Metadata, Data Models and Interoperability

An EDIF Summary Paper

Abstract 1

Introduction 1

The UOs' approach to metadata and data models 2

- An ontology for the UOs 2
- Current status of implementation at the different Observatories 7
  - Manchester and Birmingham 7
  - Newcastle 8

What metadata solution to enable further and longer collaborations between the UOs and DfT? 8

- Metadata used by the UOs' data providers 9
- Guidelines and best practice recommendations for metadata 10
  - General considerations 10
  - Metadata in the air-quality sector 10
  - Metadata in the transport sector 12

A wider perspective on cross-domain metadata 13

- The INSPIRE Directive 13
  - OGC and STAC 14
- Open & Agile Smart Cities (OASC) 15
- How do INSPIRE, OGC and OASC relate to each other? 16

Stepping forward 17

Conclusions 18

Glossary 20

Appendix A 21

- CDBB and the National Digital Twin program (NDTp) 21

Abstract

Metadata has always been of paramount importance in any discipline that involves the production and exchange of quantifiable information, as it represents the means by which such information has to be understood and contextualised.

In 2021, this topic was the focus of a collaboration among the Urban Observatories (UOs) of Manchester, Birmingham and Newcastle. The goal was to explore options and establish practices in order to facilitate and standardise the data communication across the cities.
This joint effort led to the (partial) adoption of an ontology-based metadata solution and the implementation of APIs that expose and describe the observatories' corpus of data through a shared vocabulary.

The current collaboration between the UOs and the Department for Transportation (DfT) has rekindled the conversation about metadata and data-models, and is pushing further the exploration of viable options. These include building on the work previously done by the UOs, or adopting a different approach that would better suit the expectations of DfT in the short, medium and long term.

In this summary paper we lay down the work done in 2021 by the UOs on the issue of metadata, and the possible way forward in the light of the ongoing collaboration with DfT.

Introduction

The urban observatories (UOs) are consortia of stakeholders, usually led by an academic institution, that collect data pertaining to the different aspects of the urban environment. Their aim is to enable cross-domain collaborations through the exchange and sharing of data, facilitate research projects, and assist the local administration in designing data-driven, informed policies.

Given the breadth of scope in which the UOs operate, adopting a metadata standard able to adequately describe the variety of the data collected represents a substantial challenge. The single disciplines that constitute the corpus of data of the urban environment have their own specificities, which help shaping the way their data is presented and described. This has the potential to create a certain level of consensus around domain-specific metadata standards. However, merging these different solutions (if and when available) into one cohesive cross-domain metadata framework adds another level of complexity that data professionals need to face when working in the wider context of city data and smart-cities.

In the current state of things, smart-city solutions tend to grow without central planning and central control. Although this has arguably positive implications, the downside is that the cumulative effect of these solutions is absent, due to low structuring and a complementarity of applications that does not translate into integration. In the face of these challenges, a coherent data framework for smart-cities initiatives needs then to confront two main issues: *discoverability* and *interoperability*.

Discoverability

The term discoverability is used to describe the effect of exposing data in such a way that makes it easy for a user (i) to understand what the content of a dataset is and (ii) to navigate toward any specific area of that content. In more technical terms we could say that a data-set is fully discoverable if a piece of software is able to crawl the API that exposes that data-set and get a full picture of its entire content.

Interoperability

Once the content of a dataset is made discoverable, the issue becomes how that data can be understood, used and integrated in a wider context than its original scope.

A semi-formal definition of interoperability is provided by the Data Interoperability Standard Consortium, according to which "*data interoperability addresses the ability of systems and services that create, exchange and consume data to have clear, shared expectations for the contents, context and meaning of that data*".

For example, a measurement of outdoor air-temperature in Manchester should be understood in the same way as a measurement of outdoor air-temperature in Birmingham. This translates into agreeing on a set of terms – a vocabulary – whose elements have a specific meaning that is shared across multiple and diverse stakeholders (data providers, brokers, data consumers, etc).

The UOs' approach to metadata and data models

The issues of discoverability and interoperability both point toward the notion of *ontologies* and the *Semantic Web* they enable. An "*Ontology provides a formal, explicit specification of a shared conceptualisation of a domain*" [1]. In other terms, it provides a common vocabulary and a grammar that gives unambiguous meaning to the entities of a specific domain and their relations. By providing a shared understanding of common domains, an ontology facilitates knowledge sharing and discoverability over distributed systems and plays a major role in solving the problem of interoperability between applications across different organisations[2].

An ontology for the UOs

The data pertaining to the urban environment is extremely diverse in nature. A possible segmentation can be done across its generic sub-domains and their features of interest, like the built environment, the natural elements (including the atmosphere), the local population, etc. Another possible classification can be done according to the propensity of the data to change over time, where we distinguish between static data (describing static assets, like the number of lamps along a street), persistent data (like the socio-economic metrics captured every few years through a census) and streaming data (describing a constant flow of information, like air-quality measurement, traffic data, etc).

---

[1] M. Paolucci, T. Kawamura, T.R. Payne, and K.P. Sycara, "Semantic Matching of Web Services Capabilities", In Proceedings of International Semantic Web Conference, 2002, pp.333-347.

D. Tidwell, "Web Services-The Web's Next Revolution", IBM Web Service Tutorial, 29 Nov. 2000, http://www106.ibm.com/developerworks/edu/ws-dwwsbasics-i.html.

[2] M.M. Taye, "Understanding Semantic Web and Ontologies: Theory and Applications", Journal of Computing 2(6): 182-192 (2010).

Given this high heterogeneity, devising an all encompassing ontology for city data is a very challenging endeavour. A comprehensive review of smart-city ontologies and their applications revealed that the vocabularies used in smart-city initiatives remain mainly confined to domain-specific sectors, like energy, health, economy, environment, crisis management, security and privacy, etc.[3] The adoption (or development) of an ontology seems then to be strongly bound to the relevant context, target groups and intended use-cases.

As the UOs' activities are in good part still exploratory in nature, the adoption of a given ontology (or set of ontologies) was initially hampered by the lack of clearly identifiable use-cases and target groups that could have guided such a choice. The UOs thus opted a more pragmatic approach based on the observation that the vast majority of the data collected by the observatories are sensor-generated, streaming data. The quest then turned into finding an ontology that was expressive enough to describe the corpus of assets (sensing devices) in each observatory, the capabilities and specifications of such assets (e.g. reading frequency, observable measured, etc.) and the context in which these assets are employed.

The UOs recognised in the SSN/SOSA ontology[4] a good candidate. This ontology was in fact created to describe sensors, their observations, the involved procedures, the studied features of interest, the observed properties, as well as actuators[5]. Also, because SSN/SOSA is domain-agnostic, it can be used to describe data pertaining to different areas of interest, such as traffic, air-quality, meteorology, etc.

The SSN ontology is very rich and articulated. Here we simply present some of the terms and entities that the UOs found particularly relevant for the scope of their activities.

Deployment - *An observatory hosts a number of activities, each of which represents a pocket of specific interests serving a particular purpose (e.g. a research project). The term deployment denotes such a notion, and operatively can be thought as consisting of a set of devices (platforms) that are employed for the particular purpose of a given project.*

Platform - *A Platform is an entity (generally a sensing device, like an air-quality monitoring station) that hosts other entities, particularly Sensors, Actuators, Samplers, and other Platforms.*

Sensor - *A Sensor is a device or agent (including humans) that performs the act of measuring some ObservableProperty (see below). Sensors can be hosted by Platforms. An example is the particulate matter sensor hosted by an air-quality monitoring station (the platform).*

ObservableProperty - *An observable quality, property or characteristic of a FeatureOfInterest (see below). An example is the outdoor air temperature, where the temperature is the ObservableProperty and the air the FeatureOfInterest.*

[3] A. De Nicola, M.L. Villani, "Smart City Ontologies and Their Applications: A Systematic Literature Review", Sustainability 13(10):5578 (2021)

[4] Semantic Sensor Network (SSN) / Sensor, Observation, Sample, and Actuator (SOSA)

[5] https://www.w3.org/TR/vocab-ssn/

FeatureOfInterest - *The thing whose property is being estimated, calculated or measured.*

Observation - *Act of carrying out a Procedure (see below) to estimate or calculate a value of an ObservedProperty of a FeatureOfInterest.*

Procedure - *A workflow, protocol, plan, algorithm, or computational method specifying how an Observation is made. Although this term is part of the SOSA ontology it was redefined and simplified for the use of the UOs (see entity schema below).*

To the above basic set of terms, the UOs agreed to add some other elements to further enrich their vocabulary.

Units - *The unit of measurements used to describe the value of an Observation. Although there exists an entire ontology dedicated to the units of measurements (QUDT), the UOs agreed to create the term units for its own use, without relying on external vocabularies.*

Discipline - *The discipline, or field of investigation, that an Observation refers to. For example, the observations made by a particulate matter sensor pertain to the discipline "atmospheric chemistry".*

Location - *The location of a Platform, described as a geoJson object.*

Aggregation - The aggregation state of the data as sent by the sensing device. The values of this field can be "aggregated" or "instantaneous". For example, if the device sends averages of multiple observations taken in a given time window, the value sent will be aggregated.

Timeseries - *A collection of observations referring to the same ObservableProperty and generally sharing some other characteristics (for example the sensor that made the observations). The term Timeseries was introduced to minimise the redundant information possibly associated with sets of observations taken in (or referring to) similar contexts.*

What follows is the data model of the entities listed above.

Entity	Attribute	Data type
Deployment	@context	array of uris
	@id	uri
	@type	string
	label	string
	description	string
	startDate	string (simplified extended ISO format)
	endDate	string (simplified extended ISO format)
	deployedOnPlatform	array of uris
	hosts	array of uris

Entity	Attribute	Data type
Platform	@context	array of uris
	@id	uri
	@type	string
	label	string
	description	string
	static	boolean
	inDeployment	uri (of deployment)
	location	geoJson object

Entity	Attribute	Data type
Sensor	@context	array of uris
	@id	uri
	@type	string
	label	string
	description	string
	isHostedBy	uri (of hosting platform)
	hasDeployment	uri (of deployment)
	location	geoJson object

Entity	Attribute	Data type
ObservableProperty	@context	array of uris
	@id	uri
	@type	string
	label	string
	description	string
	recommendedUnits	Array of strings

Entity	Attribute	Data type
FeatureOfInterest	@context	array of uris
	@id	uri
	@type	string
	label	string
	description	string

Entity	Attribute	Data type
Observation	@context	array of uris
	@id	uri
	@type	string
	resultTime	string (simplified extended ISO format)
	hasResult	{value: number, unit: string}
	madeBySensor	uri of Sensor
	inTimeseries	uri of Timeseries
	observedProperty	string
	hasFeatureOfInterest	string

	usedProcedures	string

Entity	Attribute	Data type
Procedure	@context	array of uris
	@id	uri
	@type	string
	label	string
	description	string

Entity	Attribute	Data type
Units	@context	array of uris
	@id	uri
	@type	string
	label	string
	description	string

Entity	Attribute	Data type
Discipline	@context	array of uris
	@id	uri
	@type	string
	label	string
	description	string

Entity	Attribute	Data type
Aggregation	@context	array of uris
	@id	uri
	@type	string
	label	string
	description	string

Entity	Attribute	Data type
Timeseries	@context	array of uris
	@id	uri
	@type	string
	startDate	string (simplified extended ISO format)
	endDate	string (simplified extended ISO format)
	hasObservations	collection of Observations
	observedProperty	string
	aggregation	string
	units	string
	madeBySensor	uri of Sensor
	hasDeployment	uri of Deployment
	disciplines	array of strings
	usedProcedures	Array of strings

The graph shown in Fig.1 depicts the relations between the entities listed above.

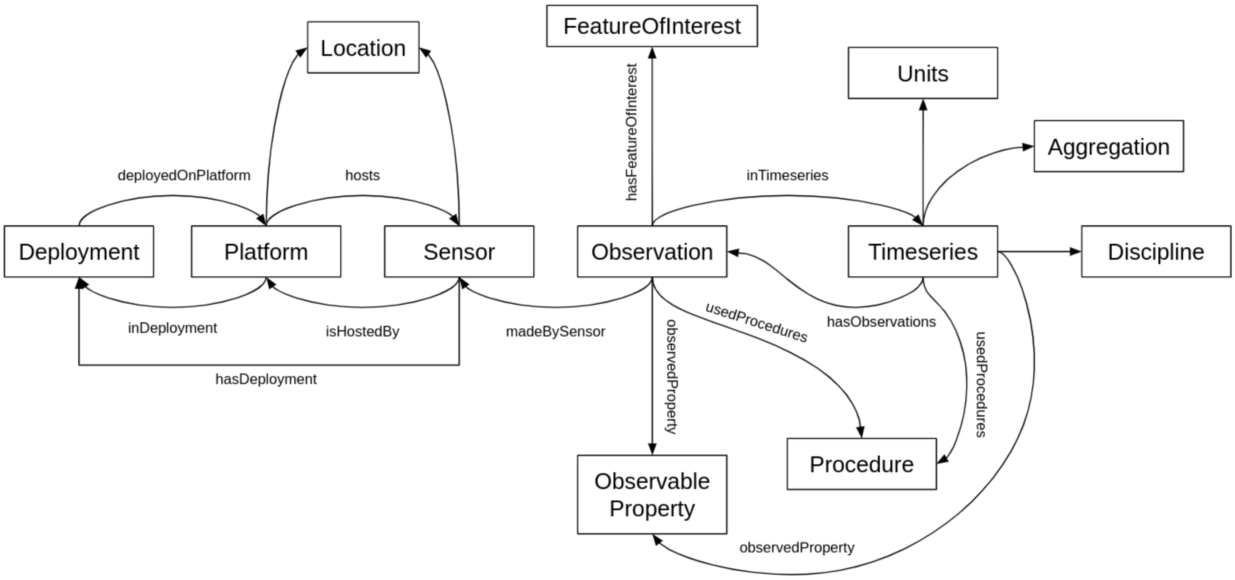

Figure 1 - Entity-relation diagram of the SSN/SOSA terms used by the Manchester and Birmingham Observatories.

Current status of implementation at the different Observatories

The ontology the UOs agreed upon was not required to be identical across the different observatories, nor was the APIs that would expose it to describe the underlying assets and data. The Birmingham Observatory, for example, added some extra attributes to some of the ontology's entities. This was not considered an issue as long as the overlapping set of concepts that all Observatories need to describe are referred to by the same terms, and that the meaning of the overlapping set of attributes is not changed.

Manchester and Birmingham

The APIs implemented by Manchester and Birmingham expose their ontology in a similar fashion. These two APIs can be explored starting from their entry points:

- Manchester → https://muo-backend.cs.man.ac.uk/
- Birmingham → https://api.birminghamurbanobservatory.com/

These two APIs, although useable for any practical purpose, may need some extra work in order to:

- Fix possible orphan hyperlinks (possibly due to some error in the code that implements some of the APIs endpoints)
- Implement a common and wider set of parameters that allow for more specific queries.

Newcastle

The Newcastle Observatory seems not to have implemented as yet an API that exposes the ontology agreed upon by the UOs. This is due to the fact that the person leading that line of work left his position before the ontology could be integrated into the new API.

The Newcastle Observatory, however, does have a public API, although not compliant with the discussed ontology specifications (https://newcastle.urbanobservatory.ac.uk/api_docs/). This API was used in the collaboration between the Newcastle Observatory and DfT during the Covid-19 pandemic to gather traffic and public transport data and better understand travel patterns and behaviours.

The main entities that the API describes are the following:

> **Sensor** - *It represents a physical sensor. It is the homologous of the Sensor entity in the ontology, but it is not structured, nor referred to, in the same way.*
>
> **Variable** - S*imilar to ObservableProperty.*
>
> **Theme** - *Similar to Discipline.*

The documentation of the API can be found at the following address:
https://newcastle.urbanobservatory.ac.uk/api_docs/doc/variables-json/

What metadata solution to enable further and longer collaborations between the UOs and DfT?

Through its collaboration with the UOs, DfT is looking to share and access data pertaining to traffic flow, active travel, micro-mobility, meteorology and air quality.

As noted previously, choosing a metadata solution strongly depends on the discipline the data pertains to and the intended use-cases. If the data of interest covers multiple disciplines and there is a lack of clear use-cases, a possible choice for DfT would consist of building on the work done by the UOs (specifically Birmingham and Manchester). The advantage of adopting this solution is two-fold:

1. The ontology and API have already been implemented to a good degree in two of the three observatories involved in the collaboration with DfT.

2. The ontology used is domain-agnostic and only needs minimal adaptation (if any) to be applied to different disciplines, as long as these rely on sensor-generated, streaming data.

Another possibility consists of adopting a new set of metadata, tailored more specifically around the fields of transport and air quality, being these DfT's main focus.

In the section below, we provide an overview of possible metadata solutions in the field of transportation and air quality. In particular we report on (i) metadata used by the different providers (brokers) from which UOMs collect data; (ii) current guidelines and best practices as recommended by relevant governing bodies.

Metadata used by the UOs' data providers

The metadata used by the UO's data providers is detailed in an accompanying document, titled "Brokers' metadata" and available at the following address:

https://docs.google.com/spreadsheets/d/1oD4nK9fhSMkMZoHeDsXE0rOC29gRzEFS59pKtUpSmy8/edit#gid=0

We highlight that, as a general trend, these metadata solutions do not rely on any ontology, nor on linked-data for discoverability, and seem to be built around the specific characteristics of the devices and the providers' business model. Hence, the amount and type of information can vary substantially from one provider to another, still within the same domain. For example, Vivacity provides data (and metadata) not only about traffic volume, but also journey times. Ca-Traffic, instead, does not provide journey times, but provides a certain number of statistics on vehicle-counts and speed, such as mean average, 85 quintile, standard deviation, times and values of peak traffic.

As for air quality, the data is provided with a set of metadata descriptors usually consisting of:

- name of the quantity measured (i.e., NO2, PM2.5, etc.)
- units
- time of the reading
- location from which the reading was originated (this can be associated with the reading itself or the device that generated the reading).

Information is also provided about the devices generating the data. This usually consists of:

- id of the device
- location
- timestamp of the last reading and /or timestamp of last seen on network
- measurement count
- power status (battery voltage, battery charge, external voltage)
- operational status (the current operation status of the device, e.g. calibration, active, etc.)

Guidelines and best practice recommendations for metadata

General considerations

The Central Digital and Data Office of the UK government has compiled a guide for APIs and data standards. With regard to metadata, the recommendations are the following:

- The response of an API query (with the metadata that such a response includes) should be provided in JSON format. JSON objects are vastly preferred to, for example, simple arrays of values, as the latter can limit the ability to include metadata about results and limit the API's ability to add additional top-level keys in the future.

- Data attributes and descriptors should be devised to meet user needs.

- It is preferable to use linked-data in the API responses. This makes the data more programmatically accessible through its metadata. Linking data is achieved by returning Uniform Resource Identifiers (URIs) instead of strings to cross reference different resources within a dataset or across different datasets. Using existing hypermedia standards, like JSON-LD, makes it easier to find such resources.

Although the above guidelines do not mention the concept of ontology, they point towards enhancing data accessibility through linked data. In this regard it must be noted that the adoption of an ontology in publishing linked data helps in data integration and schema alignment. In other terms, by providing a shared vocabulary to refer to the entities described by the metadata, it is possible to go from better data accessibility – as enabled by linked data – to data discoverability and interoperability[6].

Metadata in the air-quality sector

Historically, air quality data in the UK have been captured, processed and used in various formats and for a range of purposes. This resulted in a lack of structure in the architecture of the overall system of datasets, lack of standard formats and inconsistent metadata. This makes it difficult to quickly know if data are compatible or comparable[7].

In a scoping study dated 2010 and commissioned by DEFRA, AEA Technology recommended a plan of action aiming to:

- standardise the approach to linking data to a point in space (spatial) and time (temporal)
- standardise the approach to linking data to other datasets
- refine and define UK data formats, metadata standards and overall architecture.
- reach compliance with the INSPIRE Directive (Spatial Information in the European Community)

[6] Dutta, B. (2014). Symbiosis between Ontology and Linked Data. Librarian, 21(2), pp. 15-24
[7] https://uk-air.defra.gov.uk/assets/documents/reports/cat09/1102161123_Data_Integration_Report_v`1-2.pdf

Presumably as a result of that study, today DEFRA provides a searchable catalogue of UK air quality monitoring and emissions datasets[8]. This catalogue can be queried for available resources nationwide and new datasets can be added.

To be registered to the catalogue, a resource is required to comply with the INSPIRE Directive, a set of rules on interoperability that define how to publish and share spatial data among public sector organisations. These specifications cover a number of themes including environmental monitoring facilities and transport networks[9].

The INPSIRE data specification for environmental facilities can be found at this address: https://inspire.ec.europa.eu/file/1535/download?token=nbWSNl_l. A simple example is provided where an air quality monitoring station with one sensor monitoring ozone concentration (O3) is installed on a tower hosting meteorological equipment. The monitoring station, which we call AQ_Station is an instance of the *EnvironmentalMonitoringFacility* Class. This object describes the Environmental Monitoring Facility itself, so its focus is more on organisational information pertaining to this facility such as the legal background leading to its establishment or the organisation responsible for maintenance. A second instance of the *EnvironmentalMonitoringFacility* Class is used to describe the O3 sensor, here called O3_Sensor, within the station. This object contains information specific to the O3 measurement process, partially through attributes within this class and partially through its link to an object of the class *ObservingCapability* (O3_ObservingCapability), which provides the more measurement specific parts of this information. The O3_Sensor is linked to the AQ_Station via the *hierarchy* association. The O3_ObservingCapability contains references to:

- an *ObservableProperty* object, describing the phenomenon being measured);
- a *FeatureOfInterest* object, describing the thing whose property is being measured;
- a *INSPIRE_OM_Process* object, detailing the measurement process.

Finally, a third object of the *EnvironmentalMonitoringFacility* class, “Meteo_Tower”, is created to represent the measurement tower.

Fig.2 shows the entity-relation diagram for the example just described. It is worth noticing that some terms used in this example are the same as in the SSN/SOSA ontology, although they may have a different schema.

[8] https://uk-air.defra.gov.uk/data/data-catalogue
[9] https://inspire.ec.europa.eu/Technical-Guidelines/Data-Specifications/2892

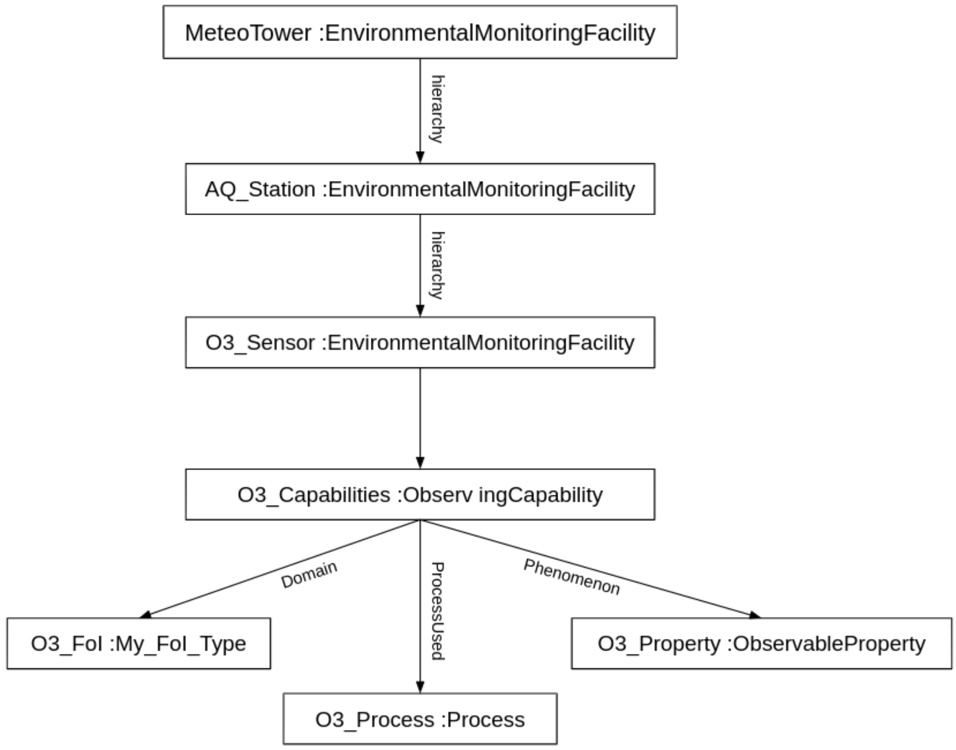

Figure 2 - Entity-relation diagram of some of the terms used in INSPIRE for the environmental sector.

Metadata in the transport sector

Road data is not easily accessible or discoverable in the UK, as there is not a unified platform through which the corpus of available data is made accessible. Data discoverability is also hampered by poor quality metadata and out of date / orphaned datasets.

A study carried out by FutureGov and DfT to better understand the needs of both publishers and users of transport data concluded that a *minimum-viable approach to metadata* should be adopted[10]. This means a short-term focus on promoting discoverability over metadata and data quality. This approach recognises that asking for large amounts of metadata increases the burden on data publishers. This in turn translates into a lower number of datasets and, as a result, less overall discoverability. Improving and deepening metadata should be a long-term goal.

This conclusion seems somewhat counterintuitive, as it is generally accepted that a comprehensive and systematic approach to metadata would facilitate data discoverability. The proposed approach should probably be understood as an invitation to be pragmatic, with the implementation of metadata standards being driven by how data consumers intend to use the data, and then building on this for further refinement and enrichment of the standard. This approach then seems to be advocating for a *de facto* standard built on use cases of traffic data.

[10] https://wearefuturegov.com/case-study/dft-national-access-point

Although INSPIRE (see previous section) provides a set of terms to describe transport networks, it does not seem to provide a vocabulary for traffic data as such. The metadata proposed appears to be intended specifically for the description of transport infrastructure. Even so, it is not clear to what extent this has been implemented in datasets and resources exposing transport / traffic data.

A wider perspective on cross-domain metadata

The INSPIRE Directive

In the previous section we mentioned the INSPIRE Directive in the context of metadata for air-quality. However the INSPIRE Directive has a broader scope. Specifically it was created to address and solve the issues that typically prevent the widespread use of spatial data across Europe[11]. The directive specifies a number of data management principles to overcome the difficulties associated with sharing and using spatial data:

- Data should be collected only once and kept where it can be maintained most effectively.
- It should be possible to combine seamless spatial information from different sources across Europe and share it with many users and applications.
- It should be possible for information collected at one level/scale to be shared with all levels/scales; detailed for thorough investigations, general for strategic purposes.
- Geographic information needed for good governance at all levels should be readily and transparently available.
- Easy to find what geographic information is available, how it can be used to meet a particular need, and under which conditions it can be acquired and used.

To ensure that the spatial data infrastructures of the Member States are compatible and usable, the INSPIRE Directive requires that common Implementing Rules are adopted. These rules pertain to:

- Metadata
- Data Specifications
- Network Services
- Data and Service Sharing
- Spatial Data Services
- Monitoring and Reporting

These Implementing Rules are binding in their entirety.

In addition to the Implementing Rules, non-binding Technical Guidance documents[12] describe detailed implementation aspects and relations with existing standards, technologies, and practises. These non-binding guidelines and good practises[13] are given as examples of how to implement the legal provisions

[11] https://inspire.ec.europa.eu/inspire-policy-background/27902
[12] https://inspire.ec.europa.eu/Technical-Guidelines
[13] https://inspire.ec.europa.eu/portfolio/good-practice-library

of the INSPIRE Directive. The examples provided in the Technical Guidance cover a wide range of themes and domains, [14] laying down the foundation for a cross-domain approach to metadata.

OGC and STAC

The Open Geospatial Consortium (OGC) is a worldwide community consisting of over 500 businesses, government agencies and research organisations committed to improving access to geospatial information[15]. OGC creates free, publicly available geospatial standards that cover a variety of topics related to geospatial information.

OGC standards are divided into three types:

- **Abstract Specification** – The Abstract Specification provides the conceptual foundation for most OGC specification. Open interfaces and protocols are built and referenced against it, thus enabling interoperability between different brands and different kinds of spatial processing systems. The Abstract Specification provides a reference model for the development of OGC Implementation Standards.
- **Implementation Standards** – The Implementation Standards detail the interface structure between software components. An interface specification is considered to be at the implementation level of detail if, when implemented by two different software engineers in ignorance of each other, the resulting components plug and play with each other at that interface.
- **Community Standards** – A Community standard is an official position of the OGC endorsing a specification or standard developed external to the OGC. Community standards can serve two purposes:
    - to bring *de facto* standards from the larger geospatial community to be a stable reference point;
    - to bring new, but implemented, standards to the OGC to form the basis for further refinement and development of interoperability between other OGC standards.

OGC maintains a public Standards Roadmap for every standard currently in development[16].

In relation to the work done by the UOs, it is worth mentioning that some of the OGC standards pertain to observations and measurements. The document detailing this standard provides specifications for data models describing concepts such as Deployment, Host, Observation, ObservableProperty, ObservingProcedures and FeatureOfInterest among others. These terms and the concepts they refer to are similar to those used in SOSA/SSN. However, the level of maturity (or consensus) of the OCG standard for observations and measurements does not seem to have reached that of SSN/SOSA. At the time of writing, the OGC standard for observations and measurements has still to be approved and is a few steps away from public release.

[14] https://inspire.ec.europa.eu/Themes/Data-Specifications/2892
[15] https://www.ogc.org/
[16] https://www.ogc.org/roadmap

OGC has also created a set of API specifications, the *OGC API-Features*, for creating, modifying, and querying spatial data on the Web. Part of these specifications pertains to the discovery and use of observation data. In particular, the OGC Observation-oriented API leverages the Observations and Measurements model to directly allow filtering on featureOfInterest, observedProperty and procedure.

Further work on the OGC API-Features has been promoted and implemented by the Radiant Earth Foundation[17]. This resulted in the creation of the STAC (Spatio Temporal Asset Catalog) API standard. STAC adopts a set of core principles, such as REST, JSON over XML[18], smaller specification approach, and the use of GitHub for the creation and evolution of the API specs[19]. Moreover STAC adds some functionalities to OGC API-Features, resulting in richer filtering, transactions, sorting, versioning, and cross-collection search.

Currently OGC and Radiant Earth Foundation are working to align the two API standards.

Open & Agile Smart Cities (OASC)

Open & Agile Smart Cities – OASC in short – is the international city network working with local public administrations of all sizes to support their digital transformation. OASC aims to overcome the chicken-and-egg situation of smart-cities where no system can scale and spread because there are no standards, and there are no standards because there is no widespread deployment.

OACS works towards a minimal common technical ground for cities and communities, called Minimal Interoperability Mechanisms, or MIMs. Implementations of data solutions can be different as long as crucial interoperability points in any given technical architecture use the same interoperability mechanisms. The MIMs are vendor-neutral and technology-agnostic, meaning that anybody can use them and integrate them in existing systems and offerings.

There are currently seven MIMs addressing different aspects of data interoperability and governance. Details about the MIMs are given at this webpage. We notice that different MIMs are at different stages of definition and development. At the time of writing some MIMs – like MIM6, pertaining to data security – are just placeholders while others – like MIM1 and MIM2 – are at a more mature state of development. Here we briefly describe the first two MIMs, which deal with the core aspects of data interoperability:

MM1 - Context
Context information contains information about the status of real-world entities, either physical or conceptual (e.g. a sensing device, a vehicle, a building, the weather conditions, etc.). MIM1 focuses on the specifications of the API in charge of creating, retrieving, updating and deleting (CRUD) such information. In particular, MM1 requires that such an API follows the NGSI-LD standard, designed by ETSI.

[17] https://medium.com/radiant-earth-insights/spatiotemporal-asset-catalogs-and-the-open-geospatial-consortium-659538dce5c7
[18] https://jsonapi.org/
[19] https://github.com/radiantearth/stac-spec

Context information is provided through a Context Information Manager (CIM), or Context Broker, a service that implements an NGSI-LD compliant API. Some of the CIMs recommended by OASC are Orion-LD (FIWARE foundation), Scorpio (NEC) and Stellio (EGM).

MIM2 - Data Models

Data models are formal representations of the entities in the real world. The adoption of agreed-upon data models is essential to support a digital market of interoperable and replicable smart solutions across multiple sectors. The set of data models suggested by OASC is the Smart Data Models. This was created to support the adoption of common, compatible representation of context entities to foster better interoperable smart solutions. In particular, Smart Data Models are designed to meet the needs of different domains based on concrete use cases and common sector-specific applications. Smart Data Models follow the NGSI-LD specifications.

Importantly, OASC has established a formal partnership with the FIWARE foundation, which provides a technical implementation of the MIMs and the NGSI-LD standard through a curated framework of open source data-platform components.

How do INSPIRE, OGC and OASC relate to each other?

The EU INSPIRE Directive introduces general rules to establish an infrastructure for spatial information. In particular it ensures that spatial datasets and services are available in a consistent format, making them more easily shared. This is done through the guidance of “Implementing Rules”.

If these rules are to be adopted, it is necessary to rely on and apply a set of standards. The OGC enables the implementation of the INSPIRE Directive by providing the necessary standards for geospatial data. In particular, geospatial data is offered either by a Web Coverage Service (WCS) as *features*[20], or by a Sensor Observation Service (SOS) as *observation results*[21].

It is worth mentioning that the OGC standards and APIs were originally developed following specifications that became outdated, for example the use of XML for data models in the API responses. As the web world has clearly moved towards different specifications (such as JSON), the OGC standards and APIs are being reworked. This endeavour was partly promoted by Radiant Earth Foundation and resulted in the STAC API standard.

OASC introduced the concept of Minimum Interoperability Mechanisms (MIM) to formalise and standardise data resources and sharing protocols within and across different sectors. This refers to specifications and architectures that do not attempt to encompass every conceivable aspect of a city but instead allow systems to be connected only when and where needed. They enable future innovation and adaptation without requiring wholesale reconstruction of existing functioning systems. The MIMs represent an agile approach to data interoperability and their definition is informed by the feedback received from concrete use-cases applications and stakeholders operating in relevant sectors.

[20] https://www.ogc.org/standards/wcs
[21] https://www.ogc.org/standards/sos

It must be noted that OGC has standards that could be chosen to implement some of the OASC MIMs, for example APIs (MIM1) and data models (MIM2). Currently OASC recommends NGSI-LD (MIM1) and Smart Data Models (MIM2) which is a parallel solution stack to the OGC standards and as such partially competing. In this sense, both OGC and OASC recommended standards could be used (at least partially) to implement the guidelines for best practices specified in the INSPIRE Directive.

Stepping forward

During the exploratory phase of the collaboration between the Observatories and DfT, the adoption of a MIMs-based approach to data description (metadata) and provision (API) was deemed as the preferable way forward. This was due to a number of different but concomitant reasons:

- The Open & Agile Smart Cities (OASC) – the entity that defines and promotes the adoption of the Minimal Interoperability Mechanisms (MIMs) – is building the foundation for a global market that enables digital services for cities and communities to scale sustainably. At the time of writing more than 150 cities worldwide have already joined OASC.

- The Smart Data Models (MIM 2) are designed by the OASC community based on real world use-cases, hence meeting the needs of specific sectors while avoiding unnecessary overabundance of details.

- The MIMs represent an agile approach to data interoperability and data integration across different domains, allowing for manageable changes whenever the relevant data models need to be updated.

- The NGSI-LD API (MIM 1) allows users to provide, consume and subscribe to context information by providing advanced geo-temporal queries, and subscription mechanisms.

The adoption of this MIMs-based approach to data description and provision is facilitated by the Urban Data Exchange (UDX), a third party organisation whose mission is to harmonise the data coming from different sources (in our case the Observatories) by mapping it to the relevant Smart Data Models (MIM 2) and exposing it through an NGSI-LD API (MIM 1).

It should be noted that the definition of the Smart Data Models is shaped and adjusted according to the experience on the field of the stakeholders that make use of them, leading to an ever increasing and informed consensus. This applies in the specific to the data models that the Observatories and DfT agreed on adopting to describe air quality and traffic data. The schema of the relevant Smart Data Models (AirQualityObserved and TrafficFlowObserved) can be found at the Smart Data Model GitHub repository. However, the Observatory, DfT and UDX are engaged in a conversation aiming to identify any further piece of information (metadata) that these data models do not currently capture. An interest emerged in including some extra information to describe the ownership and provenance of the data as well as the context in which such data is produced. This is intended to provide end users with a set of factual pieces of information that allows them to understand the data in terms of reliability as well as present/future availability.

Conclusions

Historically data resources of public utility have been developed without following any metadata standards. As a result the vast array of datasets available corresponds to a comparable variety in metadata solutions.

It is universally acknowledged that this is not a desirable situation, as it hampers the discoverability, usability and interoperability of important public data resources. Studies commissioned by the governing bodies of relevant domains (*i.e.* DfT for transport, DEFRA for the environment) agree on the necessity to move towards a shared approach to metadata, but they seem to focus on different strategies, presumably depending on the level of maturity of the data resources and infrastructure pertaining to the different disciplines.

The environmental sector seems to be one in which the adoption of a metadata standard has occurred to a considerable extent. As recommended by AEA Technology in their scoping study of 2010, the standard adopted is INSPIRE, a rich ontology capable of describing the facilities and infrastructure of different domains of public interest. Unfortunately, this process of standardisation does not seem to go together with the implementation of open APIs that everyone can access. For example, although DEFRA had to adopt the INSPIRE standard to be registered to the UK Air Quality Data Catalogue (https://uk-air.defra.gov.uk/data/data-catalogue), it does not provide users with an API that allows for programmatic access to air-quality data. This data is rather presented in tabular form as html documents and necessitates manual querying.

As for the transport sector, the approach recommended by FutureGov is a pragmatic one, where the design of a metadata standard should start with the short term goal of meeting the needs of specific target groups and projects. It is suggested that a richer and more expressive metadata standard should be the focus of a long term effort. This would be achieved through a feedback loop between data publishers and users, which is intended to guide the resource publishers in creating, maintaining and continuously improving their service. However there seems to be no clear roadmap leading to a wide-consensus over a more articulated metadata solution that grows organically from meeting shorter term needs.

Regardless of how close a given domain of public interest is to reaching consensus about a metadata solution (or how far it is from it), the general trend is towards the adoption of at least some basic guidelines as detailed in the API technical and data standards, published in the UK government website. In this document the suggestion to use linked data is particularly relevant to the topic of data accessibility, although it does not necessarily lead – in and of itself – to data discoverability and interoperability.

In this effort towards standardisation, the Urban Observatories of Newcastle, Birmingham and Manchester have worked on the definition of a limited set of terms – borrowed or inspired by the SSN/SOSA ontology – to be used as their metadata. This solution has the advantage of being domain-agnostic as it only defines terms that pertain to sensor data and the facilities (devices, sensors, etc.) that generate it.

In regard to the collaboration between the Observatories and DfT, the adoption of a metadata solution was guided by the desiderata of (i) adopting an agile approach that would require minimal restructuring to accommodate future innovation in existing functioning systems and (ii) using data models informed by relevant stakeholders' feedback and based on concrete sector-specific use cases. The Open and Agile Smart Citiess's (OASC) approach to metadata and data models seemed the most suitable to meet those criteria. In particular, the OASC approach focuses on Minimal Interoperability Mechanisms (MIMs) to achieve interoperability of data, systems, and services based on an inclusive list of baselines and a minimal common ground.

The data provided by the Observatories is harmonised (in particular, made MIMs compliant) by the Urban Data Exchange (UDX), a third party organisation that aims to facilitate the sharing and access of live urban data streams on a large scale. This is done by collecting data from the Observatories, re-formatting it through Smart Data Models and re-exposing it via an NGSI-LD API. By using UDX' services, the Observatory and DfT can demonstrate how urban data can be collected and exposed using a unified metadata / data-model solution.

Glossary

Linked Data - The term refers to structured and interlinked pieces of data that refer to each other through their corresponding URIs.

Metadata - Metadata is data that provides information about other data (the underlying data), but not its content. This information helps to extract meaning from the underlying data and describes different aspects of the context in which such data has to be understood.

NGSI-LD - Information model and API for publishing, querying and subscribing to context information. It is meant to facilitate the open exchange and sharing of structured information between different stakeholders.

Ontology - Set of concepts pertaining to a subject area that are described through their properties and relations among them.

Semantic web - Extension of the existing World Wide Web, which provides software programs with machine-interpretable metadata of the published information and data. As a result, computers are able to make meaningful interpretations similar to the way humans process information to achieve their goals.

URI - Short for Uniform Resource Identifier, it is a unique sequence of characters that identifies a physical, virtual or logical resource on the internet. URIs may be used to identify anything, including real-world objects, such as people and places, concepts, or information resources such as web pages and books.

Appendix A

CDBB and the National Digital Twin program (NDTp)

The Centre for Digital Built Britain (CDBB) is a consortium that brings together industry, academia, and policy makers in order to consider the wider effects of the digital agenda on society and the economy. Its mission is to develop and demonstrate policy and practical insights that exploit emerging technologies in order to enhance the natural and built environment, drive up commercial productivity, and improve citizens' well-being.

CDBB convenes multiple programmes of work, most notably the Construction Innovation Hub (specifically concerned with the construction sector) and the National Digital Twin Programme (NDTp). These programmes work together toward the development of digitally enabled infrastructure that serves as a platform for economic, social and environmental sustainability.

The focus of the NDTp is on (i) creating an ecosystem of connected digital twins to foster better outcomes from our built environment; (ii) ensuring secure resilient data sharing and effective information management; (iii) providing coordination and alignment among key players.

The two survey papers "A survey of Top-Level Ontologies" and "A Survey of Industry Data Models and Reference Data Libraries", were used as the foundation for the work described in "The Pathway towards an Information Management Framework" (IMF). The survey papers identify the requirements and inform the ontological choices for a Foundation Data Model (FDM). The FDM, built upon a top-level ontology, is a key component of the IMF and a basis for ensuring consistent data across the National Digital Twin.

Following the publication of the Survey documents the pragmatic and technical requirements for the Foundation Data Model have been developed and consideration has been given as to whether any existing Top-Level Ontologies could be used as a suitable start-point.

There are four Top-Level Ontologies that meet all the technical requirements: BORO, IDEAS, HQDM and ISO 15926-2. They are distinct from the other Top-Level Ontologies in being 4-dimensionalist. These allow us to see individual objects as four-dimensional, having both spatial and temporal parts.