant with DCMI specifications for networked information resource (other specification can be added in future);
   B. Resources content application schema must be expressed on the top of a *well-known* (i.e. common) set of reference models;
2. THREDDS internal application model must be able to conceptualise all the *well-known* (i.e. common) reference models data structures;
3. The *well-known* set of reference model must be easily extensible.

These requirements call for a particularly *expert* internal THREDDS application model.

THREDDS Internal Application Model

THREDDS Interoperability Services

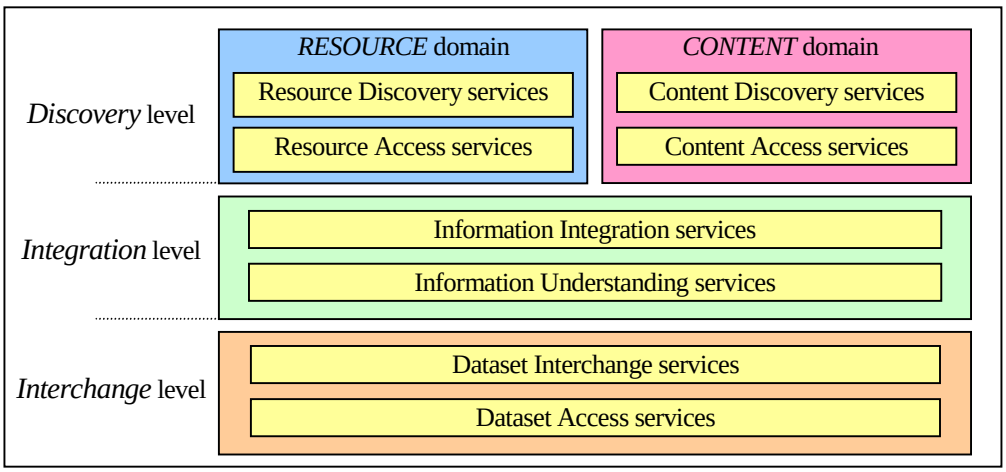
A THREDDS application must be able to *import* the transferred information in order to generate knowledge; that is possible by mapping the data structures of the common model into the THREDDS internal application schema concepts.

Therefore, the THREDDS application schema must be at least as general as the common application schema (see the GR schema); moreover, it must be able to integrate digital library domain concepts with Geospatial information concepts, in a computational framework.

Furthermore, the THREDDS system is required to perform several other interoperability services (e.g. discovered content browsing and query, dataset presentation, etc.), which can be subdivided into useful main categories.

Service Architecture

As far as interoperability objectives are concerned, it is possible to introduce the following services architecture overview for the THREDDS application:



Interoperability Service Architecture Overview

In particular, three services levels are outlined:

1. Resource Discovery level;
2. Information Integration level;
3. Dataset Interchange level.

Two main information domains are reported:

1. Resource domain;
2. Resource Content domain.

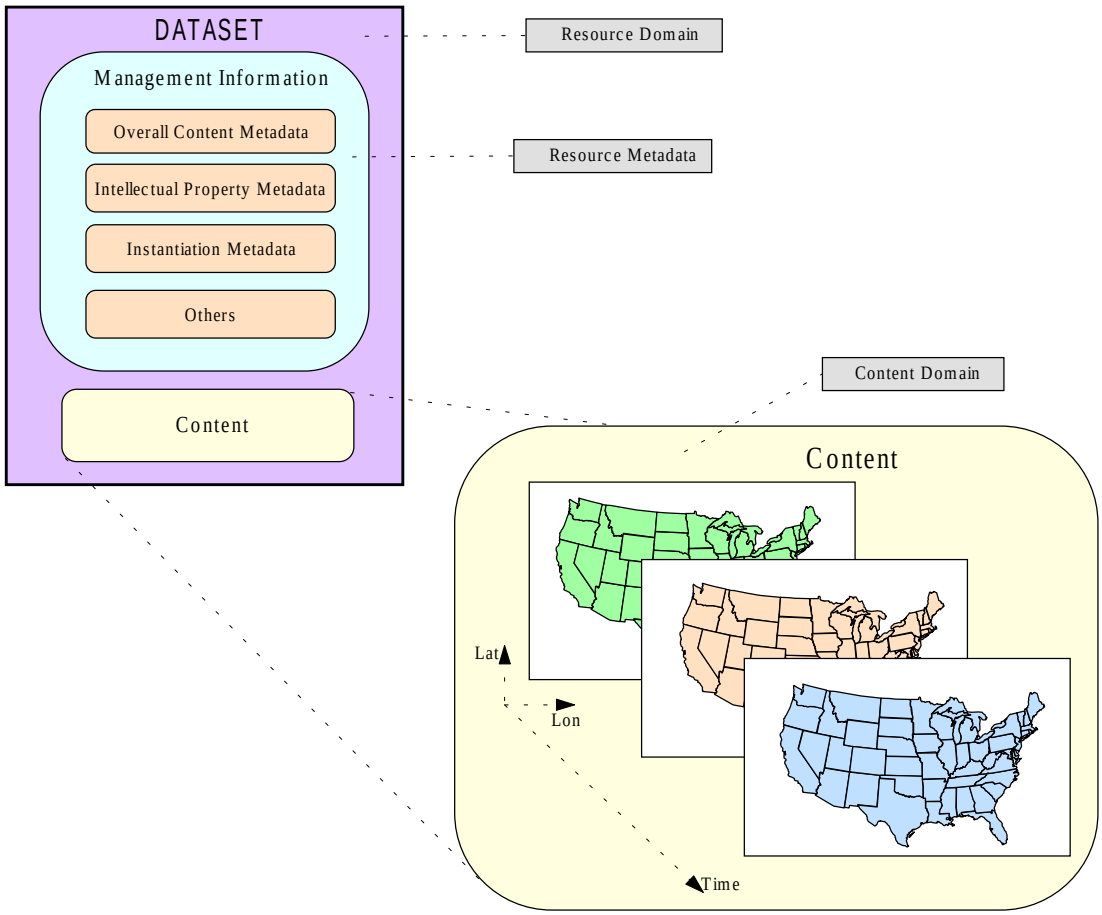
Examples of Information Resource/Content granularity

It is possible to consider information resource examples at different levels of granularity, such as: a dataset, a collection of dataset, a catalogue of collection, a registry of catalogues, etc.

The corresponding resource contents, are: the logical information content managed by the dataset, the logical information content managed by the collection, the logical information managed by the registry, and the registry content.

Therefore, the resource domain deals with the *management* of information (i.e. the content), including a description of managed information, according to the management point of view. In such perspective, the managed information (i.e. the content) is treated as a whole; the description of its internal structure and meaning constitutes the Content Domain.

A typical example of THREDDS resource is depicted in the following diagram.



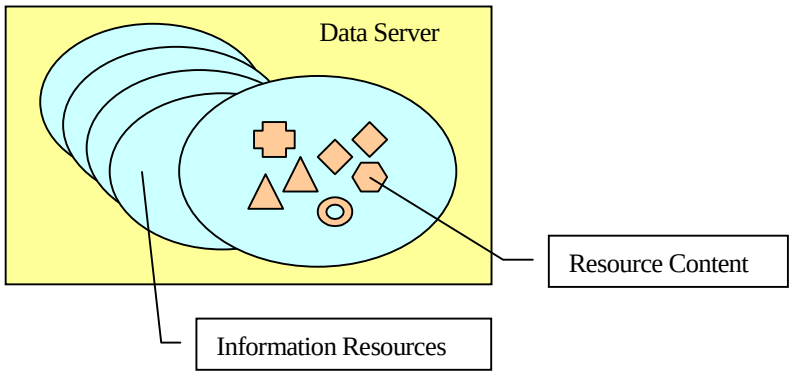
Internal Model Information Domains

Resource Domain

According to the DCMI, with the (Information) resource concept, it is to intend any entity characterized by a URI and which manages a logical collection of information –i.e. the

resource content: from library catalogues and world-wide directories to aggregations of dataset, maps, software, and content to personal collections of multimedia data.

A data server can contain several different information resources, as shown in the following figure.



Information Resources Domain

For THREDDS, the resource domain deals with at least four important aspects:

1. the description of the resource itself (e.g. scope, title, subject, Instantiation, intellectual property, etc.);
2. the description of managed information content, including the reference model utilised for the content modelling -where necessary.
3. the reference to the encoding schema utilized for creating dataset.
4. The structure of interfaces to achieve resource data transfer or data transactions.

It is possible to divide networked information resources into two broad categories:

1. Resource whose content is structured according to a reference model;
2. Resource whose content is not structured according to any reference model.

Examples of the first category of resources are: information systems, data servers, Databases, XML documents, WSDL-based services.

Example of the second categories of resources may be: web sites, HTML pages, undocumented services.

Naturally, only for the first kind of resource the reference model will be provided by the resource content description.

Resource Content Domain

The resource content domain deals with the structure of data contained in the resource.

According to the THREDDS approach of sharing a common application model knowledge, information resources must model such domain by means of a supported reference model.

Information resource content model describes:

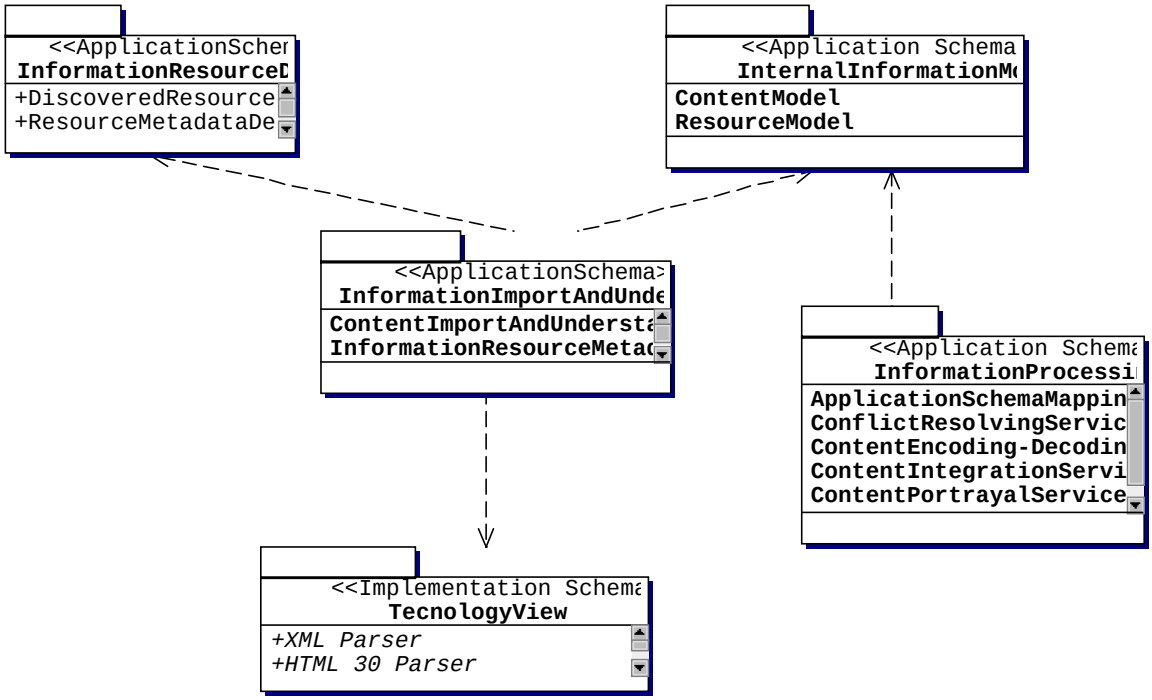
1. The structure and content of datasets to be transferred –modeled according to supported conceptual reference model;
2. The structure of interfaces involved in dataset transfer or transaction.

Internal Application Model Specification

A possible internal application model context (i.e. a contextual model) for THREDDS is herewith drafted. It is possible to access the complete specification at the following URLs:
http://www.unidata.ucar.edu/projects/THREDDS/Nativi/ApplicationModel/ApplicationModelFrameset.htm (Application schema)
http://www.unidata.ucar.edu/projects/THREDDS/Nativi/ApplicationModel/ApplicationModelFrameset.htm (GI content);

The online specification is interactive, and can be navigated as a normal hyper-textual document.

Packages Overview

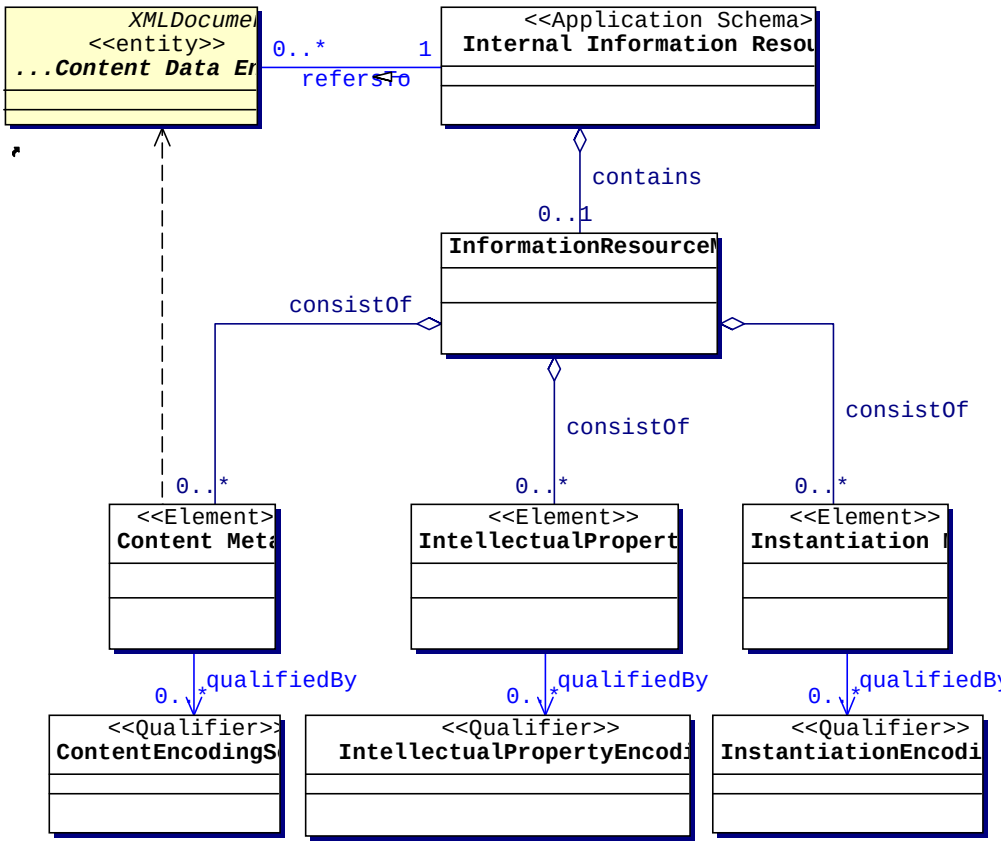


Internal Information Model

Resource Domain

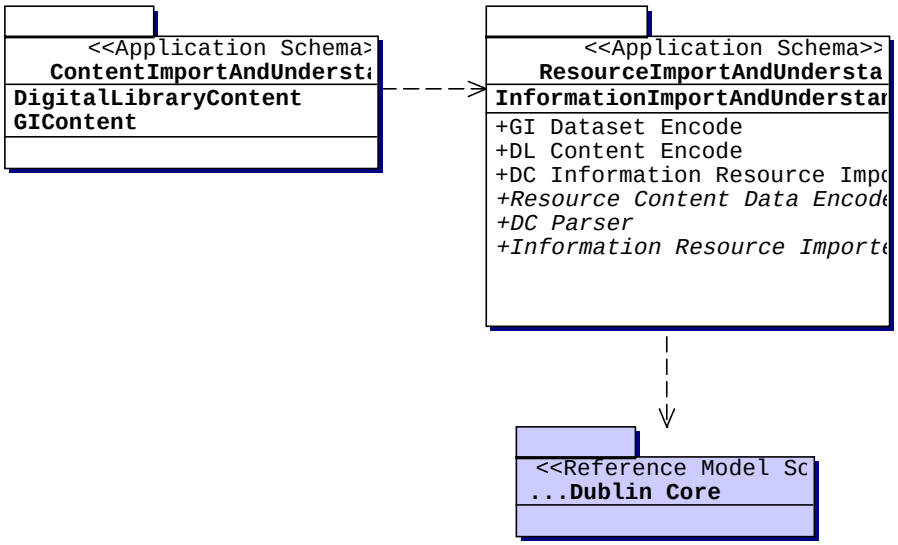
Resource domain conceptualisation has been based on the DC metadata initiative.

The following schema reports the UML-based conceptual schema of the proposed Resource Domain reference model.

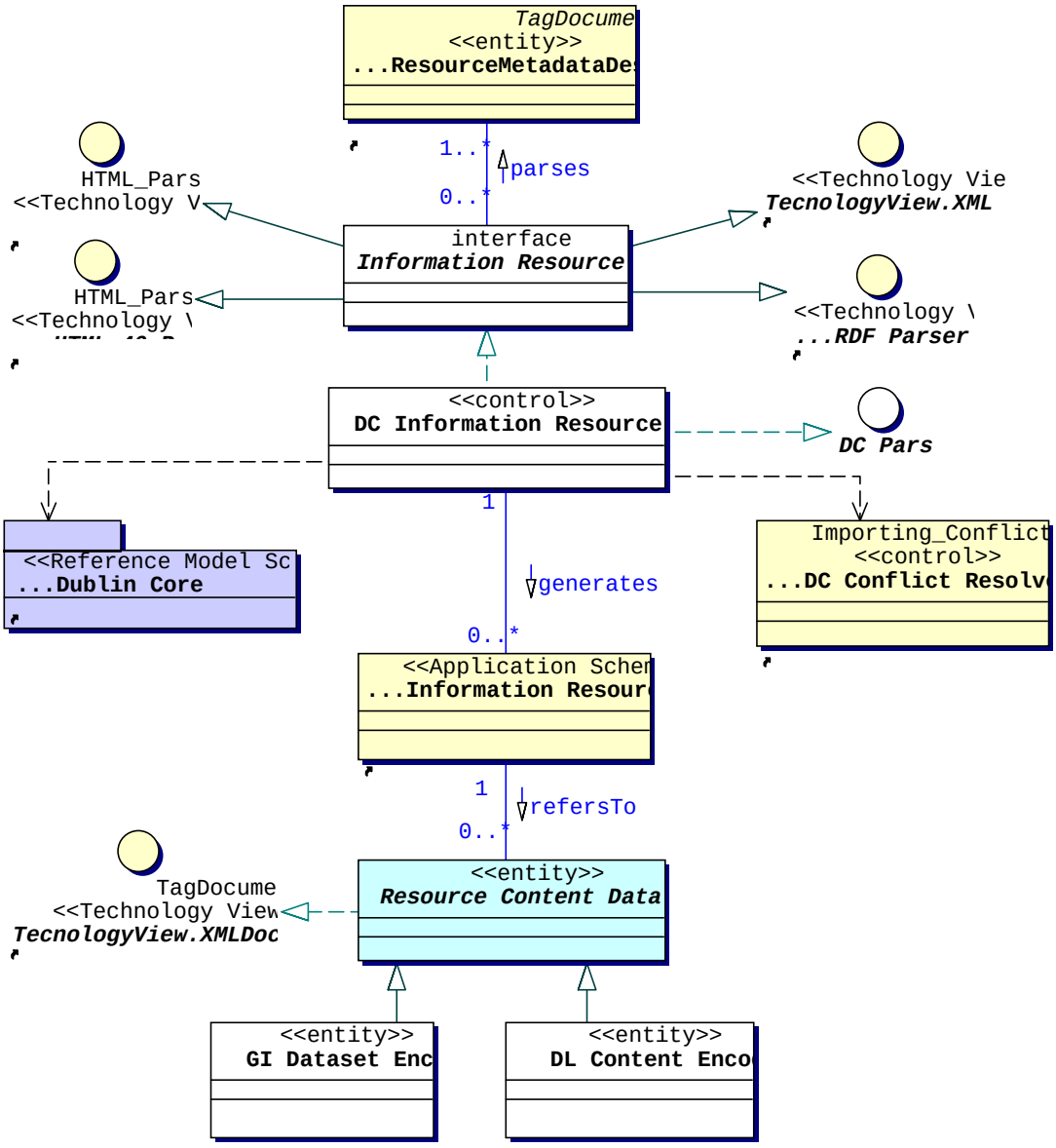


Information Resource conceptual schema

Information Import and Understanding

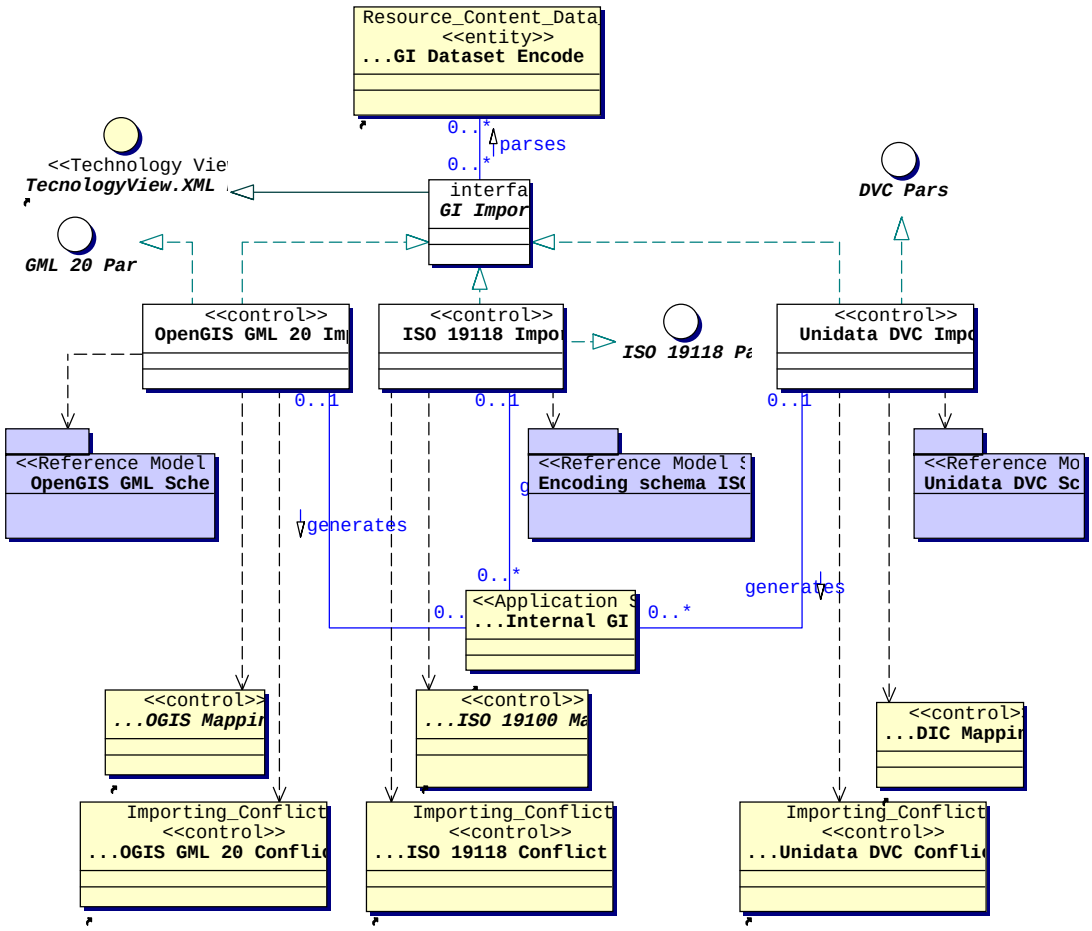


Resource Import and Understanding

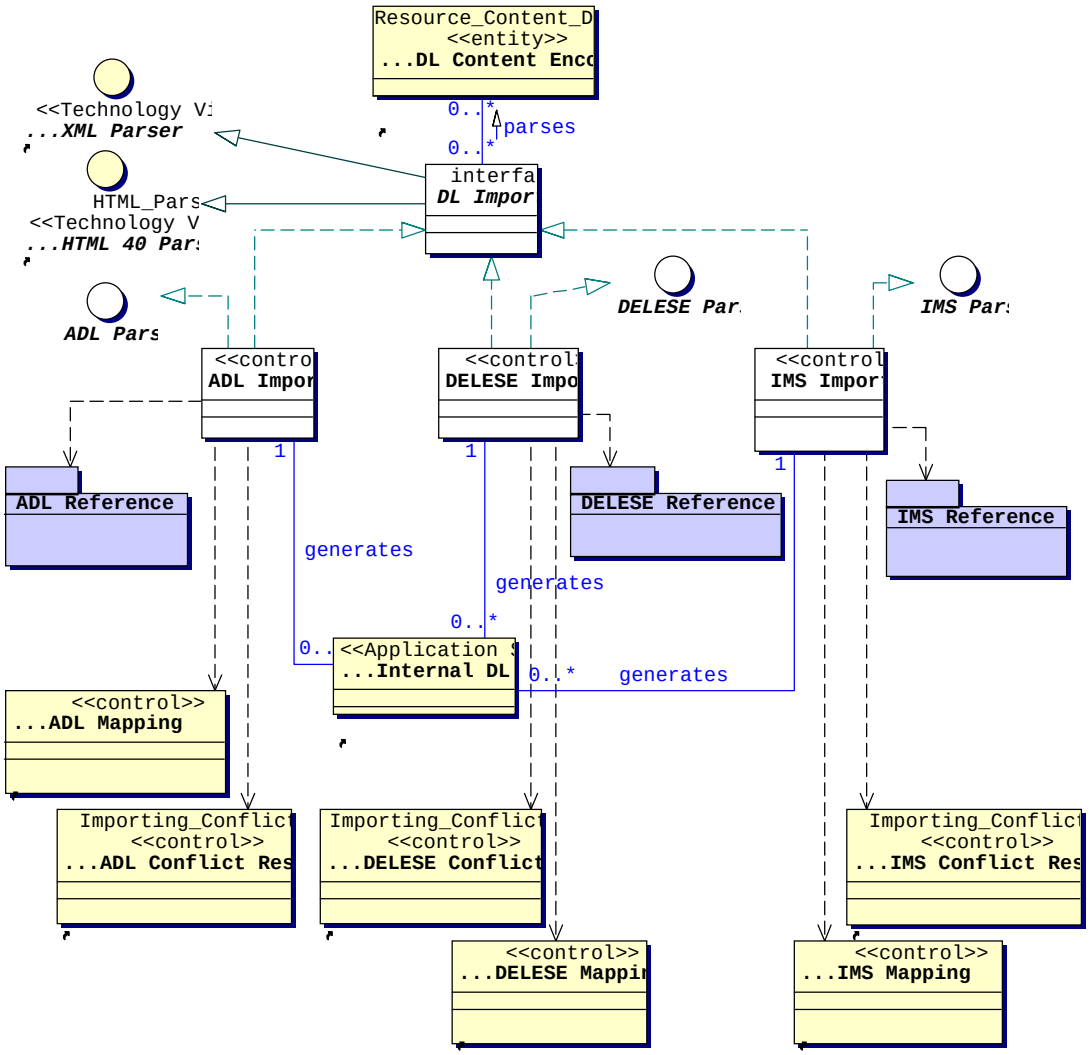


Content Import and Understanding

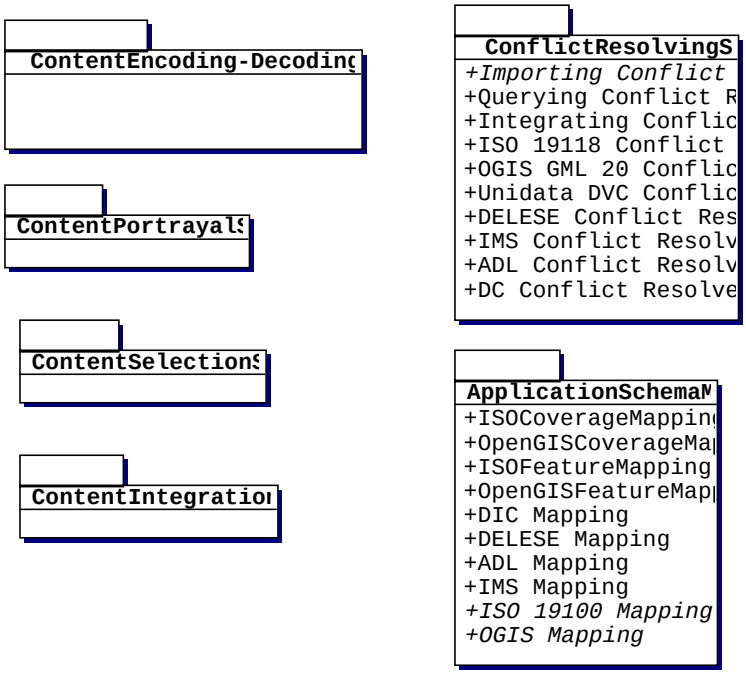
GI Content



Digital Library Content



Information Processing



For more details, go to:

- Application Model Description;
- Application Model Presentation;

Resource Content

Only the GI content domain has been conceptualised.

GI Content

This reference model has been derived from the exhaustive reference models released by the existing international standardisation initiatives (i.e. OpenGIS® and ISO TC211)

The GI notion
According to OGC (the OpenGIS Consortium) “Geospatial information is anything that you can learn by looking at maps -- not just traditional maps, but new, creative, and annotated maps. A map, after all, is simply a metaphor for the Earth itself. We therefore accept raster Earth imagery as a kind of map, and even less structured collections of samples of Earth phenomena with any kind of instrumentation as acceptable maps.”

The fundamental unit of geospatial information is called “feature”.
The definition of a feature is not easy as features can be recursively, so there can be considerable variation in feature granularity.

Example of features –depending on application or information scope- are:

- A segment of a road between consecutive intersections;
- A numbered highway consisting of many road segments;
- A georeferenced satellite image;
- A single pixel of a georeferenced satellite image;
- A temperature overlay on a weather map;
- A triangulated irregular network;
- A dynamically segmented road

- A drainage network;
- A set of flush/flood events;
- elevation contours.

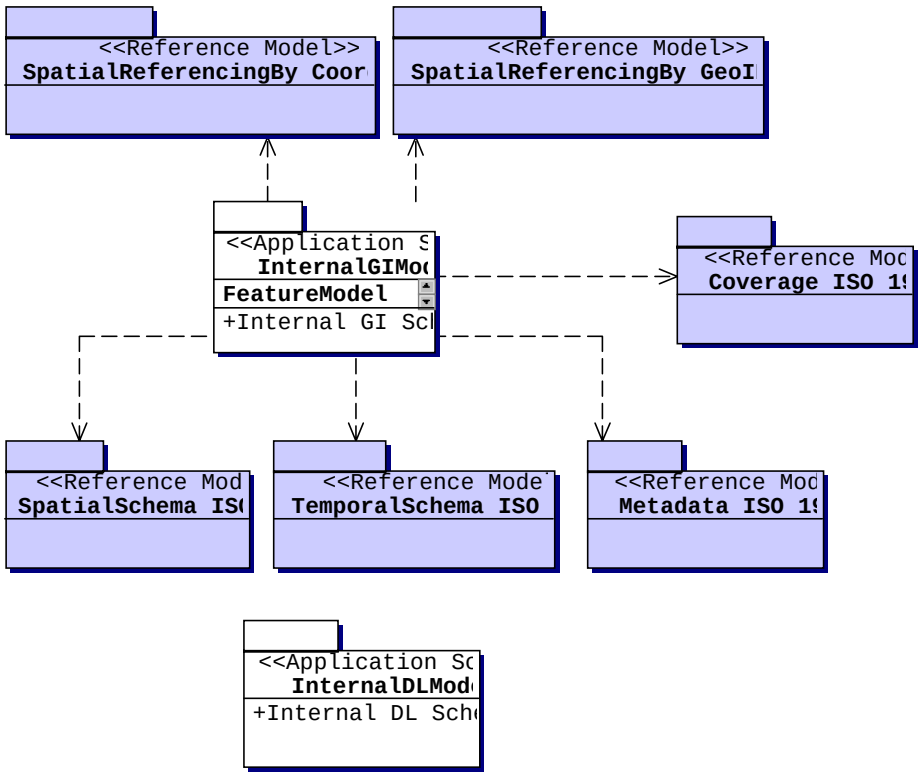
Modelling Features
For reference purpose it is useful to split the geospatial features modelling according two well-accepted technologies:

A. "Features with Geometry" technology:
   such approach models the spatial extent of a feature with point, lines, polygons, and other geometric primitives that come from a list of well-known types. Features modeled in this fashion are typical of GIS-based application and tools.
B. Coverage technology:
   such approach supports mapping from a spatial domain to attribute values, where attribute types are common to all geographic positions within the spatial domain. A spatial domain consists of a -finite/infinite- collection of direct positions in a coordinate space. Examples of coverages include rasters, triangulated irregular networks, point coverages, polygon coverages, volumetric coverages and 4D coverages. Coverages are the prevailing data structures in a number of application areas, such as remote sensing, meteorology, and bathymetric, elevation, soil, and vegetation mapping.

"Features with Geometry" and Coverages are intimately related, yet distinct concepts.

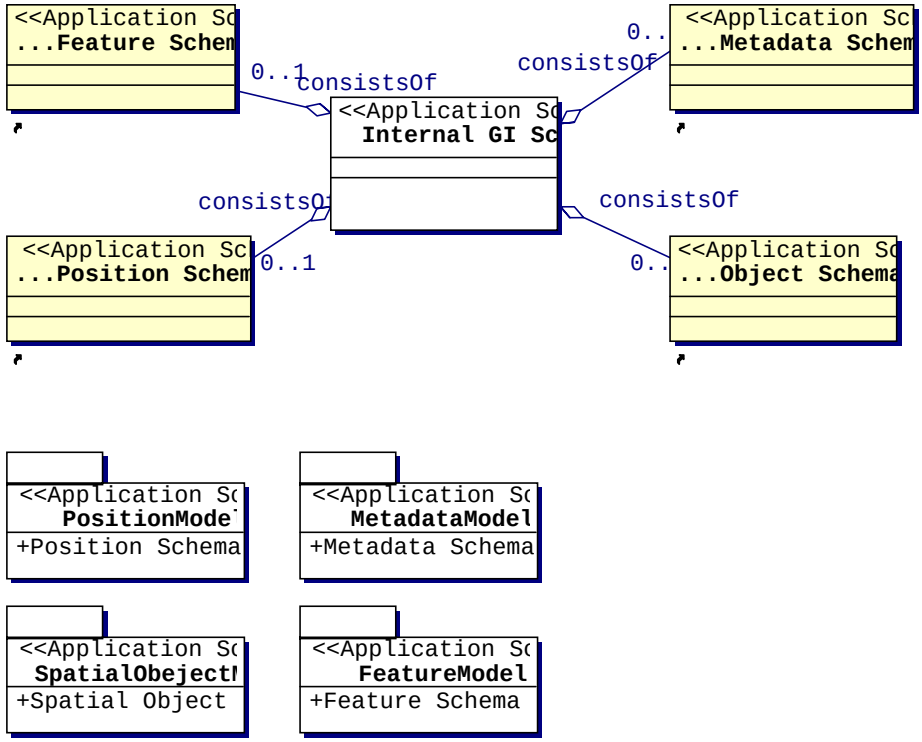
The semantics of Coverage term is often more general than intended here. Besides, Coverage is generally modelled as a specialisation of Feature concept, and it has a proper geometry; therefore it is only for convenience that it is referred as distinct from "Features with Geometry" category.

A General GI content Schema
The following architectural schema reports the different components which constitute a very general GI content model.



Architecture for general GI content model

Internal GI Schema



Extensibility Scientific Community extensibility

- Domain extensibility
- Standardised Reference Model extensibility
- Encoding Languages extensibility
- Transfer Protocol extensibility

Annex A

UNIDATA/THREDDS Dataset Inventory Catalogue Application Model

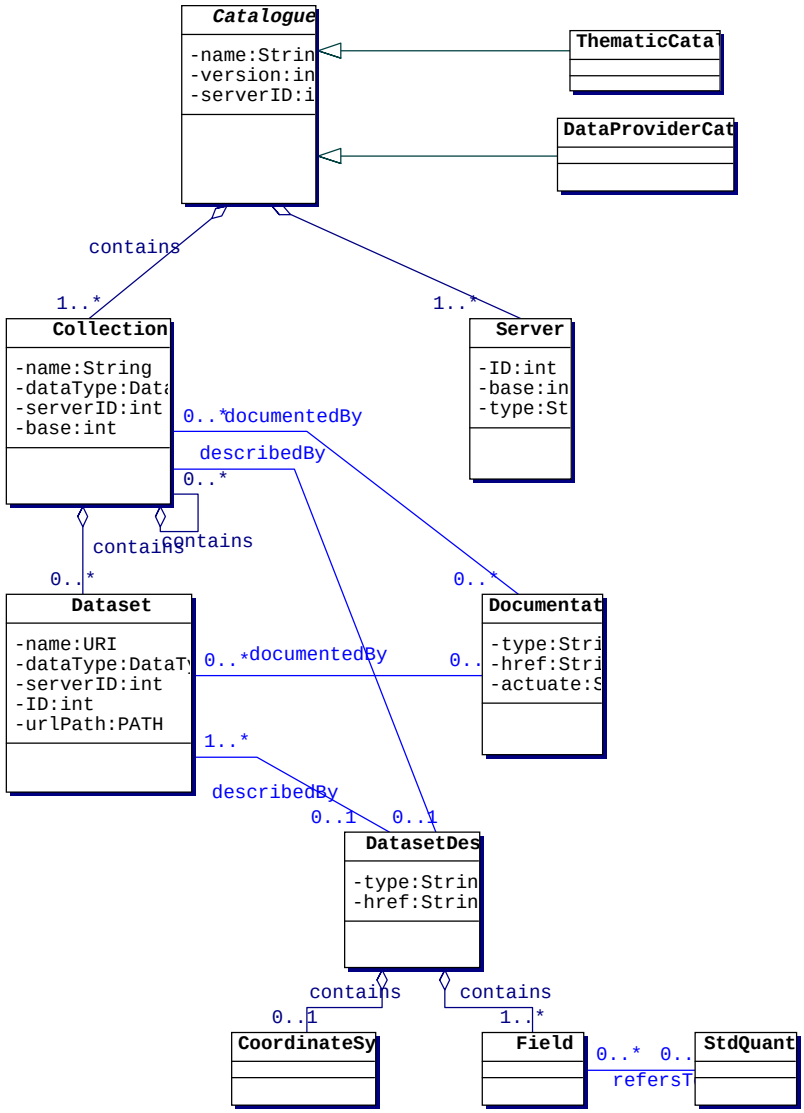
The UNIDATA Reference Model

In the framework of the THREDDS project, an application model -which has become the reference model for UNIDATA enabled data servers- has been released for enabling interoperability among data servers managed by the UNIDATA Community.

Such reference model is based on the inventory catalogue concept. An inventory catalogue is a collection of datasets.

Such dataset are those archived and managed by the following data servers: NetCDF, ADDE, DODS; these servers are supposed to be the main information sources for the THREDDS system.

A schema of such model is depicted in the following figure.

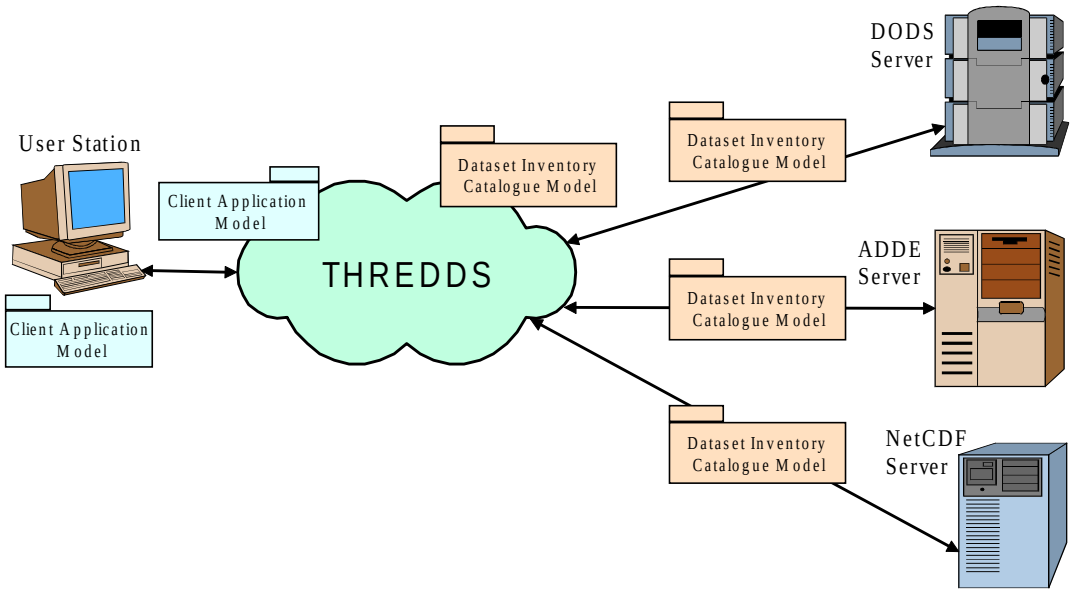


Dataset Inventory Catalogue application schema

The Dataset Inventory Catalogue model is based on the following concepts:

- ***catalog***: the dataset inventory catalog;
- ***collection***: a collection of datasets or nested collections;
- ***dataset***: the atomic unit of selectable content, associated with an URI;
- ***datasetDesc***: a link to a Dataset Description XML document;
- ***documentation***: inline comments and/or a link to an html page for more information;
- ***server***: an on-line source of datasets.

This domain application model allows THREDDS to discover and access traditional data servers, as represented in the following picture.



THREDDS Interoperability View for UCAR traditional servers

In particular, each dataset server is able to map from its own application schema to the Dataset Inventory Catalogue reference model.

The THREDDS system is able to map from the Dataset Inventory Catalogue Model to the Client SW application model (e.g. MetApps application model).

References

ISO/DIS 19101 Geographic Information – Reference Model, © International Organization for Standardization, 2000.

ISO/DIS 19109 Geographic Information – Rules for Application Schema, © International Organization for Standardization, Jul 2001.

J. Caron, Dataset Inventory Catalog Version 0.4, UNIDATA document, http://www.unidata.ucar.edu/projects/THREDDS/tech/InvCatalog.html, 2001.

ISO/DIS 19109 Geographic Information – Rules for Application Schema, © International Organization for Standardization, 2000.

ISO/DIS 19123 Geographic Information – Coverage geometry and functions, © International Organization for Standardization, 2000.

The OpenGIS™ Abstract Specification Topic 5: Features, © Open GIS Consortium, Inc., 1999.

The OpenGIS™ Implementation Specification : Grid Coverage Rev 1.00, © Open GIS Consortium, Inc., Jan 2001.

THREDDS Project, http://www.unidata.ucar.edu/projects/THREDDS.

S. Nativi, A Modelling Conceptual Framework, http://www.unidata.ucar.edu/projects/THREDDS/Nativi/Standards/StandardsFrameset.htm, Oct. 2001

S. Nativi, Notes on THTREDDS Application Model, http://www.unidata.ucar.edu/projects/THREDDS/Nativi/ApplicationModel/ApplicationModelFrameset.htm, Nov. 2001.