DAFNI Project Report

Infrastructure Research Ontologies – Data Discovery

Professor Steve Hallett, Ian Truckell
Cranfield University

Table of Contents

Glossary of Terms *4*

The 'Data Analytics for National Infrastructure' facility - DAFNI *5*

Use Cases for DAFNI in use, and consequent project requirements **5**

The need to classify data by source and subject *6*

DAFNI Data Sources **7**

How data is currently classified and how this links to current data standards **7**

INSPIRE Directive **8**

INSPIRE Topic categories in accordance with EN ISO 19115 ..... **8**

Spatial data themes according to Annex I of the INSPIRE Directive 9

Spatial data themes according to Annex II of the INSPIRE Directive 10

Spatial data themes according to Annex III of the INSPIRE Directive 11

Alternative taxonomic classification systems **13**

EuroSciVoc 14

Global Change Master Directory Keywords (GCMD) 14

Universal Decimal Classification (UDC) 15

Research Fish classifications 16

Research council Joint Electronic Submissions (Je-S) 16

Standard Industrial Classification of economic activities (SIC) 18

The requirement for handling geospatial datasets **18**

Metadata in DAFNI **19**

Adoption of the DCAT-2 MetaData standards and a comparison with UK GEMINI **21**

DCAT-2 mapping to UK GEMINI **22**

Discussion *23*

Conclusions *25*

Recommendations *26*

References *27*

Appendix One – Metadata Standards *28*

MARC Standard **28**

Dublin Core **28**

DCLite4G **30**

FGDC CSDGM **30**

e-GMS **32**

ISO19115 and ISO19139 ..... **32**

UK GEMINI **35**

Data Catalogue Vocabulary (DCAT) - Version 2 38

Comité Européen de Normalisation (CEN) 43

Tables

Table 1 Distinct Use Cases in DAFNI 6
Table 2 Current DAFNI Data Sources 7
Table 3 Current DAFNI Data Headings 7
Table 4 The INSPIRE Directive Topic Categories 8
Table 5 The INSPIRE Directive Annex 1 Themes 9
Table 6 The INSPIRE Directive Annex 2 Themes 10
Table 7 The INSPIRE Directive Annex 3 Themes 11
Table 8 Comparison between DAFNI Subjects and INSPIRE Directive spatial data themes 13
Table 9 EuroSciVoc Top Level Classes 14
Table 10 GCMD Category Keywords 15
Table 11 UDC main tables – top level classification 15
Table 12 Research-Fish Definitions of research databases or models 16
Table 13 Research-Fish Definitions of software or technical products 16
Table 14 NERC Research Areas 17
Table 15 EPSRC Research Area Themes 17
Table 16 DAFNI Metadata Thematic Sections 20
Table 17 Metadata example – DCAT-2 21
Table 18 Metadata mapping – DCAT-2 and UK GEMINI 22
Table 19 Dublin Core Fields 28
Table 20 Metadata standards developed by ISO 32
Table 21 ISO19115 Information Categories 33
Table 22 ISO19115 Fields 34
Table 23 UK GEMINI Version 2.3 Schema 35
Table 24 DCAT-2 Classes 39

Figures

Figure 1 Structure of CSDGM 31
Figure 2 DCAT-2 Classes 40
Figure 3 DCAT Implementation Schema in DAFNI 43

Citation:

Hallett, S.H., Truckell, I.G. (2020) Infrastructure Research Ontologies – Data Discovery. DAFNI Project Report, Cranfield University, 43pp

Acknowledgements:

With thanks to Bethan Perkins and Noel Vizcaino, STFC, for DAFNI-specific advice. Graphics: academic by pam pyke from the Noun Project; decision by Ralf Schmitzer from the Noun Project; Business opportunity by Vectors Market from the Noun Project.

Glossary of Terms

The following terms are either used in the text or are relevant to the themes of the report:

DAFNI Data and Analytics Facility for National Infrastructure programme
DCAT is the main vocabulary for describing datasets.
DCAT-2 is the latest version of DCAT retaining backward compatibility while aligning with schema.org.
DCAT-AP European wide dataset description format that is a profile of DCAT.
FAIR adjective to label the data that meets the principles of findability, accessibility, interoperability, and reusability.
FOSS Free and Open Source.
GDPR EU law transposed to UK law about individuals retaining control of their own personal information.
FOI Freedom Of Information (Request).
GCMD Global Change Master Directory (NASA).
GEMINI A legal requirement to publish geospatial data for public institutions.
INSPIRE European wide version of GEMINI and its superset, also legally required for describing geospatial datasets.
IRI Internationalised URI. A URI can be URL or an URN. PIDs must be IRIs.
IRO Infrastructure Research Ontologies project of DAFNI.
Je-S Joint Electronic Submissions system of UKRI
JSON-LD is a serialisation format based on JSON syntax incorporating linked data abilities.
LWEC Living with Environmental Change
NLP Natural Language Processing.
PID Permanent ID.
RDF Resource Description Framework.
SIC Standard Industrial Classification of economic activities.
SKOS Simple Knowledge Organization System.
STFC Science and Technology Funding Council.
UDC Universal Decimal Classification.
W3C World Wide Web Consortium.

After: https://docs.secure.dafni.rl.ac.uk/docs/reference/dafni-ld-metadata

The 'Data Analytics for National Infrastructure' facility - DAFNI

The Data and Analytics Facility for National Infrastructure (DAFNI, www.dafni.ac.uk), led by Oxford University and operated by the UKRI Science and Technology Funding Council (STFC) in Harwell Oxfordshire, as described by Hall (2019) DAFNI offers a powerful research infrastructure platform able to receive and operate simulation models and visualise the results, as well as to hold and make available many hundreds of substantial 'legacy' data volumes, concerning themes such as built infrastructure networks, environmental, socio-demographic profiles and cadastral information. Combining this capability with high throughput, real-time sources of infrastructure and environmental sensored data opens the route to data-informed decision support tools, such as those embedded within 'Digital Twinning' approaches.

Instrumented and archived data need to be held together in ways that allow its decoupling from the observed systems, permitting scenario modelling of various management strategies and exploration and visualisation of the potential environmental consequences. The DAFNI facility offers a platform that can provide this, it will be possible to ingest real-time sensor data to be placed alongside extant static data volumes, with all these being then being made directly available to suites of modelling and visualisation tools. However, to enable timely and efficient use of these data resources, and to enable their compliance with the 'FAIR' data principles (Findability, Accessibility, Interoperation, Reuse)[1], then the use of data 'ontologies' are required. Ontologies permit the "means to formally model the structure of a system, i.e., the relevant entities and relations that emerge from its observation, and which are useful to our purposes" (Guarino et al, 2009), whereby "an ontology is a formal, explicit specification of a shared conceptualization" (Studer et al., 1998). Thus, when approaching a data intensive modelling task, for example in the area of infrastructure management, then related ontologies can be consulted and collated as enablers.

Substantive datasets are included within DAFNI, and an approach is required to facilitate data discovery, and linking of key datasets in modelling applications. The DAFNI Infrastructure Research Ontologies (IRO) project seeks to provide a framework to achieve this. Within this scope of this is the express need to review (1) the ontologies used to serve and improve the data classification and search tools (taxonomies), and to develop (2) an approach for metadata descriptions for data holdings.

Use Cases for DAFNI in use, and consequent project requirements

It may be considered firstly how the DAFNI system can be accessed, and the various models of use that are anticipated for this, within the scope of infrastructure modelling. The following lists a series of distinct Use Cases for DAFNI, embracing academic researchers, policy users and commercial interests, Table 1.

[1] https://www.go-fair.org/fair-principles/

Table 1 Distinct Use Cases in DAFNI

1.	**Academic Researcher** 	A user seeking to develop a *new* infrastructure using data from the platform. They wish to learn best practices for achieving this but also which datasets are available and how to find and access them and integrate them in models.
2.	**Academic Researcher** 	A user seeking to develop, fork or port an *existing* infrastructure model onto the platform. They wish to learn best practices for achieving this, but also which datasets are available and how to find and access them and integrate them in the model.
3.	**Academic Researcher** 	A user who has developed an external modelling workflow but who is needing to *access datasets*. They need to understand which datasets are available and how best to find and access them.
4.	**Policy-oriented user** 	A user who will access DAFNI as a tool by which to *run a model* they are seeking to be informed about. They will need to know how to select a model how to run it and visualise summary outputs and how to find, access and potentially download results.
5.	**Business user/Commercial interest** 	A user who will access DAFNI as a tool by which to *run a model* they have developed or which they wish to operate. They will need to know how to select a model, how to run it and visualise summary and detailed outputs and how to find, access and potentially download results.

The need to classify data by source and subject

Given the anticipated Use Cases and user categories noted, it is important that there are proper provisions made to enable data discovery, to enable classifications and targeted searches to be made across the data holdings. Of these datasets, a significant number have a spatial/geographical component.

Geographical information is increasingly prevalent in our lives, affecting personal leisure activities as much as the workplace. Geographical information represents a key theme in infrastructure and environmental management, and it may be estimated that some 80% of the data used for environmental, business and policy-oriented decision making is geographical in nature[2]. The exact number in any case is significant. Such spatial data requires a structured approach in its management if the maximum benefit is to be derived from analysis and dissemination.

[2] http://povesham.wordpress.com/2010/02/22/the-source-of-the-assertion-that-80-of-all-organisational-information-is-geographic/

The 'World Wide Web Consortium' (W3C) are responsible for many of the common web standards in widespread use. The key issues stated in Spatial Data on the Web Best Practices, W3C Working Group Note[3] are enabling discoverability, accessibility and interoperability of spatial data with the aim to enable spatial data to be integrated within the wider Web of data; providing standard patterns and solutions that help solve these problems as part of a 'semantic web'.

DAFNI Data Sources

Current DAFNI data has been sourced from a number of providers, Table 2.

Table 2 Current DAFNI Data Sources

Data Source (having 1 or more items)	No. of data items
Office for National Statistics	620
Newcastle University	28
Department for Transport	4
National Grid	4
Ordnance Survey Ltd	2
Office of Rail and Road	2
Companies House	1
DAFNI Workflows	1
Department for Business, Energy & Industrial Strategy	1
Department for Environment, Food & Rural Affairs	1
Historic England	1
Office of Rail and Road	1
INET, Oxford	1
Ordnance Survey Ltd	1
University of Oxford	1

Note: Figures from 5/12/2020, showing principal sources
Several sources are the same, but with slightly different spellings.

How data is currently classified and how this links to current data standards

DAFNI data descriptions are currently described using INSPIRE Topic Categories (see Table 4) to list subjects. Table 3 shows the distribution of the current holdings by these categories, indicating a heavy prevalence of data pertaining to society (ONS data).

Table 3 Current DAFNI Data Headings

Subject (INSPIRE Topic Categories)	No. of data items
Society	621
Transportation	18
Climatology / Meteorology / Atmosphere	16
Utilities/Communication	9

[3] https://www.w3.org/TR/sdw-bp/#bp-indexable

Economy	3
Oceans	2
Environment	1
Biota	0
Boundaries	0
Elevation	0
Farming	0
Health	0
Planning / Cadastre	0
Structure	0
Test being able to publish to NID4	0
Total	***670***

Note: Figures from 5/12/2020

INSPIRE Directive

The existing DAFNI data organisation and classification approaches derive in large part from those outlined in the EU INSPIRE Directive. The INSPIRE Directive[4] aims to "create a European Union spatial data infrastructure for the purposes of EU environmental policies and policies or activities which may have an impact on the environment. This European Spatial Data Infrastructure will enable the sharing of environmental spatial information among public sector organisations, facilitate public access to spatial information across Europe and assist in policy making across boundaries".

> *INSPIRE defines a series of 'Topic Categories' and 'Themes'. DAFNI currently uses the Topic Categories to both classify newly added data items, and for searching out existing data (referred to in DAFNI as 'Subjects'). INSPIRE Themes are also captured on data entry but are not presently utilised directly.*

INSPIRE Topic categories in accordance with EN ISO 19115

Within the INSPIRE Directive there are defined nineteen broad 'Topic Categories' (EN ISO 19115), used to classify newly added data items and to support thematic searches for data. They are as follows[5], Table 4.

Table 4 The INSPIRE Directive Topic Categories

Topic Categories
Biota
Boundaries
Climatology / Meteorology / Atmosphere
Economy
Elevation
Environment

[4] https://inspire.ec.europa.eu

[5] https://inspire.ec.europa.eu/metadata-codelist/TopicCategory and https://apps.usgs.gov/thesaurus/thesaurus-full.php?thcode=15

Farming
Geoscientific Information
Health
Imagery / Base Maps / Earth Cover
Inland Waters
Intelligence / Military
Location
Oceans
Planning / Cadastre
Society
Structure
Transportation
Utilities / Communication

INSPIRE is based on the infrastructures for spatial information established and operated by the Member States of the European Union. The Directive further addresses 34 spatial data themes needed for environmental applications.

Datasets in scope of INSPIRE are those which fall under one or more of the 34 spatial data themes set out in the INSPIRE Directive. Interoperability in INSPIRE means the possibility to combine spatial data and services from different sources across the European Community in a consistent way without involving specific efforts of humans or machines. Interoperability may be achieved by either changing (harmonising) and storing existing data sets or transforming them via services for publication in the INSPIRE infrastructure. DAFNI captures INSPIRE themes for newly entered data items, but this is presently not used subsequently.

Annexes 1 to 3 of the INSPIRE Directive contain the themes of data which have been successively addressed with data specifications to harmonise sources of 'interoperable' data across member states. The themes are as follows:

Spatial data themes according to Annex I of the INSPIRE Directive

The following 9 themes comprise INSPIRE Annex 1, Table 5.

Table 5 The INSPIRE Directive Annex 1 Themes

Annex 1 INSPIRE Themes		Description
1.	**Coordinate reference systems**	Systems for uniquely referencing spatial information in space as a set of coordinates (x, y, z) and/or latitude and longitude and height, based on a geodetic horizontal and vertical datum.
2.	**Geographical grid systems**	Harmonised multi-resolution grid with a common point of origin and standardised location and size of grid cells.

3.	**Geographical names**	Names of areas, regions, localities, cities, suburbs, towns or settlements, or any geographical or topographical feature of public or historical interest.
4.	**Administrative units**	Units of administration, dividing areas where Member States have and/or exercise jurisdictional rights, for local, regional and national governance, separated by administrative boundaries.
5.	**Addresses**	Location of properties based on address identifiers, usually by road name, house number, postal code.
6.	**Cadastral parcels**	Areas defined by cadastral registers or equivalent.
7.	**Transport networks**	Road, rail, air and water transport networks and related infrastructure. Includes links between different networks. Also includes the trans-European transport network as defined in Decision No 1692/96/EC of the European Parliament and of the Council of 23 July 1996 on Community Guidelines for the development of the trans-European transport network (1) and future revisions of that Decision.
8.	**Hydrography**	Hydrographic elements, including marine areas and all other water bodies and items related to them, including river basins and sub-basins. Where appropriate, according to the definitions set out in Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 establishing a framework for Community action in the field of water policy (2) and in the form of networks.
9.	**Protected sites**	Area designated or managed within a framework of international, Community and Member States' legislation to achieve specific conservation objectives.

Spatial data themes according to Annex II of the INSPIRE Directive

The following 4 themes comprise INSPIRE Annex 2, Table 6.

Table 6 The INSPIRE Directive Annex 2 Themes

Annex 2 Themes		Description
1.	**Elevation**	Digital elevation models for land, ice and ocean surface. Includes terrestrial elevation, bathymetry and shoreline.
2.	**Land cover**	Physical and biological cover of the earth's surface including artificial surfaces, agricultural areas, forests, (semi-)natural areas, wetlands, water bodies.
3.	**Orthoimagery**	Geo-referenced image data of the Earth's surface, from either satellite or airborne sensors.
4.	**Geology**	Geology characterised according to composition and structure. Includes bedrock, aquifers and geomorphology.

Spatial data themes according to Annex III of the INSPIRE Directive

The following 21 themes comprise INSPIRE Annex 3, Table 7.

Table 7 The INSPIRE Directive Annex 3 Themes

Annex 3 Themes		Description
1.	**Statistical units**	Units for dissemination or use of statistical information.
2.	**Buildings**	Geographical location of buildings.
3.	**Soil**	Soils and subsoil characterised according to depth, texture, structure and content of particles and organic material, stoniness, erosion, where appropriate mean slope and anticipated water storage capacity.
4.	**Land use**	Territory characterised according to its current and future planned functional dimension or socio-economic purpose (e.g., residential, industrial, commercial, agricultural, forestry, recreational).
5.	**Human health and safety**	Geographical distribution of dominance of pathologies (allergies, cancers, respiratory diseases, etc.), information indicating the effect on health (biomarkers, decline of fertility, epidemics) or well-being of humans (fatigue, stress, etc.) linked directly (air pollution, chemicals, depletion of the ozone layer, noise, etc.) or indirectly (food, genetically modified organisms, etc.) to the quality of the environment.
6.	**Utility and governmental services**	Includes utility facilities such as sewage, waste management, energy supply and water supply, administrative and social governmental services such as public administrations, civil protection sites, schools and hospitals.
7.	**Environmental monitoring facilities**	Location and operation of environmental monitoring facilities includes observation and measurement of emissions, of the state of environmental media and of other ecosystem parameters (biodiversity, ecological conditions of vegetation, etc.) by or on behalf of public authorities.
8.	**Production and industrial facilities**	Industrial production sites, including installations covered by Council Directive 96/61/EC of 24 September 1996 concerning integrated pollution prevention and control (1) and water abstraction facilities, mining, storage sites.
9.	**Agricultural and aquaculture facilities**	Farming equipment and production facilities (including irrigation systems, greenhouses and stables).
10.	**Population distribution — demography**	Geographical distribution of people, including population characteristics and activity levels, aggregated by grid, region, administrative unit or other analytical unit.

11.	**Area management/restriction/ regulation zones and reporting units**	Areas managed, regulated or used for reporting at international, European, national, regional and local levels. Includes dumping sites, restricted areas around drinking water sources, nitrate-vulnerable zones, regulated fairways at sea or large inland waters, areas for the dumping of waste, noise restriction zones, prospecting and mining permit areas, river basin districts, relevant reporting units and coastal zone management areas.
12.	**Natural risk zones**	Vulnerable areas characterised according to natural hazards (all atmospheric, hydrologic, seismic, volcanic and wildfire phenomena that, because of their location, severity, and frequency, have the potential to seriously affect society), e.g., floods, landslides and subsidence, avalanches, forest fires, earthquakes, volcanic eruptions.
13.	**Atmospheric conditions**	Physical conditions in the atmosphere. Includes spatial data based on measurements, on models or on a combination thereof and includes measurement locations.
14.	**Meteorological geographical features**	Weather conditions and their measurements; precipitation, temperature, evapotranspiration, wind speed and direction.
15.	**Oceanographic geographical features**	Physical conditions of oceans (currents, salinity, wave heights, etc.).
16.	**Sea regions**	Physical conditions of seas and saline water bodies divided into regions and sub-regions with common characteristics.
17.	**Bio-geographical regions**	Areas of relatively homogeneous ecological conditions with common characteristics.
18.	**Habitats and biotopes**	Geographical areas characterised by specific ecological conditions, processes, structure, and (life support) functions that physically support the organisms that live there. Includes terrestrial and aquatic areas distinguished by geographical, abiotic and biotic features, whether entirely natural or semi-natural.
19.	**Species distribution**	Geographical distribution of occurrence of animal and plant species aggregated by grid, region, administrative unit or other analytical unit.
20.	**Energy resources**	Energy resources including hydrocarbons, hydropower, bio-energy, solar, wind, etc., where relevant including depth/height information on the extent of the resource.
21.	**Mineral resources**	Mineral resources including metal ores, industrial minerals, etc., where relevant including depth/height information on the extent of the resource.

It is useful to consider how these themes can match with the existing classification, Table 8.

Table 8 Comparison between DAFNI Subjects and INSPIRE Directive spatial data themes

'Subject' from DAFNI	Possible relevant INSPIRE Theme
Biota	- Habitats and biotopes - Species distribution
Boundaries	- Cadastral parcels - Administrative units - Area management/restriction/regulation zones and reporting units
Climatology / Meteorology / Atmosphere	- Atmospheric conditions - Meteorological geographical features
Economy	Economic factors
Elevation	Elevation
Environment	Many themes including Soil, Land cover, Geology, Environmental monitoring Facilities etc.
Farming	- Agricultural and aquaculture facilities - Land Use - Land cover
Health	Human health and safety
Oceans	- Oceanographic geographical features - Sea regions - Hydrography
Planning / Cadastre	- Cadastral parcels - Addresses - Buildings
Society	Population distribution and demography
Structure	- Buildings - Production and industrial facilities
Transportation	Transport networks
Utilities / Communication	Utility and governmental services

A classification of 372 entities (e.g., '*building*', '*duct*', '*drainage basin*' etc) has been placed within this feature concept classification[6], thus '*building*' -> '*buildings*', '*duct*' -> '*Utility and government services*', '*drainage basin*' -> '*Hydrography*' etc.

Alternative taxonomic classification systems

At present, DAFNI only utilises the nineteen Topic Categories to classify data additions. A part of this report then is to review potential alternative taxonomies that could be deployed that can retain a high-level simplicity yet have enough hierarchical granularity to be of use in characterising more accurately the majority of datasets anticipated for the system. Further to the classification taxonomies in INSPIRE, there

[6] https://inspire.ec.europa.eu/featureconcept

are a number of other classification taxonomies that can also be considered for use in partitioning the DAFNI data holdings and aiding their discoverability.

A taxonomy represents "a controlled vocabulary in which all the terms belong to a single hierarchical structure and have parent/child or broader/narrower relationships to other terms. The structure is sometimes referred to as a 'tree'. The addition of non-preferred terms/synonyms may or may not be part of a taxonomy"[7].

EuroSciVoc

European Science Vocabulary (EuroSciVoc) is the 'taxonomy of fields of science based on OECD's 2015 Frascati Manual taxonomy. EuroSciVoc was extended with fields of science categories extracted from CORDIS content through a semi-automatic process developed with Natural Language Processing (NLP) techniques. Described online at: https://op.europa.eu/en/web/eu-vocabularies/at-dataset/-/resource/dataset/euroscivoc?target=Documentation and http://publications.europa.eu/resource/dataset/euroscivoc

EuroSciVoc comprises a hierarchical taxonomy of 949 terms which at the highest level includes the following six classes, Table 9.

Table 9 EuroSciVoc Top Level Classes

EuroSciVoc Top Level Classes
agricultural sciences
engineering and technology
humanities
medical and health sciences
natural sciences
social sciences

EuroSciVoc could be usefully adopted in DAFNI as its primary (or secondary) taxonomy as it embraces concepts as broad as social sciences, engineering, smart city, environmental issues, GIS and philosophy (ontologies). Being derived from the EU CORDIS scientific reference, EuroSciVoc covers all scientific disciplines, and has a granular and hierarchical classification. A SKOS RDF schema also exists for EuroSciVoc that can be used with the DCAT-2 metadata standard.

Global Change Master Directory Keywords (GCMD)

NASA maintains the Global Change Master Directory (GCMD) Keywords[8]. This is 'a hierarchical set of controlled Earth Science vocabularies that help ensure Earth science data, services, and variables are described in a consistent and comprehensive manner', allowing for the 'precise searching of metadata and subsequent retrieval of data, services, and variables'. The categories of GCMD Keywords are as follows (Table 10).

[7] https://op.europa.eu/en/web/eu-vocabularies/taxonomies
[8] https://earthdata.nasa.gov/earth-observation-data/find-data/gcmd/gcmd-keywords

Table 10 GCMD Category Keywords

GCMD Category Keywords
Earth Science
Earth Science Services
Data Centers/Service Providers
Projects
Instruments/Sensors
Platforms/Sources
Locations
Horizontal Data Resolution
Vertical Data Resolution
Temporal Data Resolution
URL Content Types
Granule Data Formats
Measurement Names
Chronostratigraphic Units

There is little in this classification that extends to the infrastructure themes held in DAFNI.

Universal Decimal Classification (UDC)

Universal Decimal Classification (UDC) is a bibliographic system used by many libraries to organise books and other materials (such as maps) by subject[9]. The system covers disciplines, artefacts and territories (e.g., countries). The system supports linked data expressions. The system also has a strongly hierarchical classification, e.g., '*656.1 Road transport*', '*631.4 Soil science. Pedology. Soil research*' and '*625.1. Railways in general. Permanent way. Track construction*'. Table 11 presents the main classifications.

Table 11 UDC main tables – top level classification

UDC Classes
0 science and knowledge. organization. computer science. information. documentation. librarianship. institutions. publications
1 philosophy. psychology
2 religion. theology
3 social sciences
5 mathematics. natural sciences
6 applied sciences. medicine. technology
7 the arts. recreation. entertainment. sport
8 language. linguistics. literature
9 geography. biography. history

[9] http://www.udcc.org/index.php/site/page?view=about

Although the topics in UDC do cover the theme of interest in DAFNI, they are designed for bibliographic reference and omit many properties that potential users may seek.

Research Fish classifications

When academics complete annual entries in Research Fish as part of the UKRI returns[10], it will be noted that there are a number of taxonomic classifications used that could be of interest to DAFNI, and with a familiarity for academic users. Examples of these classifications are shown in Table 12 and Table 13.

Table 12 Research-Fish Definitions of research databases or models

Select the type of research database or model
Database/Collection of Data
Data Analysis Technique
Computer model/algorithm
Data handling & control

Table 13 Research-Fish Definitions of software or technical products

Select type of software or technical product
Webtool/Application
Software
e-Business platform
Grid Application
Physical Model/Kit
New Material/Compound
New/Improved Technique/Technology
Systems, Materials & Instrumental Engineering
Detection Devices

It is considered that ResearchFish does not have the discrimination required to describe DAFNI datasets.

Research council Joint Electronic Submissions (Je-S)

In completing any proposed research project on the UKRI JES system, academics will encounter the URKI classification system. This hierarchical system allows documents to be referenced by discipline. There is a classification for each of the Councils, for example NERC[11], and EPSRC[12], see Table 14 and Table 15.

[10] https://app.researchfish.com/
[11] https://nerc.ukri.org/funding/application/howtoapply/topics/
[12] https://epsrc.ukri.org/research/ourportfolio/researchareas/

Table 14 NERC Research Areas

Research area
Archaeology
Atmospheric physics & chemistry
Climate & climate change
Ecology, biodiversity & systematics
Genetics & development
Geosciences
Marine environments
Medical & health interface
Microbial sciences
Omic sciences & technologies
Planetary science
Plant & crop science
Pollution, waste & resources
Terrestrial & freshwater environments
Tools, technology & methods

Table 15 EPSRC Research Area Themes

Thematic area
Digital economy
Energy
Engineering
Global uncertainties
Healthcare technologies
ICT
LWEC
Manufacturing the future
Mathematical sciences
Physical Sciences
Quantum technologies
Research Infrastructure

These systems will not be familiar to non-academics, although being intuitive and easily understood, and further do not have the discrimination required to describe DAFNI datasets.

Standard Industrial Classification of economic activities (SIC)

The UK Government Office of National Statistics (ONS) promote the Standard Industrial Classification of economic activities (SIC)[13]. The SIC assigns numerical codes to industrial activities, e.g., '*49410. Freight transport by road*'. SIC codes could be useful in DAFNI when developing modelling approaches considering the impacts of infrastructure states on the economy. SIC codes are also useful for providing regional and sub-regional summaries of economic activity. SIC also has a simple hierarchical taxonomy, e.g., '*Section H Transportation and storage*' -> '*49410. Freight transport by road*'.

Although the topics in SIC do cover the theme of interest in DAFNI, they are designed for discriminating business categories and omit many properties that potential users may seek.

The requirement for handling geospatial datasets

The spatial aspects of data held by DAFNI are critical for its success as a data centre, and so demand particular attention. It has been estimated that some 80% of government data can be related to location (a quote that is widely discussed, but which is not unreasonable[14]). Several best practices for publishing and using spatial data on the web have been produced. However, often these may not meet the needs of spatial data, nor address the issues around data complexity and how data may change over time.

They key items for consideration when focussing on spatial data include:

- Spatial things, features and geometry;
- Spatial relationships;
- Coordinate reference systems and datums;
- Linked data;
- Spatial data infrastructures.

The term 'spatial thing' is adopted to mean anything with a spatial extent (position, size etc.) it can be real-world phenomena or their abstractions. Geometry is a property of the spatial thing and can be used to model its spatial extent or shape. Spatial relationships demonstrate how objects are related in space. This is represented as relationships based on topology, direction or distance.

Coordinate reference systems define how the geometry of the object is stored. Often web applications use a geographical coordinate system such as WGS84. However, if projected coordinate systems are adopted it is important to consider the distortions to features or related measurements when coordinate transformations are applied.

---

[13] https://www.gov.uk/government/publications/standard-industrial-classification-of-economic-activities-sic

[14] https://gis.stackexchange.com/questions/18838/80-of-data-has-a-spatial-component-says-who

Linked data uses web technologies to make linking fundamental across distributed data sources. Example ontologies exist where the spatial parameters of the data can be used to link such as geometry or topology. These ontologies can be overarching or domain specific such as those produced by Ordnance Survey[15].
Across the sector there is a general lack of a spatial vocabulary. Information from W3 suggests: "no single standardized vocabulary is available that covers all needs. A possible way forward is an update for the GeoSPARQL spatial ontology. This will provide an agreed spatial ontology, i.e., a bridge or common ground between geographical and non-geographical spatial data and between W3C and OGC standards; conformant to the ISO-19107 abstract model and based on existing available ontologies such as GeoSPARQL" (W3C, 2017).

The items making up the technology to allow information to be easily shared include the Resource Description Framework (RDF) data model, the OWL web ontology model and the RDF query language such as SPARQL. SPARQL can be used to generate simple queries from graph pattern matching to more complex queries.

GeoSPARQL[16] was developed to deal with querying geospatial data. It builds on SPARQL to include components such as a topological vocabulary to allow for topological relationships between spatial data and geometry components to allow for storage or transmission of geometry objects.

It is important to consider these best practices in relation to the storage and use of spatial data in DAFNI. The methods adopted should create a framework to facilitate data discovery and linking of key datasets, therefore meeting the needs of the Use Cases as described previously and related project requirements.

Metadata in DAFNI

Metadata may be described as "data about data" (Hillman, 2005). The National Information Standards Organization describes metadata as "structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource." (NISO, 2004). Geospatial data metadata can be used to describe data or organise and manage a database. It can also be used to publish an organisation's data holdings to clearing houses, geo-portals and data catalogues; and provide information to aid transferring of data.

Levels of metadata vary from '*discovery level*' (the basic form of metadata with which to find out the potentially available data that meets some specific criteria during a search), '*evaluation level*' (checking the quality of the information to ascertain the fitness of the data for the intended purpose) to '*usage level*' (information to use to configure a system to process the data).

[15] http://data.ordnancesurvey.co.uk/ontology
[16] https://www.ogc.org/standards/geosparql

Due to heterogeneity in the type and use of data at various times by various organisations, there are a variety of metadata standards, national and international. Although it might be seen as preferable to have only one standard, there cannot be a single universal standard that can be adapted to all forms of data. The principal standards in use with relevance to geospatial datasets are UK GEMINI, ISO19115, Dublin Core and DCAT-2, outlined in Appendix One. DAFNI currently utilises the DCAT-2 standard for metadata. The current DAFNI 'DAFNI-LD, DCAT2 Profile Metadata' description is available online[17]. DAFNI also uses the GeoNames gazetteer[18] to locate named places. DAFNI metadata is captured across six thematic sections, Table 16.

Table 16 DAFNI Metadata Thematic Sections

DAFNI Metadata Thematic Sections
Dataset-level information describing the properties and characteristics of the overall dataset – its name, a description of its content, primary language, the theme and/or subjects it encapsulates, etc.
Provenance information describing the origins of the dataset, particularly useful when a dataset has been provided via a third-party source (e.g., TfL, ONS, Ordnance Survey, a government department, etc.)
Legal information pertaining to potential licensing arrangements, access and usage rights, etc.
Temporal information describing the time range the dataset encapsulates.
Spatial information describing the geospatial extent of the dataset.
File-level information detailing properties of each file hosted within the dataset – its name, format (derived from file extension), size, any international standard it conforms to, etc.

Note: from https://docs.secure.dafni.rl.ac.uk/docs/reference/dafni-ld-metadata

The DCAT-2[19] metadata standard adopted in DAFNI 'is an RDF vocabulary designed to facilitate interoperability between data catalogues published on the Web', providing a structured, formalised means to characterise the datasets held in DAFNI. DCAT is able to describe datasets, resources, services and catalogues. DCAT-2 further permits datasets to be classified with wider thematic classifications. For this, it uses the Simple Knowledge Organization System (SKOS)[20], a common data model for sharing and linking knowledge organization systems via the Web. Using SKOS, concepts can be 'identified using URIs, labelled with lexical strings in one or more natural languages, assigned notations (lexical codes), documented with various types of note, linked to other concepts and organized into informal hierarchies and association networks, aggregated into concept schemes, grouped into labelled and/or ordered collections, and mapped to concepts in other schemes'. DCAT-2 incorporates the amendments created by DCAT-AP, which had in turn extended the original version of DCAT.

[17] https://docs.secure.dafni.rl.ac.uk/docs/reference/dafni-ld-metadata/
[18] https://www.geonames.org
[19] https://www.w3.org/TR/vocab-dcat-2
[20] https://www.w3.org/TR/skos-reference

DAFNI also incorporates GeoJSON-LD 1.0[21]standards. JSON is a data serialization and messaging format this has been used within JSON-LD which is a format used to serialize linked data. The main purpose is to use linked data in web-0based programming environments and to build interoperable web services which is a key objective of the DAFNI platform. GeoJSON is also a basis for Next Generation Service Interfaces-Linked Data (NGSI-LD) a key IoT standard with applications relating to Digital Twins. NGSI-LD is described further in the Discussion section.

The metadata system adopted in DAFNI should where possible be aligned to the adopted metadata systems in other comparable and associated data catalogues, to help ensure cross-compatible searching and other 'FAIR' characteristics. UKRI Councils operate a range of data centres and a review of metadata standards adopted in these would be useful.

Adoption of the DCAT-2 MetaData standards and a comparison with UK GEMINI

An ambition for the DAFNI website is to present the selected DCAT-2 metadata fields alongside the equivalent metadata fields from the other metadata standards described in this report. This will allow a mapping between standards, enabling ease of data discovery and interoperability where potential users are have adopted other metadata approaches. DCAT-2 fields are described with example completions in Table 17.

Table 17 Metadata example – DCAT-2

DCAT-2 Fields	Example
Title	Test
Subject	Subject: Transportation
Version	Version: May 19th 2020
Creation date	Created: May 19th 2020
Creator	Creator: f3527716-9028-4373-ac8c-9287dc717f07
Contact	Contact
Description	Description
Unique Id	Identifier: b78990f9-af61-4507-9618-6b9410808713:60a0ad95-841b-486a-9bf7-6a41de88c415:0886aa04-2f20-43ab-b448-c3af142ba1b8
Location	London, UK
Start date	Start date: May 1st 2020
End date	End date: May 19th 2020
Theme(s)	Theme(s), Transport networks
Organisation	Cranfield University
Organisation identifier	http://www.cranfield.ac.uk
Publisher	Publisher

[21] https://www.w3.org/TR/2014/REC-json-ld-20140116/

Bounding box	
Issued	Issued, May 19th 2020
Update frequency	Irregular
Rights	Rights
Language	Language, en
Standard	Standard, N/A
Update frequency	Update Frequency, N/A
Keywords	Keywords

DCAT-2 mapping to UK GEMINI

Mappings between the members of differing metadata schemes allow greater interaction between search and discovery tools. Such mappings can be automated, for example the XSLT template developed by the EU Joint Research Centre to permit automatic transformations to convert ISO 19139 metadata to DCAT-AP[22]. A further mapping of DCAT-2 to SCHEMA-ORG v3.4 is available on the DCAT portal. The following table provides a proposed mapping between DCAT-2 fields and UK GEMINI, Table 18.

Table 18 Metadata mapping – DCAT-2 and UK GEMINI

DCAT-2 Fields[23]	UK GEMINI Equivalent Field[24]
dct:title	1. Title
<?>	2. Alternative title
dct:language	3. Dataset language
dct:description	4. Abstract
dcat:theme *and/or* dct:subject[1] *and/or* foaf:primaryTopic	5. Topic Category
dcat:keyword	6. Keyword
dct:temporal *(see also temporal properties)*	7. Temporal extent
dct:issued[2]	8. Dataset reference date
prov:wasGeneratedBy	10. Lineage
dct:location (dct:spatial)	15. Extent
<?>	16. Vertical extent information

[22] https://github.com/GeoCat/iso-19139-to-dcat-ap/blob/master/documentation/Mappings.md

[23] https://www.w3.org/TR/vocab-dcat-2

[24] https://www.agi.org.uk/gemini/40-gemini/1250-element-summary and https://www.agi.org.uk/agi-groups/standards-committee/uk-gemini/40-gemini/1062-gemini-datasets-and-data-series

dct:conformsTo	17. Spatial reference system
dcat:spatialResolutionInMetres	18. Spatial resolution
dcat:accessURL	19. Resource locator
dct:type and/or dcat:packageFormat	21. Data format
dct:publisher	23. Responsible organisation
dct:accessRights	25. Limitations on public access
dct:licence	26. Use constraints
<?>	27. Additional information
dct:issued *and* dct:modified	30. Metadata date*
dct:language	33. Metadata language*
dcat:contactPoint	35. Metadata point of contact*
dct:identifier	36. Resource identifier
dcat:endpointDescription	37. Spatial data service type
dct:isReferencedBy	38. Coupled resource
dct:type and/or dct:format	39. Resource type*
dct:conformsTo	41. Conformity
<?>	43. Equivalent scale
dct:location	44. Bounding box
dct:identifier	45. File Identifier*
<?>	47. Hierarchy level name*
dct:conformsTo	48. Quality scope
dct:isReferencedBy and/or dct:qualifiedRelation and/or dct:relation	49. Parent identifier*
dct:spatial and/or locn:geometry	50. Spatial representation type
<?>	51. Character encoding
dqv:QualityAnnotation (see also dqv:QualityPolicy and dqv:QualityMeansurement)	52. Data quality
<?>	53. Maintenance information
dct:conformsTo	54. Metadata standard name
dct:conformsTo	55. Metadata standard version

Notes:

* indicates metadata on metadata

UKGEMINI Element numbers 9, 11, 12, 13, 14, 20, 22, 28, 29, 31, 32, 40 and 42 have been omitted because they were used to identify elements that have now been deleted from the Standard and have not been reallocated to avoid confusion.

The DCAT-2 normative namespace prefixes are used here, thus: dcat = DCAT-2; dct = Dublin Core Terms. <?> = no obvious correlation

[1] dcat:theme provides the link to any SKOS Concept Schemas; dcat:theme is a sub property of dcat:subject

[2] Note also the potential to use dct:created and dct:issued

Discussion

The approaches described in this report highlight both methods adopted in DAFNI, and a review of wider similar standards. However, there are a range of challenges remaining, key to which include:

How the current methods for discoverability and findability will deal with future datasets, such as real-time dynamic datasets.

> **Suggestion:** With the growing interest in 'Digital Twins' and 'IoT' data stream inclusion, a further review should be conducted to affirm that DCAT-2 remains appropriate. DCAT-2 references datasets and services, so it is anticipated this will be the case. As described by Jacoby et al. (2020) the Next Generation Service Interfaces-Linked Data (NGSI-LD) API is an IoT standard defined by the ETSI Industry Specification Group and relates very closely with the context of Digital Twins.

Much of the data in DAFNI appears to currently comprise CSV files (Comma Separated Value Files). However, in addition to data sources such as plain text documents, there are a wide range of binary and ASCII atomic (single) and complex (compound) files such as Shapefiles, Geopackages, and geodatabases, as well as linked data repositories such as PostGIS, or existing API's web services (WFS etc.). These should all be cross-referenced.

> **Suggestion:** A review of handling complex / compound files should be undertaken such that, for example a Shapefile is presented as a single entity comprising multiple files.
>
> Ideally DAFNI should further seek to permit federated cross-archive searches to enable optimum 'findability'.

How should large complex dynamic datasets be published? These are often gathered from automated sensors supplying continuous data feeds. It may be impossible to make entire datasets available to users, so it is necessary to adequately describe data structure and allow for the extraction of subsets based on spatial and temporal requirements.

> **Suggestion:** Further consideration of managing IoT data sources should be undertaken and further Use Cases be developed to extend and complement those in this report.

Commonly hypercubes are adopted to store large complex datasets with multi-dimensions. Currently the RDF Data cube ontologies support the use of sensor data but do not include the spatial/temporal aspects of the data. The extension of this ontology to include spatial/temporal components is an item on the W3C working group[25].

> **Suggestion:** A review of managing complex linked data configurations is needed. This will result in a selection of data 'Ontologies' being promoted for use.

[25] https://www.w3.org/TR/qb4st/

Conclusions

A motivation for the work in DAFNI is in supporting the growing focus on the 'green recovery', and the need to encourage and foster environmentally sustainable clean economic growth. The national infrastructure plan recognises a range of important themes in planning future capacities, and in regions such as, for example, the Oxford Cambridge Arc, the 'Spatial Framework' lays out a vision and sense of place the combines these. A key development from traditional modelling tools now is the adoption of the Digital Twinning (DT) approach. The DT requires collation (sometimes referred to as 'clashing and layering') of a multitude of active linked data sources. Ontological models are designed to aid and normalise this approach, and to permit interoperability of potential sources.

1. Recognising the spectrum of ontologies, from general scale to localised and specific, we need to 'map' application-level ontologies to the specific tasks at hand. Applications will both vary widely or may use similar (linked) data but in nuanced ways – a keyword is interoperability of sources.
2. To address this and to help users navigate their way through DAFNI, we highlight the strong role for worked Use Cases. This is particularly the case in supporting Digital Twinning applications where different '*forms*' of data (e.g., high throughput streamed data) are involved along with the '*variety*' of data themes. The ontologies are both (a) to aid sensemaking for users and increase re-use, but importantly (b) to aid automated data inclusion and data (source) swapping.
3. Key to the discovery of data resources in DAFNI will be the metadata used to characterise and describe it. The 'F' in 'FAIR' data principles[26] relates to the findability of the data, this is addressed by publishing metadata. DCAT-2 is selected as the contemporary standards-based form for this[27]. DCAT-2 as embedded in DAFNI, incorporates much of the descriptive form of the 'Dublin Core' metadata standard. Mappings can usefully be made to aid comparison to other standards, for example UK GEMINI and ISO19115 as well as other forms.
4. Key to the classification of the bodies of data now held within DAFNI are the data taxonomies, accessed by the user through the interface. There are a range of taxonomies and restricted vocabularies which can be adopted and a review of these is being conducted, ensuring that the selections feed into the ontological models. There is an issue of scale and aggregation to consider here as data themes can be clustered to represent a greater whole.

[26] https://www.go-fair.org/fair-principles/
[27] https://www.w3.org/TR/vocab-dcat-2/

Recommendations

Further to the points raised and discussed in this document, and with reference to the FAIR data principles, we make the following recommendations.

1. **Discovery and Metadata**: Continue to use DCAT-2. Seek to align metadata standards with the discoverability protocols of other data centres – for example in NERC CEDA[28], in ESRC the UK Data Service[29] and in STFC the UK EnergyData Centre[30].
2. **Taxonomies**: Consider adopting the EUSciVoc controlled vocabulary alongside the INSPIRE themes and topic categories to classify data in supporting searches. The existing INSPIRE approach is quite acceptable – but may not have the granularity required. EUSciVoc has a hierarchy that may prove easier to use as well as a broader scope of topics drawn across CORDIS.

[28] https://www.ceda.ac.uk
[29] https://esrc.ukri.org/research/our-research/uk-data-service/
[30] https://ukerc.rl.ac.uk

References

Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE), http://data.europa.eu/eli/dir/2007/2/oj

FGDC (2007) Geospatial Metadata Standards, http://www.fgdc.gov/metadata/geospatial-metadata-standards

Guarino, N., Oberle, D., enStaab, S. (2009) What is an Ontology. Accessed on 4/12/20 at https://iaoa.org/isc2012/docs/Guarino2009_What_is_an_Ontology.pdf

Hall, J. (2019) UK reveals new platform for infrastructure data analysis and simulation modelling. Proceedings of the Institution of Civil Engineers - Civil Engineering 2019, 172:3, 102-102; doi: 10.1680/jcien.2019.172.3.102.

Hillman, D. (2005) Using Dublin Core, http://dublincore.org/documents/usageguide

Jacoby, M., Usländer, T. (2020) Digital Twin and Internet of Things—Current Standards Landscape. Appl. Sci. 10, no. 18: 6519. https://doi.org/10.3390/app10186519

NISO (2004), Understanding Metadata, National information Standards Organization Press Bethesda, Maryland

OGC (2012) OGC GeoSPARQL - A Geographic Query Language for RDF Data, http://www.opengis.net/doc/IS/geosparql/1.0

Ordnance Survey (2020) Ordnance Survey Ontologies http://data.ordnancesurvey.co.uk/ontology

OWL 2 Web Ontology Language Document Overview (Second Edition) W3C Recommendation 11 December 2012. https://www.w3.org/TR/owl2-overview/

Ordnance Survey Ontologies. http://data.ordnancesurvey.co.uk/ontology

Osgeo Wiki (2007), DCLite4G, http://wiki.osgeo.org/wiki/DCLite4G

Studer, R., Benjamins, R., Fensel, D. (1998) Knowledge engineering: Principles and methods. Data & Knowledge Engineering, 25(1-2):161-198, 1998.

UK Gemini (2018) Specification for discovery metadata for geospatial data resources. AGI. V2.3. Online at https://www.agi.org.uk/gemini/40-gemini/1250-element-summary

UK GEMINI. Metadata Guidelines Part 3 April 2015. https://www.agi.org.uk/agi-groups/standards-committee/uk-gemini

W3C (2017) Spatial Data on the Web Best Practices, W3C Working Group Note 28 September 2017. https://www.w3.org/TR/sdw-bp/#intro

Appendix One – Metadata Standards

Appendix One lists and describes the principal metadata standards that have been used in recent years.

MARC Standard

The Machine-readable cataloguing (MARC) standard[31] was developed and managed by US library of congress and it is designed to store information about bibliographic materials. It stores information in machine (computer)-readable form using alpha-numeric field identifier. MARC has a large number of fields, and it can be easily extended to accommodate more fields. Although, it has fields that can store geospatial information, it is not well adapted to storing spatial information, and the large number of fields and its characteristic no-domain restriction make it less cost effective.

Dublin Core

The Dublin Core (DC) standard[32] for metadata is developed and managed by DCMI (an international open forum with participants from Academic institutions, library groups etc.). In its simple initial form DC has 15 free text elements with no mandatory fields, falling across three categories 'content', 'intellectual property' and 'instantiation'[33], as noted below in Table 19. It was designed for cross-disciplinary resource discovery. Although it has subsequently had added dozens of additional fields for increasing precision of referencing, it has limitations in handling geographic aspects of data.

Dublin Core is an open forum standard, managed by Dublin Core Metadata Initiative (DCMI), with the aim of "promoting the widespread adoption of interoperable metadata standards and developing specialised metadata vocabularies for describing resources that enable more intelligent information discovery systems"[34].
Unlike the other standards mentioned here, Dublin Core itself is not specifically designed to encompass and document geospatial data. Dublin Core was developed to be simple and concise, and to describe Web-based documents (NISO, 2004).

In contrast to the FGDC and ISO standards, Dublin Core has no specification of 'mandatory' fields. All entries are purely optional. The standard is split into two levels: *Simple* and *Qualified*. The fields are as follows, Table 19.

Table 19 Dublin Core Fields

Dublin Core Simple Fields
Contributor (intellectual property) Definition: An entity responsible for making contributions to the resource.

[31] http://www.loc.gov/marc/
[32] http://dublincore.org/
[33] https://www.ietf.org/rfc/rfc2413.txt
[34] http://dublincore.org/about/

Comment: Examples of a Contributor include a person, an organization, or a service. Typically, the name of a Contributor should be used to indicate the entity.
Coverage (Content) Definition: The spatial or temporal topic of the resource, the spatial applicability of the resource, or the jurisdiction under which the resource is relevant. Comment: Spatial topic may be a named place or a location specified by its geographic coordinates. Temporal period may be a named period, date, or date range. A jurisdiction may be a named administrative entity or a geographic place to which the resource applies. Recommended best practice is to use a controlled vocabulary such as the Thesaurus of Geographic Names. Where appropriate, named places or time periods can be used in preference to numeric identifiers such as sets of coordinates or date ranges.
Creator (intellectual property) Definition: An entity primarily responsible for making the resource. Comment: Examples of a Creator include a person, an organization, or a service. Typically, the name of a Creator should be used to indicate the entity.
Date (Instantiation) Definition: A point or period of time associated with an event in the lifecycle of the resource. Comment: Date may be used to express temporal information at any level of granularity. Recommended best practice is to use an encoding scheme.
Description (Content) Definition: An account of the resource. Comment: Description may include but is not limited to: an abstract, a table of contents, a graphical representation, or a free-text account of the resource.
Format (Instantiation) Definition: The file format, physical medium, or dimensions of the resource. Comment: Examples of dimensions include size and duration. Recommended best practice is to use a controlled vocabulary.
Identifier (Instantiation) Definition: An unambiguous reference to the resource within a given context. Comment: Recommended best practice is to identify the resource by means of a string conforming to a formal identification system.
Language (Instantiation) Definition: A language of the resource. Comment: Recommended best practice is to use a controlled vocabulary.
Publisher (intellectual property) Definition: An entity responsible for making the resource available. Comment: Examples of a Publisher include a person, an organization, or a service. Typically, the name of a Publisher should be used to indicate the entity.
Relation (Content) Definition: A related resource. Comment: Recommended best practice is to identify the related resource by means of a string conforming to a formal identification system.
Rights (intellectual property) Definition: Information about rights held in and over the resource. Comment: Typically, rights information includes a statement about various property rights associated with the resource, including intellectual property rights.
Source (Content) Definition: The resource from which the described resource is derived.

Comment: The described resource may be derived from the related resource in whole or in part. Recommended best practice is to identify the related resource by means of a string conforming to a formal identification system.
Subject (Content) Definition: The topic of the resource. Comment: Typically, the topic will be represented using keywords, key phrases, or classification codes. Recommended best practice is to use a controlled vocabulary. To describe the spatial or temporal topic of the resource, use the Coverage element.
Title (Content) Definition: A name given to the resource. Comment: Typically, a Title will be a name by which the resource is formally known.

Note: Fields in alphabetical order

Dublin Core Qualified Fields
Audience Definition: a class entity for whom the resource is intended or useful
Provenance Definition: a statement of any changes in ownership and custody of the resource since its creation that are significant for its authenticity, integrity, and interpretation
Rights Holder Definition: a person or organisation owning or managing rights over the resource.

Subsequent to the development of the original set of properties, 'ISO 15836-2, Information and documentation – The Dublin Core metadata element set – Part 2: DCMI Properties and classes, extends … these ... with 40 properties and 20 classes in order to improve the precision and expressiveness of descriptions'[35].

DCLite4G

DCLite4G[36], which stands for "Dublin Core Lightweight Profile for Geospatial" is a proposed extension of Dublin Core that will handle geospatial metadata. "It defines minimal information model for metadata exchange, (Go-Geo, 2006). It has all the elements of Dublin core, together with some elements for description of spatial data and namespace used to define extra properties needed to usefully specify the properties of geospatial data (Osgeo wiki, 2007).

FGDC CSDGM

The Content Standard for Digital Geospatial Metadata (CSDGM)[37] is the Metadata standard developed and maintained by the FGDC (Federal Geographic Data Committee) for the US spatial data user community. The elements are organised hierarchically. It was originally published in 1994 and revised in 1998 to produce its second version. According to FGDC (2007), "A key feature of the CSDGM Version 2 is the ability of geospatial data communities to customize the base CSDGM. Extensions

[35] https://www.iso.org/news/ref2474.html
[36] http://wiki.osgeo.org/index.php/DCLite4G
[37] http://www.fgdc.gov/metadata/geospatial-metadata-standards

are a set of added elements that extend the standard to better serve the community or data type." Although its use has been superseded by the ISO standard, there have been efforts to improve it to meet the ISO specifications. (FGDC, 2007).

The US Federal Geographic Data Committee (FGDC) is the custodian of the Content Standard for Digital Geospatial Data (CSDGM) with the aim to support the collection and processing of geospatial metadata.

The standard was formally approved on the 8th June 1994 since which time it has evolved and expanded with various developments and additions. The current version was adopted in June 1998 with improvements and modifications made to Section 4 (Spatial Reference), edits to the glossary and the valid format of domains to include free text as an acceptable value.

CSDGM was originally developed from the perspective of defining the information required by a prospective user to determine the availability of a set of spatial data, as well as to determine the suitability of it for an intended use. The standard is organised hierarchically into three areas namely Sections, Compound Elements and Data Elements.

Sections are classified as the main "chapters" of the standard, represented in Figure 1. The figure shows how the three supporting data elements link into the seven main compound elements that fall under the Metadata section.

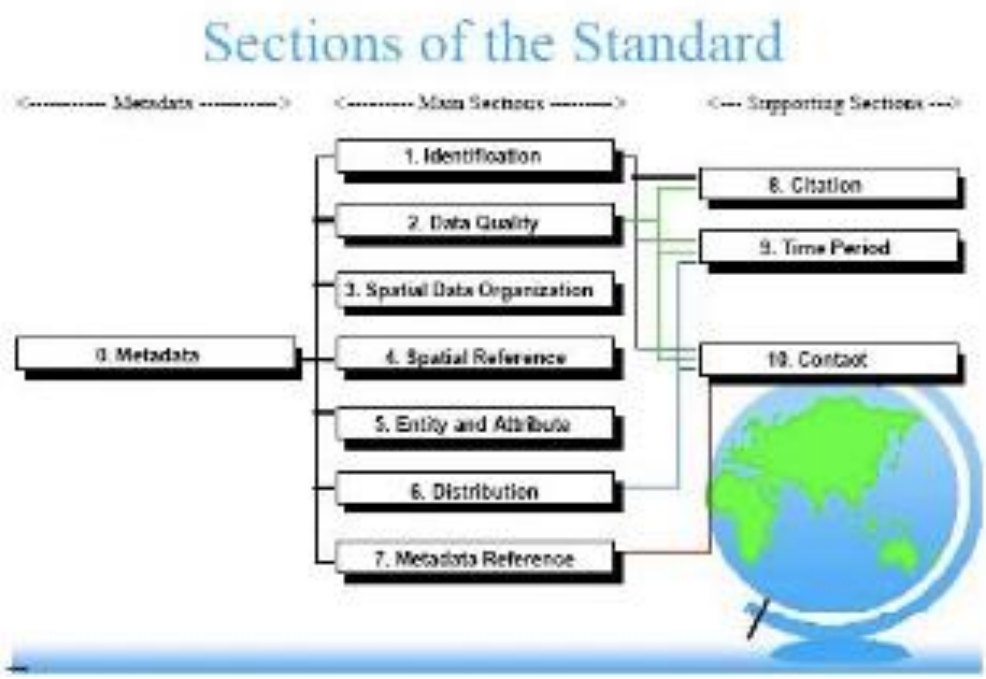

Figure 1 Structure of CSDGM

Source: FGDC website http://www.fgdc.gov/metadata/

The only two sections in CSDGM deemed to be mandatory as a whole are Section 1 (identification information) and section seven (metadata reference information). The support sections do not stand alone but contain information that is referenced more than once in the seven main sections.

e-GMS

"The UK e-Government Metadata Standard (e-GMS)[38] lays down the elements, refinements and encoding schemes to be used by government officers when creating metadata for their information resources or designing search interfaces for information systems. This forms part of the e-Government Interoperability Framework (eGIF)" (Cabinet office, 2003). The e-GMS standard is based upon Dublin core and so, it represents geospatial data poorly, but it does ensure uniformity of metadata across the UK's public sector. Note this standard is surpassed by the UK Gemini standard outlined below.

ISO19115 and ISO19139

These are the two key international standards developed by the ISO for geospatial metadata. ISO19115[39] was published in 2003 and has more than 300 elements (most of which are optional) for describing all aspects of geospatial data. "It is widely used as the basis for geospatial metadata services. However, because of the large number of metadata elements and the complexity of its data model, it is difficult to implement" (B Oates & J Rapaport, 2007). It should be noted that IS019000 series specifies the standards for geographic data representation, while the XML implementation of the ISO19115 is ISO19139, published in 2007. Due to the fact that ISO19115 standard does not sufficiently enough, represent remotely sensed imagery and gridded data metadata and, ISO19115-2 (Geographic information – Metadata – Part 2: Extensions for imagery and gridded data) is being developed and is due to be released soon.

In May 2003, as part of the ISO Technical Committee 211 (ISO/TC 211, 2004), ISO released the metadata standard ISO19115 as a comprehensive profile of information for spatial data, Table 20. ISO19115 is part of a larger suite of Geospatial standards and the standard defines a comprehensive set of more than 300 metadata elements which are grouped in 86 classes of 282 attributes and 56 relations, most of which can be applied optionally. At the uppermost level the classes are grouped in 14 packages.

Table 20 Metadata standards developed by ISO

Standard Name	Description
ISO19115:2003 – Metadata	Defines the structure required for describing geographic information and services. It provides information about the identification, extent,

[38] http://www.esd.org.uk/standards/egms/
[39] https://www.iso.org/standard/53798.html

	quality, spatial and temporal schema, spatial reference and distribution of digital spatial data
ISO/TS19139:2007 – Metadata XML schema implementation	Defines structure of the XML schema implementation derived from ISO19115

The scope of the standard is broad. The various entities within the standard can be described as variable and flexible, although mandatory information items are present within the standard. The list below describes the various information relating to the specific metadata tabs within ISO19115, highlighting which are mandatory (M) and which are optional (O), Table 21 and Table 22.

Table 21 ISO19115 Information Categories

ISO19115 Information Categories
Metadata entity information (M): is the top-level package and contains general information on the metadata, e.g., date, contact, language, etc.
Identification information (M): contains information to uniquely identify the data. Information required includes purpose, credit, resource, status, point of contact.
Constraint information (O): contains information concerning the restrictions placed on data (distribution, usage, etc).
Data quality information (O): contains a general assessment of the quality of the dataset.
Maintenance information (O): contains information about the scope and frequency of updating data.
Spatial representation information (O): contains information concerning the mechanisms used to represent spatial information in a dataset.
Reference system information (O): contains the description of the spatial and temporal reference system used in a dataset.
Content information (O): contains information identifying the feature catalogue used and/or information describing the content of a coverage dataset.
Portrayal catalogue information (O): contains information identifying the used portrayal catalogue (style) used for a given dataset.
Distribution information (O): contains information about the distributor of the dataset, and options for obtaining it.
Metadata extension information (O): contains information about used specified extensions, e.g., a reference to extended elements to the standard which may be defined by a metadata producer or a user community.
Application schema information (O): contains information about the application schema used to build the dataset.
Extent information (O): contains information that describes the spatial and temporal extent of the data.
Citation and responsible party information (O): contains information needed for citing a resource (dataset, feature, source, publication, etc) as well as information about the party responsible for the resource.

Note: O=optional, M=mandatory; C=conditional

Table 22 ISO19115 Fields

ISO19115 Fields
Dataset title (M)
Spatial representation type (O)
Dataset reference date (M)
Reference system (O)
Dataset responsible party (O)
Lineage (O)
Geographic location of the dataset (by four coordinates or by geographic identifier) (C)
On-line resource (O)
Dataset language (M)
Metadata file identifier (O)
Dataset character set (C)
Metadata standard name (O)
Dataset topic category (M)
Metadata standard version (O)
Spatial resolution of the dataset (O)
Metadata language (C)
Abstract describing the dataset (M)
Metadata character set (C)
Distribution format (O)
Metadata point of contact (M)
Additional extent information for the dataset (vertical and temporal) (O)
Metadata date stamp (M)

Note: O=optional, M=mandatory; C=conditional

The Open Geospatial Consortium (OGC) aims at providing a comprehensive suite of open interface applications to enable transparent access to heterogeneous spatial information in a network environment based on XML.

Recent advancements have seen the ISO19139 standard proposed and formalised which defines the structure of the XML schema derived from the ISO19115 content standard.

ISO 19139 is designed to provide a common XML specification for describing, validating, and exchanging geographic metadata. It is intended to promote interoperability and exploit ISO19115's advantages in a concrete implementation specification.

In the first half of 2007 the XML schema for the implementation of metadata based on the ISO19115 specification was released and provides a single UML interpretation of the metadata structure. This specification is designed to enhance interoperability by providing a common specification for describing, validating and exchanging metadata.

UK GEMINI

The UK GEMINI (GEo-spatial Metadata INteroperability Initiative) Discovery Metadata Standard[40] is a metadata standard derived by combining e-GMS and ISO19115 in the UK. Its element set can be used for describing geo-spatial, discovery level metadata. It was produced by the Association for Geographic Information (AGI), the e-Government Unit of the Cabinet Office and the UK Data Archive (Cabinet Office, 2003), it has all the core elements of ISO 19115 and e-GMS and replaced the former 'NGDF' standard.

The UK GEMINI Discovery Metadata Initiative (GEMINI) was launched in October 2004 as the result of collaboration between the Association for Geographic Information (AGI), the UK Cabinet Office e-Government Unit and the UK Data Archive at the University of Essex (UK). The latest revision is version 2.3, published in 2018, having full compliance with ISO 19115 (AGI, 2009).

The scope of the project was to identify a set of elements derived from the ISO19115 standard, to be used as a discovery level metadata system within the UK. A discovery level metadata system aims to provide a basic overview and existence of the resource whilst providing a steppingstone to further and more detailed data.

The standard is managed by a steering group representing central and local government, the private sector, academia and the geo-spatial industry at large. The standard was created from a rigorous process of national consultation, feedback and revision by all the involved stakeholders.

GEMINI proposes a profile (subset of information derived from other published standards, in this case the ISO19115 and the UK national eGMS standards) and supersedes the earlier GI-gateway Metadata Specification, previously known as the National Geospatial Data Framework (NGDF) metadata standard, which had until then been the prevailing national metadata standard within the UK.

GEMINI v2.3 comprises a set of 32 metadata elements listed below. Each element is a subject of a separate table in the GEMINI metadata structure, Table 23.

Table 23 UK GEMINI Version 2.3 Schema

UK GEMINI id	Element name	Obligation	Number of occurrences	Dataset and series	Service	Revision date
1	Title	Mandatory	Single	✓	✓	March 2019
2	Alternative title	Optional	Multiple	✓	✓	April 2020
3	Dataset language	Mandatory	Multiple	✓	n/a	September 2018
4	Abstract	Mandatory	Single	✓	✓	March 2019
5	Topic Category	Mandatory when the data resource is a	Multiple	✓	n/a	March 2019

[40] https://www.agi.org.uk/agi-groups/standards-committee/uk-gemini

UK GEMINI id	Element name	Obligation	Number of occurrences	Dataset and series	Service	Revision date
		dataset or dataset series. Not applicable to services.				
6	Keyword	Mandatory	Multiple	✓	✓	May 2019
7	Temporal extent	Mandatory for datasets and dataset series	Multiple	✓	✓	May 2018
8	Dataset reference date	Mandatory	Multiple	✓	✓	March 2019
10	Lineage	Mandatory	Single	✓	n/a	March 2019
15	Extent	Optional	Multiple	✓	✓	August 2010
16	Vertical extent information	Optional	Multiple	✓	✓	March 2019
17	Spatial reference system	At least one coordinate reference system shall be given	Multiple	✓	✓	March 2019
18	Spatial resolution	Conditional - where a resolution distance can be specified	Multiple	✓	✓	March 2019
19	Resource locator	Conditional - Must be supplied when online access is available	Multiple	✓	✓	July 2009
21	Data format	Mandatory	Multiple	✓	n/a	March 2019
23	Responsible organisation	Mandatory	Multiple	✓	✓	July 2009
25	Limitations on public access	Mandatory	Multiple	✓	✓	January 2020
26	Use constraints	Mandatory	Multiple	✓	✓	September 2018
27	Additional information	Optional	Single	✓	n/a	October 2018
30	Metadata date*	Mandatory	Single	✓	✓	May 2018
33	Metadata language*	Mandatory	Single	✓	✓	September 2018
35	Metadata point of contact*	Mandatory	Multiple	✓	✓	July 2009
36	Resource identifier	Mandatory for datasets and dataset series	Multiple	✓	n/a	May 2018
37	Spatial data service type	Mandatory	Single	n/a	✓	July 2009
38	Coupled resource	Conditional - mandatory for View and Download services,	Multiple	n/a	✓	March 2019

UK GEMINI id	Element name	Obligation	Number of occurrences	Dataset and series	Service	Revision date
		optional for other service types.				
39	Resource type*	Mandatory	Single	✓	✓	July 2009
41	Conformity	Mandatory	Multiple	✓	✓	April 2020
43	Equivalent scale	Optional	Multiple	✓	✓	July 2009
44	Bounding box	Mandatory	Multiple	✓	✓	December 2012
45	File Identifier*	Mandatory	Single	✓	✓	March 2019
47	Hierarchy level name*	Conditional, required when Resource type (ISO hierarchyLevel) is not "dataset".	Single	✓	✓	April 2020
48	Quality scope	Mandatory	Multiple	✓	✓	March 2019
49	Parent identifier*	Optional	Single	✓	✓	May 2018
50	Spatial representation type	Mandatory	Multiple	✓	n/a	May 2018
51	Character encoding	Conditional (mandatory if an encoding is used that is not based on UTF-8, otherwise optional)	Multiple	✓	n/a	May 2018
52	Data quality	Conditional: Topological consistency report is mandatory if the dataset includes types from the INSPIRE Generic Network Model and does not assure centerline topology (connectivity of centrelines) for the network, otherwise optional	Multiple	✓	n/a	May 2018
53	Maintenance information	Optional	Single	✓	n/a	April 2020

UK GEMINI id	Element name	Obligation	Number of occurrences	Dataset and series	Service	Revision date
54	Metadata standard name*	Optional	Single	✓	✓	January 2020
55	Metadata standard version*	Optional	Single	✓	✓	January 2020

** indicates metadata on metadata*

Note. Element numbers 9, 11, 12, 13, 14, 20, 22, 28, 29, 31, 32, 40 and 42 have been omitted because they were used to identify elements that have now been deleted from the standard and have not been reallocated to avoid confusion.

Source: https://www.agi.org.uk/40-gemini/1250-element-summary

Several of the metadata elements specified in UKGEMINI use enumerated code lists, e.g., pre-defined sets of values identified by codes which are useful to standardise the entries to aid searches of metadata for specific values. The code lists are taken from the ISO19115 specifications, although in some cases the code has been modified to make it more appropriate to the UK context.

Data Catalogue Vocabulary (DCAT) - Version 2

Sharing data resources among different organizations, researchers, governments and citizens requires the provision of metadata. This is irrespective of the data being open or not. DCAT is a vocabulary for publishing data catalogues on the Web, which was originally developed in the context of government data catalogues such as data.gov and data.gov.uk, but it is also applicable and has been used in other contexts.

The revision of DCAT to DCAT-2[41] has extended the previous version to support further use cases and requirements [DCAT-UCR; DCAT-AP etc.]. These include the possibility of cataloguing other resources in addition to datasets, such as data services. The revision also supports describing relationships between datasets as well as between datasets and other catalogued resources. Guidance on how to document licenses and rights statements associated with the catalogued items is provided.

DCAT provides RDF classes and properties to allow datasets and data services to be described and included in a catalogue. The use of a standard model and vocabulary facilitates the consumption and aggregation of metadata from multiple catalogues, which can:

- increase the discoverability of datasets and data services
- allow federated search for datasets across catalogues in multiple sites

Data described in a catalogue can come in many formats, ranging from spreadsheets, through XML and RDF to various specialized formats. DCAT does not make any assumptions about these serialization formats of the datasets but it does distinguish between the abstract dataset and its different manifestations or distributions.

[41] https://www.w3.org/TR/vocab-dcat-2/

Data is often provided through a service which supports selection of an extract, sub-set, or combination of existing data, or of new data generated by some data processing function. DCAT allows the description of a data access service to be included in a catalogue.

Complementary vocabularies can be used together with DCAT to provide more detailed format-specific information. For example, properties from the VoID vocabulary [VOID] can be used within DCAT to express various statistics about a dataset if that dataset is in RDF format.

DCAT is an RDF vocabulary for representing data catalogues. DCAT is based around six main classes, Table 24 and Figure 2.

Table 24 DCAT-2 Classes

DCAT-2 Classes
dcat:catalog represents a catalogue, which is a dataset in which each individual item is a metadata record describing some resource; the scope of dcat:catalog is collections of metadata about datasets or data services.
dcat:resource represents a dataset, a data service or any other resource that may be described by a metadata record in a catalogue. This class is not intended to be used directly, but is the parent class of dcat:dataset, dcat:dataService and dcat:catalog. Member items in a catalogue should be members of one of the sub-classes, or of a sub-class of these, or of a sub-class of dcat:resource defined in a DCAT profile or other DCAT application. dcat:resource is effectively an extension point for defining a catalogue of any kind of resource. dcat:dataset and dcat:dataService can be used for datasets and services which are not documented in any catalog.
dcat:dataset represents a dataset. A dataset is a collection of data, published or curated by a single agent. Data comes in many forms including numbers, words, pixels, imagery, sound and other multi-media, and potentially other types, any of which might be collected into a dataset.
dcat:distribution represents an accessible form of a dataset such as a downloadable file.
dcat:dataService represents a data service. A data service is a collection of operations accessible through an interface (API) that provide access to one or more datasets or data processing functions.
dcat:catalogRecord represents a metadata item in the catalogue, primarily concerning the registration information, such as who added the item and when.

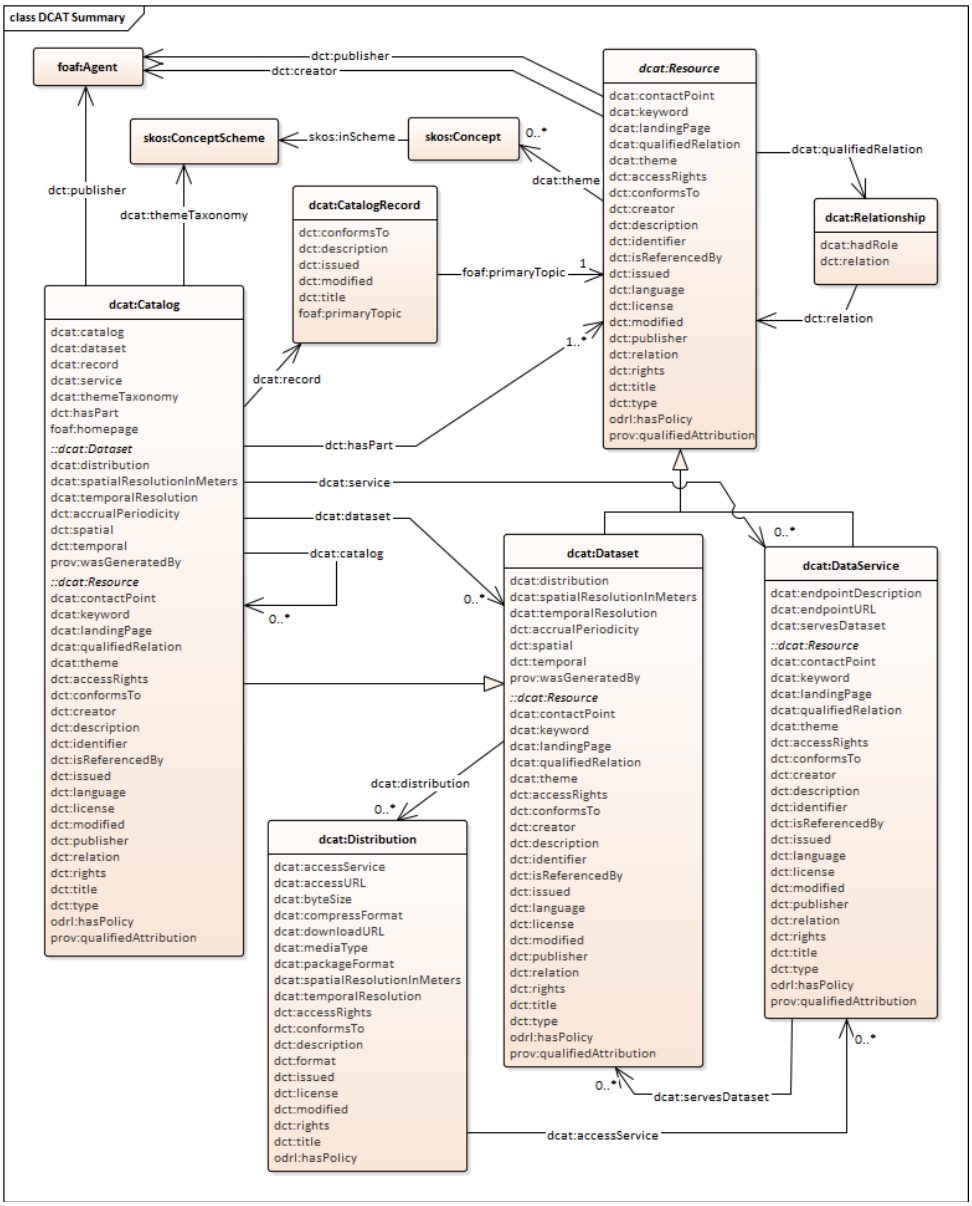

Figure 2 DCAT-2 Classes

The specific DCAT-2 profiler schema in JSON as adopted in DAFNI is presented as follows, Figure 3. This is a snapshot captured on the 14th December 2020 and is subject to change, being work in progress.

```
{
"@context": {
        "@version": 1.1,
        "@vocab": "http://www.w3.org/ns/dcat#",
        "adms": "http://www.w3.org/ns/adms#",
        "dc": "http//purl.org/dc/elements/1.1/",
        "dcat": "http://www.w3.org/ns/dcat#",
        "dct": "http://purl.org/dc/terms/",
        "owl": "http://www.w3.org/2002/07/owl#",
        "rdf": "http://www.w3.org/1999/02/22-rdf-syntax-ns#",
        "dctype": "http://purl.org/dc/dcmitype/",
        "xsd": "http://www.w3.org/2001/XMLSchema#",
```

```
"skos": "http://www.w3.org/2004/02/skos/core#",
"rdfs": "http://www.w3.org/2000/01/rdf-schema#",
"vcard": "http://www.w3.org/2006/vcard/ns#",
"prov": "http://www.w3.org/ns/prov#",
"foaf": "http://xmlns.com/foaf/0.1/",
"spdx" : "http://spdx.org/rdf/terms#",
"csvw": "http://www.w3.org/ns/csvw#",
"locn": "http://www.w3.org/ns/locn#",
"odrl": "http://www.w3.org/ns/odrl/2/",
"time": "http://www.w3.org/2006/time#",
"ssn": "http://www.w3.org/ns/ssn/",
"sosa": "http://www.w3.org/ns/sosa/",
"gj": "https://purl.org/geojson/vocab#",
"title": "dct:title",
"identifier": "dct:identifier",
"description":{
    "@id":"http://purl.org/dc/terms/description",
    "@type":"http://www.w3.org/2001/XMLSchema#string"
},
"distribution":{
    "@id":"http://www.w3.org/ns/dcat#distribution",
    "@type":"http://www.w3.org/ns/dcat#Distribution"
},
"dct:subject": {
    "@id":"http://purl.org/dc/terms/subject",
    "@type":"http://www.w3.org/2001/XMLSchema#string"
},
"mediaType" :"dcat:mediaType",
"downloadURL": {
  "@id": "dcat:downloadURL",
  "@type": "@id"
},
"accessURL": {
  "@id": "dcat:accessURL",
  "@type": "@id"
},
"fileName":{
  "@id":"spdx:fileName",
  "@type":"spdx:fileName"
},
"byteSize":{
  "@id":"dcat:byteSize",
  "@type":"xsd:decimal"
},
"conformsTo": {
              "@id": "dct:conformsTo",
              "@type": "Standard",
              "@context": {
                    "label": "rdfs:label"
            }
},
"language": {
                "@id": "dct:language",
                "@type": "@id"
},
"license": {
            "@id": "dct:license",
            "@context": {
                    "label": "rdfs:label"
            }
},
"rights": {
    "@id":"http://purl.org/dc/terms/rights",
```

```
    "@type":"http://purl.org/dc/terms/RightsStatement"
},
"internalID":{
                "@id": "adms:identifier",
                "@type": "@id"
},
"created": {
  "@id": "dct:created"
    },
"published": {
  "@id": "dct:issued"
},
"modified": {
  "@id": "dct:modified"
},
"accrualPeriodicity":{
    "@id":"http://purl.org/dc/terms/accrualPeriodicity",
    "@type":"http://purl.org/dc/terms/Frequency"
},
"temporal": {
    "@id":"http://purl.org/dc/terms/temporal",
    "@type":"http://purl.org/dc/terms/PeriodOfTime"
},
"creator": {
    "@id":"http://purl.org/dc/terms/creator",
    "@type":"http://xmlns.com/foaf/0.1/Agent"
},
"publisher": {
    "@id":"http://purl.org/dc/terms/publisher",
    "@type":"http://xmlns.com/foaf/0.1/Agent"
},
"keyword": {
    "@id":"http://www.w3.org/ns/dcat#keyword",
    "@type":"http://www.w3.org/2000/01/rdf-schema#Literal"
},
"landingPage":{
    "@id":"http://www.w3.org/ns/dcat#landingPage",
    "@type":"http://xmlns.com/foaf/0.1/Document"
},
"spatial": {
    "@id":"dct:spatial",
    "@type":"http://purl.org/dc/terms/Location"
},        "Feature": "gj:Feature",
"FeatureCollection": "gj:FeatureCollection",
"GeometryCollection": "gj:GeometryCollection",
"LineString": "gj:LineString",
"MultiLineString": "gj:MultiLineString",
"MultiPoint": "gj:MultiPoint",
"MultiPolygon": "gj:MultiPolygon",
"Point": "gj:Point",
"Polygon": "gj:Polygon",
"bbox": {
  "@container": "@list",
  "@id": "gj:bbox"
},
"coordinates": {
  "@container": "@list",
  "@id": "gj:coordinates"
},
"features": {
  "@container": "@set",
  "@id": "gj:features"
},
```

```
        "geometry": "gj:geometry",
        "properties": "gj:properties",
        "type": {
          "@id": "@type"
        },
        "id": {
          "@id": "@id"
        }
}
```

Figure 3 DCAT Implementation Schema in DAFNI

Comité Européen de Normalisation (CEN)

"Comité Européen de Normalisation (CEN)[42] was the key European standard of geospatial metadata representation." It was operational from February 1992 until September 1999 when it became dormant, although it was revived again. CEN/TC 287 was formed to decide how to adopt and apply ISO standards to Europe" (ISO/TC 211, 2003) producing the European metadata standard for the INSPIRE project.

--ooOOoo--

[42] http://centc287.eu/