Dealing with diversity in a smart-city datahub

Conference or Workshop Item

How to cite:

d'Aquin, Mathieu; Adamou, Alessandro; Daga, Enrico; Liu, Shuangyan; Thomas, Keerthi and Motta, Enrico (2014). Dealing with diversity in a smart-city datahub. In: Proceedings of the Fifth Workshop on Semantics for Smarter Cities, CEUR Workshop Proceedings, CEUR-WS.org, pp. 68–82.

Dealing with Diversity in a Smart-City Datahub

Mathieu d'Aquin, Alessandro Adamou, Enrico Daga, Shuangyan Liu, Keerthi Thomas, and Enrico Motta

The Open University, United Kingdom
{mathieu.daquin, alessandro.adamou, shuangyan.liu, keerthi.thomas, enrico.motta}@open.ac.uk

Abstract. In this paper, we present the data curation approach taken by the MK:Smart project, creating a large data repository of datasets about all aspects of the city of Milton Keynes in the UK and its citizens. The issues faced here, which we believe will become more and more common to large, data-centric smart-cities initiatives, is the one associated with the diversity of these thousands of datasets in terms of the licenses, policies and terms they are associated with them. We describe this repository of datasets, the MK Datahub, and its architecture to create data workflows from original sources to applications. We focus on the approach taken to record, in a structured, ontology-based way the components of the licenses and policies of each dataset, as well as the tools we are developing to manage such representations and to reason with them.

Keywords: Data curation, licenses, policies, ODRL, smart-city

1 Introduction

Data and the intelligent processing of data is at the center of smart-city initiatives, whether they rely on the simple publication of open data by local authorities (see for example [2]) or on more complex systems including big data and big computation (see for example [15]). Once a certain scale is reached, such projects have to face not only the amount of data to store, process and deliver, but also their variety. Indeed, they need to deal with data regarding many different aspects of the functioning of the city, its inhabitants, services, infrastructures, etc. Semantic technologies obviously have a role to play in dealing with this variety, supporting the integration of data in a meaningful way accross sources.

However, beyond dealing with semantic interoperability, there is a need to manage the number and diversity of these datasets. Indeed, in projects such as MK:Smart[1], which plan to deal with several thousands of datasets, the curation process of these datasets needs to be streamlined, to avoid for the system and its operators to become overwhelmed with having to deal with the mechanisms

[1] `http://mksmart.org`

associated with the import, management and redistribution of data from diverse sources, in diverse formats, having diverse requierements for storage and update, and most importantly, being associated with diverse policies, licenses and terms of use, which need to be instantiated and managed across the whole workflow of the information services associated with the data repository.

In this paper, we describe the initial approach taken by the MK:Smart project to deal with this diversity, with the intent to create a datahub that scales not only in the amount of data it can hold, but in the number of different situations which dimensions affect the effort associated with curating the datasets and managing the associated workflows. We first give an overview of the MK:Smart project and discuss the building of a large scale datahub to store and re-deliver data from a large variety of sources. We then present the approach currently being deployed in this project, as part of this datahub, to deal with the diverse set of policies, terms and licenses attached to these datasets.

2 The MK:Smart Project and the MK Datahub

In this section, we briefly introduce the MK:Smart project in Milton Keynes (UK), the role of the large data repository which the project is creating (the MK Datahub) and how this repository will support the import, integration and delivery a large amounts of data comning from a large number of sources.

2.1 The MK:Smart Project

Milton Keynes (MK) is a new town in Buckimghamshire, which has been officially designated in 1967. It is now one of the fastest growing cities in the UK, with a population of approximetly 230,000 inhabitants. However, the challenge of supporting sustainable growth without exceeding the capacity of the infrastructure, and whilst meeting key carbon reduction targets, is a major one. MK:Smart is a large collaborative initiative, partly funded by HEFCE (the Higher Education Funding Council for England) and led by The Open University, which aim to develop innovative solutions to support economic growth in MK.

Central to the project is the creation of a state-of-the-art “MK Datahub” which will support the acquisition and management of vast amounts of data relevant to city systems from a variety of data sources. These will include data about energy and water consumption, transport data, data acquired through satellite technology, social and economic datasets, and crowdsourced data from social media or specialised apps.

On this basis, the project is looking to support the creation of new sensor and analytics applications and services, putting a particular emphasis on the areas of transport, energy and water. Indeed, while MK was planned out to support traffic relativly efficiently compared to other towns, mobility is critical to economic activities and intelligent, hybrid transportation systems that intelligently facilitate mobility within the area are needed. Also, with the constant growth of the area in terms of population, there is a risk that the current infrastructures for the delivery of water and energy will reach their limit capacity in the future. Smart technologies here are thought to be used to improve energy and water efficiency, reducing consumption (as well as making generation more efficient) both through data analytics applied at the level of the infrastructure, and through supporting citizens in changing their behaviour.

2.2 The MK Datahub and Data Workflows

The MK Datahub is the central IT and data management infrastructure of the MK:Smart project. It is based on the simple idea that innovative solutions in the three domains mentioned above, and in many others, will require data that can be drawn from a large varietry of sources. The idea is therefore to build a common facility to efficiently manage, integrate and re-deliver such data for applications and services to rely on, reducing development costs for all of these applications, and enabling intelligent data processing mechanisms (mining, analytics, aggregation, alignment, linking) at the scale of the entire city, in a common data infrastructure.

The MK Datahub is made out of several layers. It relies on a hardware infrastructure physically located in MK, but which is then configured similarly to a private cloud (using common virtualisation technologies, and cloud orchestration services) to enable optimal use by the layers above, as well as to meet the (changing) requirements of applications. The data management layer is the main focus of this paper, and is being described in more details below. Its role is to implement the workflow that connect data in their original sources to applications that might want to exploit these data. Finally, additional development and service management mechanisms sit on top of the data management layer, for the management of the users/customers of the datahub, and of the developers of applications.

An overview of the architecture of the data management layer of the MK Datahub is shown in Figure 1. This architecture is itself based on three main layers: The data import layer, the storage layer and the data delivery layer. Starting from the middle, one of the choices made in the design of this architecture is not to rely on one single type of storage technology, but to implement data storage in a distributed and hybrid way (RDMSs as well as graph-based and triple stores). The reason for this choice is first that maintenance and management of smaller storage components distributed amongst dedicated servers is easier and more robust than it would be with a unique warehouse when talking about thousands of datasets, the sources of most of them being out of our control. Also, the development cost of import pipelines are reduced when we can choose the most appropriate storage format amongst a number of options for each data feed to be considered.

The main goal of the bottom layer is to create the pipelines to import each of the thousands of data sources to be included in and re-delivered through the datahub. Here already, the challenge is in the diversity of these data sources. Indeed, each might rely on a different format, a different mode of transfer, have

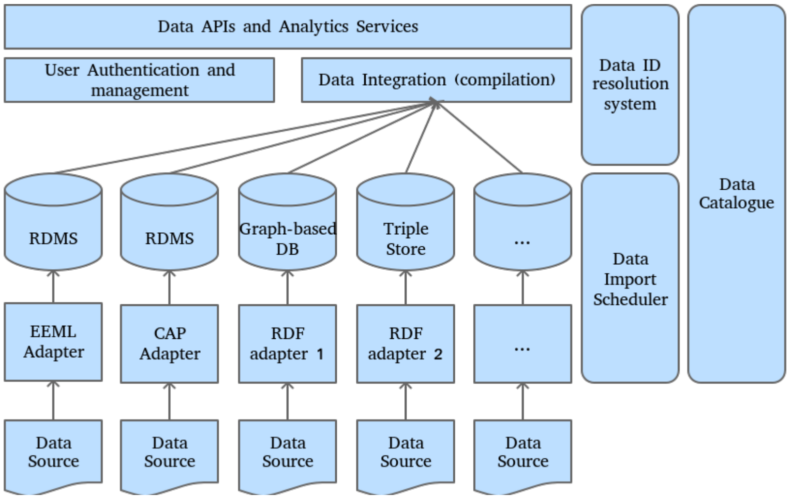


Fig. 1: Overview of the architecture of the data management layer of the MK:Datahub

different constraints attached to it, a different update rate, etc. The challenge here is therefore to create a data processing pipeline framework which allows at the same time the creation of ad-hoc processing pipelines for each of the considered data sources, but also the reuse of processing components and their automatic invocation to reduce their development and maintenance costs. The approach taken here is to create a generic wrapper for data processing services that can be distributed amongst the various (internal) processing servers available to the datahub. Each processing service is developed independetly, but uses standard communication mechanisms (here, we rely on Apache ActiveMQ[2]) as well as conventions for their naming and the naming of communication channels (queues and topics in ActiveMQ) to facilitate the creation of simple data import pipeline as sequences of components that can be executed automatically by a scheduler. The scheduler, takes information from the data catalogue of the datahub (see Section 3.2 about the data catalogue, especially focusing on policies and terms) about the original data feed, the pipeline, the storage destination and the update rate of the data. If decomposed in a sufficiently granular way, generic components developed for one pipeline (e.g. transformation of statistical data out of a business intelligence tools into RDF Data Cube) can therefore be reused in the creation of other pipelines, and pipelines are naturally distributed in the services that implement the components.

Of course, the "lazy" approach taken in the import and storage layer have an impact on the top layer regarding re-delivery. Indeed, the goal of this layer is to provide data to external applications and services in a homogeneous and convenient way, through both feed-centric and entity-centric (URI dereferenc-

[2] `http://activemq.apache.org/`

ing) Web APIs. This can indeed be called a “lazy” approach to data integration as we here only integrate data at the last step of re-delivery, leaving data feeds isolated in independent databases in the datahub. Therefore, data integration in the MK Datahub can be likened to dynamic compilation, where information about a particular entity (using a global URI) is drawn at runtime from various databases, put together in a common canonical format using canonical, global identifiers (URIs) for connected objects, and then transfered to the client in the specific format of their choice (RDF, JSON, XML, CSV, etc). The (ongoing) implementation of this has therefore to face many technical challenges, including the manipulation of an efficient intermediary format to manipulate results of distributed queries from many, diverse databases, the caching of these intermediary results and their update, as well as the tracing of the sources of data included in each integrated result, in particular to enable the propagation of terms and policies as discussed in the next section.

3 Dealing with Diverse Data Sources, with Diverse Policies

Here, we present the data curation approach thought for the MK:Smart project, focusing in particular on the management of the various terms and conditions of use and re-distribution of data that might be associated with each dataset. From a technology point of view, we take an approach that follows the principles of structured data and semantic technologies, ensuring the cataloguing and recording of these terms and conditions in an easily exploitable format. Based on these cataloguing practices, we set up a process for the reviewing and updating of the curated datasets that directly feed into the structured representation of datasets. We also describe the ontology-based approach thought to manage the propagation of the policy rules that represent these terms and conditions as datasets are being processed.

3.1 Data Cataloguing with Structured Data

As described above, one of the key challenges for the management and maintenance of the MK Datahub is the large number of datasets to be included, that are attached to a very diverse set of policies, that might relate to open data and open licenses, as much as to commercial Web APIs, corporate data and personal data. For this reason, establishing a small set of common policies that datasets would have to adhere to is not feasible, but the complexity of manually handling all of the different policies on a case by case basis would also amount to an unfeasible task. The approach taken here is therefore to automate, as much as possible, the recording, tracking and updating of each of the policies attached to each of the datasets, through relying on a data cataloguing system that includes an ontology-based, structured representation of the components of these policies. The idea here, once again, is that many policies from many datasets will share components which, while not necessaraly expressed in the exact same way, represent the same rule or constraint (e.g. attribution, no modification, retention for a certain period of time). Therefore, policies that might be unique to particular datasets might be described using components (rules/constraints) that are shared with many others.

There has been an increasing interest lately in the structured and semantic representation of rights, policies and terms for various kinds of objects, from the ones that focus on specific types of policies (e.g. privacy policies as in P3P [3]) to some focusing on specific types of terms (e.g. open licenses as in CC REL[3]). ODRL (the Open Digital Rights Language) [7] and the ODRL ontology [5] have been recently proposed to provide *"flexible and interoperable mechanisms to support transparent and innovative use of digital content in publishing, distribution, and consumption of of digital media across all sectors and communities. The ODRL Policy model is broad enough to support traditional rights expressions for commercial transaction, open access expressions for publicly distributed content, and privacy expressions for social media."*[4]. ODRL has been used in several initiatives recently to represent policies attached to datasets, and the ODRL ontology provides a simple, yet flexible vocabulary to represent and manipulate both high level objects such as policies, as well as more granular components including rules to represent permissions, prohibitions and duties with respect to the use, re-use and re-distribution of digital assets. It is therefore an ideal base to be used to record, keep track and manipulate information regarding the datasets included in the MK Datahub. As an example, the code below is a simplified representation of the Open Government License (OGL[5]) used by many open data providers following the ODRL ontology:

```
<http://mksmart.org/dc/policy/ogl>
     a odrl:Set;
     odrl:permission odrl:copy ;
     odrl:permission odrl:publish ;
     odrl:permission odrl:reproduce ;
     odrl:permission odrl:commercialize ;
     odrl:duty       odrl:attribute.
```

The requirements for the data cataloguing features of the MK datahub are therefore that they should support not only the description of the datasets, of their sources and of their import pipelines, but should also make it possible to describe the policies attached to the datasets at a granular levels (including their components) in a way compatible with the ODRL model. While several platforms exist for data cataloguing (including, most prominently, CKAN[6] and Socrata[7]), these only support the general description and annotation of datasets,

[3] http://wiki.creativecommons.org/CC_REL

[4] http://www.w3.org/ns/odrl/2/

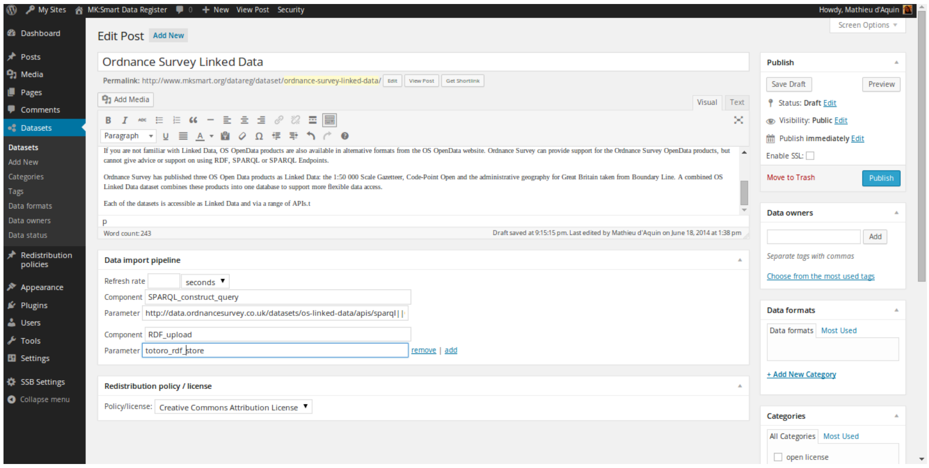
[5] http://www.nationalarchives.gov.uk/doc/open-government-licence/version/2/

[6] http://ckan.org/

[7] http://www.socrata.com/

without going into the details of the processes to manipulate them and, crucially, of the attached policies (beyond a link to the corresponding license).

We therefore developed a data cataloguing system that reproduce similar features to CKAN, but also includes features for data cataloguers to enter the details of other dimensions that are not fully supported in CKAN. This system is being developed as a plugin of the Wordpress content management system (CMS)[8]. Wordpress is the most used CMS on the web, and is open source. Relying on Wordpress meant that we could reuse the basic data model and content organisation it implements, and reuse the interaction patterns Wordpress users are familiar with and that have proven effective through the popularity of the CMS with both professional web developers and general members of the public.



Ordnance Survey Linked Data

Permalink: http://www.mksmart.org/datareg/dataset/ordnance-survey-linked-data/

If you are not familiar with Linked Data, OS OpenData products are also available in alternative formats from the OS OpenData website. Ordnance Survey can provide support for the Ordnance Survey OpenData products, but cannot give advice or support on using RDF, SPARQL or SPARQL Endpoints.

Ordnance Survey has published three OS Open Data products as Linked Data: the 1:50 000 Scale Gazetteer, Code-Point Open and the administrative geography for Great Britain taken from Boundary Line. A combined OS Linked Data dataset combines these products into one database to support more flexible data access.

Each of the datasets is accessible as Linked Data and via a range of APIs.t

p

Word count: 243

Draft saved at 9:15:15 pm. Last edited by Mathieu d'Aquin on June 18, 2014 at 1:38 pm

Data import pipeline

Refresh rate seconds

Component SPARQL_construct_query

Parameter http://data.ordnancesurvey.co.uk/datasets/os-linked-data/apis/sparql||

Component RDF_upload

Parameter totoro_rdf_store

Redistribution policy / license

Policy/license: Creative Commons Attribution License

Status: Draft Edit

Visibility: Public Edit

Publish immediately Edit

Enable SSL: □

Data owners

Separate tags with commas

Data formats

Categories

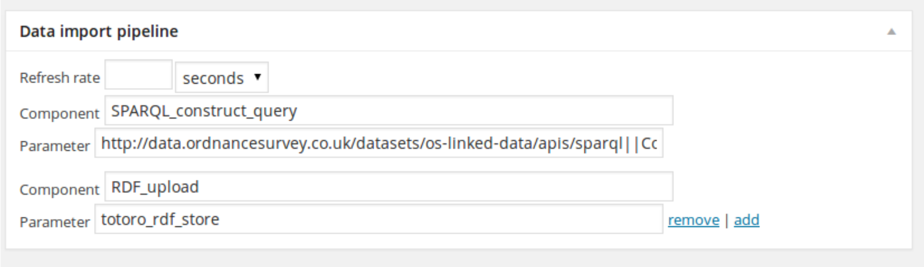
□ open license

Fig. 2: Dataset description editing interface.

The implementation of the *MKS Data Cataloguing* plugin for wordpress is mainly centered arround two “custom post types”: Datasets and Policies. Custom post types in wordpress are ways to create posts that are dedicated to the representation of particular types of objects. In this way, for example, a dataset is described in a specific type of posts which can directly reuse the standard Wordpress post interface for editing its description (see Figure 2). The information associated with a dataset (and so editable through the Wordpress interface) is then customised in two ways: By introducing specific taxonomies and by adding custom attributes. Taxonomies are used in Wordpress to represent information attached, as annotations or metadata to the posts. The standard taxonomies provided by Wordpress for common posts are tags and categories (which we also use for datasets). However, new taxonomies can also be introduced that repre-

[8] http://wordpress.org

sent different aspects to link the dataset to. Here we use taxonomies to represent the owner of a dataset (i.e. its source or origin), its orginal format and a “status” tag to represent the level of inclusion the dataset should have in the datahub. Interestingly, only using this simple mechanism allows us to reproduce most of the features of system such as CKAN, since relying on taxonomies makes standard Wordpress search and browsing functions directly useable on our custom dataset post type.



Data import pipeline

Refresh rate: seconds

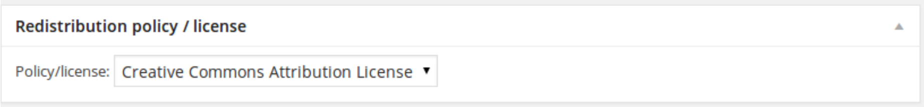
Component: SPARQL_construct_query

Parameter: http://data.ordnancesurvey.co.uk/datasets/os-linked-data/apis/sparql||Cc

Component: RDF_upload

Parameter: totoro_rdf_store

Fig. 3: Description of the import pipeline of a dataset in the data cataloguing facility.



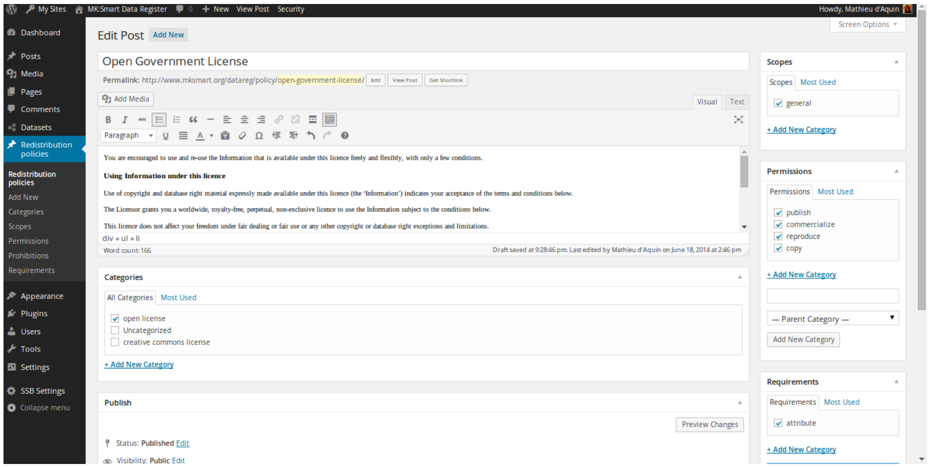
Redistribution policy / license

Policy/license: Creative Commons Attribution License

Fig. 4: Linking a dataset to the corresponding license/policy.

Custom attributes are used for information that requires more expressivity than a simple taxonomy. For dataset, they are used to represent the data import pipeline associated with each dataset (Figure 3), as well the link the the policy/license associated with the dataset (Figure 4).

Similarly to datasets, policies and licenses are represented using a custom post type. The Policy post type also includes the basic fields for describing each policy (e.g. to include the text of the license being described). Four taxonomies are then used to represent each policy in a way that is compatible with the ODRL model (see Figure 5 for the interface to edit/create a policy). The “Scope” taxonomy is used to represent the applicability of the policy. This is especially useful in cases where the dataset/data source is associated with more than one policy, for example with one policy for personal use and another for commercial use. The three other taxonomies (“Permissions”, “Prohibitions” and “Duties”) are used to represent the three types of rules that form components of the policies in the ODRL model. They are pre-populated with “Actions” from the ODRL



Open Government License

Permalink: http://www.mksmart.org/datareg/policy/open-government-license/

You are encouraged to use and re-use the Information that is available under this licence freely and flexibly, with only a few conditions.

Using Information under this licence

Use of copyright and database right material expressly made available under this licence (the 'Information') indicates your acceptance of the terms and conditions below.

The Licensor grants you a worldwide, royalty-free, perpetual, non-exclusive licence to use the Information subject to the conditions below.

This licence does not affect your freedom under fair dealing or fair use or any other copyright or database right exceptions and limitations.

div » ul » li

Word count: 166

Draft saved at 9:28:46 pm. Last edited by Mathieu d'Aquin on June 18, 2014 at 2:46 pm

Categories

All Categories | Most Used

☑ open license
□ Uncategorized
□ creative commons license

+ Add New Category

Publish

Status: Published Edit

Visibility: Public Edit

Scopes

Scopes | Most Used

☑ general

+ Add New Category

Permissions

Permissions | Most Used

☑ publish
☑ commercialize
☑ reproduce
☑ copy

+ Add New Category

— Parent Category —

Add New Category

Requirements

Requirements | Most Used

☑ attribute

+ Add New Category

Fig. 5: Policy description editing interface.

ontology (e.g. "Attribute", "Modify", "Copy", etc). Here again, the advantages of implementing it this way are that it makes it easy for a user to describe a policy or license reusing elements from these taxonomies that have been used to describe other policies and licenses, that it is relatively easy for users to describe the complex objects of policies, and that the long-term use of this system will facilitate the creation of a large, queriable and searchable "knowledge base" of policies and rules, and of their connection with thousands of relevant datasets.

The data associated with datasets and policies in the MKS Data Cataloging plugin are naturally stored in the intermal database of the Wordpress CMS. However, in order to make these data directly usable for the scheduler and data integration components of the datahub (see Figure 1), they are also automatically transformed in RDF and sent to a dedicated triple store using the ODRL ontology for policies, and a mix of other ontologies (VoID [1], DC [17], Prov-O [9], etc.) for datasets, everytime a user saves an entry of one of the two custom types.

3.2 Process to Manage and Update the Data Catalogue

The previous section describes the technical approach to managing diverse datasets, associated with diverse policies and licenses. On the basis of the tool created, a process can therefore be setup to take care of the update and review of the catalogue when new datasets are being included, and when the policies of existing datasets have changed, as summarised in Figure 6.

At the center of this process is the "Data Governance Board", which is in charge of reviewing the policies and licenses attached to each dataset put forward as candidate for inclusion. The board is formed of members of the project management team and of at least one "data cataloguer" who has the specific role to input the results of deliberations of the board into the data catalogue,

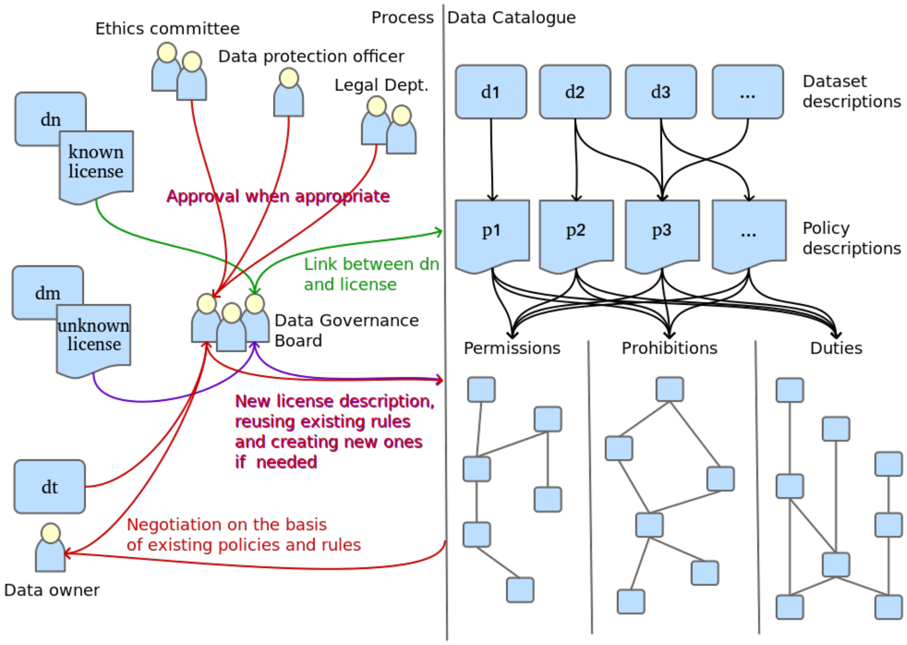


Fig. 6: Overview of the data cataloguing and review process.

through the tool presented in the previous section. For every new dataset, three different situations might appear:

1. *The dataset is associated with a known license* (green arrows): This is the simplest case, as the license would have already been reviewed and created in the data catalogue. In this case, the only task is to select the license from the list when creating the dataset description.
2. *The dataset is associated with a license which is not already known* (blue arrows): The new license therefore needs to be reviewed and described within the data catalogue to be associated with the newly created description of the corresponding dataset. Describing the license corresponds to first including a descriptive text in the catalogue (which might be the whole text of the license) and second, choosing the set of rules that apply in each of the three taxonomies of permissions, prohibitions and duties. ODRL already includes some of the most common actions that can be associated with rules, and these are pre-loaded into the catalogue (e.g. permission to copy, prohibition to modify, duty to attribute, etc). Also, the data cataloguer is encouraged, as much as possible, to reuse existing rules from other licenses at the basis of the description, so that the work of creating new components is reduced, the taxonomies remain manageable and the license descriptions are consistent. It is expected however that some licenses will include rules that have not been encountered before, which will therefore need to be created (and described)

by the data cataloguer at the time of describing the license (using the simple mechanisms for editing provided by Wordpress).

3. *The dataset is not yet associated with a license* (red arrows): This is the most complex case, as the policy for inclusion in the Datahub will in this case need to be negotiated directly with the data owner. An interesting aspect of this however is that this negotiation can be done on the basis of existing licenses in the datahub, and reuse existing rules (permissions, prohibitions and duties) as a basis for creating a new policy specific to the negotiated Dataset, therefore increasing reuse and reducing the complexity of the negotiation step. It will also facilitate the review of the established license by the Legal Department of The Open University in the process of approving it.

It is important to notice here that, beyond the licenses and policies directly attached to the Dataset by/with the data owner, there are other aspects of the datasets that might require the creation of specific policies. These include:

– *When the data have human or animal sujects included:* An Ethics Committee will have to review not only the license, but also the data itself, to check that the data is collected, processed and re-distributed in a way that is compatible with the ethical standards in place.

– *When the data includes personal data:* The Data Protection Officer of the organisation will also have to review the license and dataset. In this case, they will not only be in charge of approving the inclusion of the dataset with the established license, but also to support the Data Governance Board in creating a privacy policy (in accordance with the Data Protection Act [13]) for the dataset. This privacy policy will be represented in the same way as the license and other policies associated with each dataset.

While this process appears very complex (and it is), it also demonstrates the value of the structured representation with clear semantics (based on ODRL) and of the availability of a simple tool to create, reuse and manipulate these structured representations. Indeed, here, it is expected that the set of licenses and taxonomies will grow progressivly as datasets are being included, creating a knowledge base of policies and rules becoming more and more useful in the description of new licenses, and therefore supporting scaling the data catalogue to thousands of datasets by reducing the effort required to review their policies.

3.3 Propagating Dataset Characteristics Accross Dataflows

The previous sections discuss the way to manage policies and rules that apply to the distribution and use of data as static objects. However, as described before, a key role of the MK Datahub is also to provide the basic services required to process such data, including integrating different sources with each other and analysing their content to create new information from several, raw datasets. Every processing step, manipulation and combination of data naturally has an effect on the policy rules in place: Combining datasets might require to combine rules in a non-trivial way; Anonymising datasets might reduce some of the permissions and prohibitions associated with the privacy policy and data protection; Modifying the data might nullify some of the permissions and contradict some of the prohibitions. Having a structured representation of the policy rules that apply to a dataset (in ODRL as described above) naturally helps with this issue. However, it also requires a semantic representation of the dataflows, i.e. a representation of what has happened, fundamentally, to data in such a dataflow, in order to understand how it might affect the permissions, prohibitions and duties that are attached to it.

There are several approaches to represent workflows on data. One of the most related to our issue here is Prov-O [9], the provenance ontology, which includes the representation of processes that might affect data beyond their original sources. However, while the process applied might, in itself, affect and/or propagate the policy rules attached to the dataset being processed, these processes only represent the mechanisms that implement a functional relationship between the dataset(s) in input and the dataset in output. While processes might be organised and structured according to a variety of dimensions, it is really these relationships that dictate the way the policy rules are affected and propagated from one (raw) dataset to another (processed) one. A meaningful representation of these relationships and of their taxonomic organisation is therefore what is needed to establish policy rule propagation mechanisms.

For this reason, and other similar use cases, we created the Datanode ontology [4][9], which can be described, in a nutshell, as an ontology of relationships between data artefacts. In this ontology, Datanode is a class of objects which is meant as an abstract class for different types and forms of data artefacts (datasets, databases, streams, feeds, etc). The rest of the ontology consists of the formal definition of a taxonomy of relationships between datanodes (as properties), including various aspects such as the description of datanodes (metadata), datanodes being derivations from others (e.g. from mining), datanodes overlaping in terms of their interpretations, capabilities, etc.

Through the Datanode Ontology, it is possible to represent, in a very simple and basic form, dataflows that focus on the fundamental relationships that exist between the origin datasets, intermediary ones and the final result. It is also possible to express rules for propagating or affecting policies and policy rules in these dataflows. For example, it is possible to express that a datanode that is an anonymised version of another one does not need to include in the attached policy the duty rules that are related to data protection. It can also express that duties such as attribution should be propagated to any dataset that are derived from any that included this policy rule originally, but not necessaraly through other relations. A prohibition to modify the dataset means that relations such as being a copy, a version, a description or overlaping with the considered one should not affect the policy, but any relation encapsulating a modification (derivation) would nullify any permission rule attached to the same policy. In

[9] `http://www.enridaga.net/datanode/0.3/docs/`

other words, relying on a semantic, structured representation of both the policies attached to datasets and of the dataflows in which they might be included allows us to reason upon the way policy rules propagate wihtin these dataflows.

4 Related Work

MK:Smart is certainly neither the first, nor the last smart city to be developed in Europe and the world. Many of these initiatives have very similar focuses, including the collection of large amounts of data from sources including sensors, local governements and companies. While the notion of a smart city goes beyond technology and include aspects of policy and governance which are certainly connected to the work presented here (see e.g. [12]), to the best of our knowledge, very little attention has been given so far to structuted approaches for the management of the very diverse licenses and policies that can be attached to datasets in such enviroments. Dublin City is an interesting example, benefiting from the work of IBM to develop the necessary technical infrastructures and tools (see e.g. [10]), and from a clear ambition from the local authorities to release data (see Dublinked[10]), especially as open data. Some of these technical architectures do actually mention aspects strongly related to the ones we focus on here, such as provenance (to track the sources of data, and the processes associated with them), privacy and general governance policies (see for example [11]). However, little details are actually available about the way in which these aspects are handled, and whether they benefit, like in our approach, from a structured, ontology-based representation of data policies and of their components. The result is that, while Dublin might be one of the most “data-centric” of smart city iniatitives, licenses and restrictions are still represented in the Dublinked portal, as in most other data catalogs out there, as basic, textual metadata fields.

While the issues described above, related to the management of highly diverse data sources in smart cities from the point of view of licenses and policies have surprisingly received little attention, the idea of employing a structured, machine readable representation of these policies is not new. While we use ODRL here, other rights representation languages exist, including for example CC-Rel (the Creative Commons Rights Expression Language[11]), focusing on open licenses. It is actually the case that such policies could be represented using more or less any kind of rule language, including for example AIR [8], which includes the representation of policies as a core use case. ODRL has the advantage to be dedicated to the explicit representation of digital rights, which are the most related to the variety of data licenses of interest here.

Naturally, there has been many other uses of ODRL, as an ontology to represent rights and usage policies, in different contexts [14], for the enforcement of access rights [6] or to check the compatibilities of the licenses associated with web data [16].

[10] `http://www.dublinked.com/`

[11] `https://wiki.creativecommons.org/CC_REL`

5 Conclusion

In this paper, we presented the approach taken by the MK:Smart project to deal with the diversity of the thousands of datasets to be included in the common data infrastructures of a smart city (the MK Datahub). We show in particular how relying on an ontology-based, semantic and structured representation of the description of each dataset, including the associated policies and the permission, prohibition and duty rules attached to them can help in managing and curating these diverse datasets in a formal and robust way, while reducing the effort required in keeping track of all the different situations.

While what we are proposing here is still preliminary, and will have to prove its value as the project evolves to include more and more datasets, it is engrained in the broader, timely challenges faced by many similar initiatives: Dealing with more and more data, from more and more sources, and with the difficulty of facing increasingly diverse situations regarding regulation, rights, terms and policies in data curation. In other terms, from the point of view of innovation, while this growth in the accessibility of data creates great new opportunities, it is also creating increasing complexity in the necessary process to curate these data. While currently these issues are too often overlooked, the area of smart cities will have to address them, making the need for structured, semantic representations more prominent not only for the data themselves, but also for all other aspects associated with them.

References

1. Keith Alexander and Michael Hausenblas. Describing linked datasets-on the design and usage of void, thevocabulary of interlinked datasets. In *In Linked Data on the Web Workshop (LDOW 09), in conjunction with 18th International World Wide Web Conference (WWW 09*. Citeseer, 2009.
2. Tuba Bakıcı, Esteve Almirall, and Jonathan Wareham. A smart city initiative: the case of barcelona. *Journal of the Knowledge Economy*, 4(2):135–148, 2013.
3. Lorrie Cranor. *Web privacy with P3P*. " O'Reilly Media, Inc.", 2002.
4. Enrico Daga, Mathieu d'Aquin, Aldo Gangemi, and Enrico Motta. From datasets to datanodes: An ontology pattern for networks of data artifacts. *Semantic Web Journal (submitted), http://semantic-web-journal.net/content/datasets-datanodes-ontology-pattern-networks-data-artifacts*, 2014.
5. Roberto García, Rosa Gil, Isabel Gallego, and Jaime Delgado. Formalising odrl semantics using web ontologies. In *Proc. 2nd Intl. ODRL Workshop*, pages 1–10, 2005.
6. Susanne Guth, Gustaf Neumann, and Mark Strembeck. Experiences with the enforcement of access rights extracted from odrl-based digital contracts. In *Proceedings of the 3rd ACM workshop on Digital rights management*, pages 90–102. ACM, 2003.
7. Renato Iannella. Open digital rights language (odrl) version 1.1. *W3c Note*, 2002.
8. Ankesh Khandelwal, Jie Bao, Lalana Kagal, Ian Jacobi, Li Ding, and James Hendler. Analyzing the air language: a semantic web (production) rule language. In *Web Reasoning and Rule Systems*, pages 58–72. Springer, 2010.

9. Timothy Lebo, Satya Sahoo, Deborah McGuinness, Khalid Belhajjame, James Cheney, David Corsar, Daniel Garijo, Stian Soiland-Reyes, Stephan Zednik, and Jun Zhao. Prov-o: The prov ontology. *W3C Recommendation, 30th April*, 2013.
10. Freddy Lecue, Spyros Kotoulas, and Pól Mac Aonghusa. Capturing the pulse of cities: A robust stream data reasoning approach. *Position paper, IBM Research, Smarter Cities Technology Centre, Dublin, Ireland*, 2011.
11. Vanessa Lopez, Spyros Kotoulas, Marco Luca Sbodio, Martin Stephenson, Raymond Lloyd, Aris Gkoulalas-Divanis, and Pol Mac Aonghusa. Queriocity: Accessing the information of a city. In *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.
12. Taewoo Nam and Theresa A Pardo. Conceptualizing smart city with dimensions of technology, people, and institutions. In *Proceedings of the 12th Annual International Digital Government Research Conference: Digital Government Innovation in Challenging Times*, pages 282–291. ACM, 2011.
13. British Parliament. Data protection act of 1998, 1998.
14. Carlos Serrão, Miguel Dias, and Jaime Delgado. Using odrl to express rights for different content usage scenarios. In *ODRL2005–2nd International ODRL Workshop*, pages 7–8, 2005.
15. Simone Tallevi-Diotallevi, Spyros Kotoulas, Luca Foschini, Freddy Lécué, and Antonio Corradi. Real-time urban monitoring in dublin using semantic and stream technologies. In *The Semantic Web–ISWC 2013*, pages 178–194. Springer, 2013.
16. Serena Villata and Fabien Gandon. Licenses compatibility and composition in the web of data. In *Proceedings of the Consuming Linked Data workshop, COLD*, 2012.
17. Stuart Weibel, John Kunze, Carl Lagoze, and Misha Wolf. Dublin core metadata for resource discovery. *Internet Engineering Task Force RFC*, 2413(222):132, 1998.