H2020-INSO-2014
INSO-1-2014 ICT-Enabled open government
YDS [645886] "Your Data Stories"

YourDataStories

D2.7 Data Management Plan v1.0

Project Reference No	645886 — YDS — H2020-INSO-2014-2015/H2020-INSO-2014
Deliverable	D2.7 Data Management Plan v1.0
Workpackage	WP2: Conceptual Architecture, User Needs Analysis and Design
Nature	Report
Dissemination Level	Public
Date	28/07/2015
Status	Final v1.0
Editor(s)	Bert Van Nuffelen (TF), Paul Massey (TF)
Contributor(s)	Niall Ó Brolcháin (NUIG), Michalis Vafopoulos (NCSR-D)
Reviewer(s)	Niall Ó Brolcháin (NUIG), Anna Triantafillou (ATC)
Document description	This report will describe how data will be managed, taking into account their characteristics. The final version of the document will form the YDS Data Management plan.

Document Revision History

Version	Date	Modifications Introduced	
		Modification Reason	Modified by
V0.01	03/05/2015	Toc	TF
V0.03	20/06/2015	Draft	TF
V0.06	15/07/2015	Align to template	TF, NCSR-D
V0.07	16/07/2015	Review	NUIG, ATC
V0.08	22/07/2015	Corrections	TF. NCSR-D
V0.09	24/07/2015	Final Corrections	TF
V1.00	28/07/2015	Final version for submission to EC	ATC

Executive Summary

The YDS data management plan describes the activities around the data that are processed in the YDS platform. It identifies the most important stakeholders involved in these activities. These stakeholders realize the YDS content. Key is the management of the associated meta-data of each dataset that is used as input or which is produced. That provenance data enables to make informed decisions w.r.t. to legal, access rights and usage conditions. By using DCAT-AP as meta-data vocabulary the interoperability with the European Open Data space is guaranteed.

This is the first version of the YDS data management plan. In the subsequent version the application of the data management plan to the pilot cases will precise the procedures. The YDS platform is a growing data eco system getting more source data as aggregated results and hence more data challenges will rise up and must be tackled adequately.

Table of Contents

List of Figures

List of Terms and Abbreviations

Abbreviation/Term	Definition
API	Application Programming Interface
CSV	Comma Separated Values
Data	The actual information
Data Catalogue	A catalogue consisting of dataset descriptions.
Dataset	A collection of data
Dataset catalogue record	A record in the catalogue describing the provenance of the dataset description
Dataset description	The meta data about a dataset.
DM	Data Management
DMP	Data Management Plan
ETL	Extraction, Translation/Transformation, Load
LOD	Linked Open Data
LOD2	EC funded R&D project (http://stack.lod2.eu/blog/)
Meta-data	Meta information about some entity. It typically describes identification information, usage conditions, publisher & ownership information.
OGD	Open Government Data
WP	Work Package
YDS	Your Data Stories

1 Introduction

1.1 Purpose and Scope

A Data Management Plan (DMP) is a formal project document which outlines the handling of the data sources at the different project stages[1]. The H2020 guidelines [22] provide an outline that has to be addressed. The DMP covers how data will be handled within a project frame, during the research and development phase but also details the intentions for the archiving and availability of the data once the project has been completed [5,8]. As the project evolves, the DMP will need to be updated to reflect the changes in the data situations and the understanding of data source becomes more concrete.

YDS as project aims to create a data eco system bringing together state of the art data processing technology with recent content about governmental budgetary and economical transparency in a platform that facilitates European citizens and in particular journalists creating stories based on factual data.

The technological foundations of the data management platform being established within YDS are such that it is intended to be multi-purpose and domain-agnostic. Within YDS this generic data management platform will be piloted using three closely related data domains: the financial transparency in the Greek and Irish governments and governmental development aid. This core activity of collecting and aggregating data from many different (external to the YDS project) data sources makes that meta data management of the used and produced datasets is key. By applying the DCAT-AP [16] standard for dataset descriptions and making these publicly available, the YDS DMP covers 4 out of the 5 key aspects (dataset reference name, dataset description, standards and metadata, data sharing) as specified in [22] as integral parts of the platform.

1.2 Approach for Work Package and Relation to other Work Packages and Deliverables

This deliverable is related to D3.1 “Data Source Assessment Methodology” since many of the questions identified here will need to be answered as part of the data source assessment (prior to trying to harvest the data source).

1.3 Methodology and Structure of the Deliverable

The initial version of the YDS DMP life-cycle is outlined in Section 2 which will elaborate the general conditions and data management methodology. As the YDS pilots will mostly handle manually created content (tables, reports, analysis …) the tooling will often require manual intervention and hence the complete data integration process from source discovery to published aggregated data cannot be completely automated. Therefore an important aspect of the YDS DMP is the general methodology. During the project progress the YDS DMP will be furthermore detailed by taking into account the experiences of the pilot cases. The remainder of this report is structured as follows:

- *Basic data information* - section 3 providing description of the basic information required for of the about the datasets that are going to be used in the YDS project.

[1] https://www.nsf.gov/eng/general/dmp.jsp – indicates it should be quite short (max 2 pages) but this would not cover the situation for the multiple data sources which would be typically used in linked data application domains.

- *Metadata Management* - Each data source and each resulting dataset of the YDS aggregation process will be described with meta-data. This meta-data can be used on the one hand for automating the YDS data ingestion process, but on the other hand also for external users to understand better the published data. This is further described in Section 4.
- *Access, sharing and re-use policies* - An important challenge in the YDS platform is the ambition to combine data from datasets having different usage and access policies. Interlinking data having payment requirements with data that is publicly and freely available impacts the technological and methodological approaches in order to implement the desired access policy. Section 0 outlines this further.

As the YDS pilots are still being defined, some questions relating to the Data management and storage (long term) are somewhat premature. Section 6 will, however, provide some direction in the sort of questions each data source and Pilot will need to answer.

2 The YDS data lifecycle

The YDS platform is a Linked Data platform; therefore the data ingested and managed by the YDS platform will follow the Linked Data life cycle [4]. The Linked Data life cycle describes the technical processing steps which are possible to create and manage a quality web of data. In order to smoothen the process best practices are described to guide data contributors in their usage of the YDS platform. This is further discussed in section 2.2 “The generic YDS data value chain”, while the common best practices [12, 13] are quoted in section 2.3 Best Practices.

Prior to the linked-data approach to the use of data, data management was perceived as being an act done by a single person or unit. Responsibilities (involving completeness, consistency, coverage, etc. of the data) were bound to the organizations duties. Today, with the used of the Internet and the distribution of data sources, this has changed: data management is seen as being a living service within a larger ecosystem with many stakeholders across internal and external organization borders. For instance, Accenture Technology Vision 2014 indicated this as the third most important trend in 2014 [9].

2.1 Stakeholders

For YDS, the key stakeholders have been identified which influence the data management. Their main interaction routes are depicted in Figure 1: DMP Role Interactions.

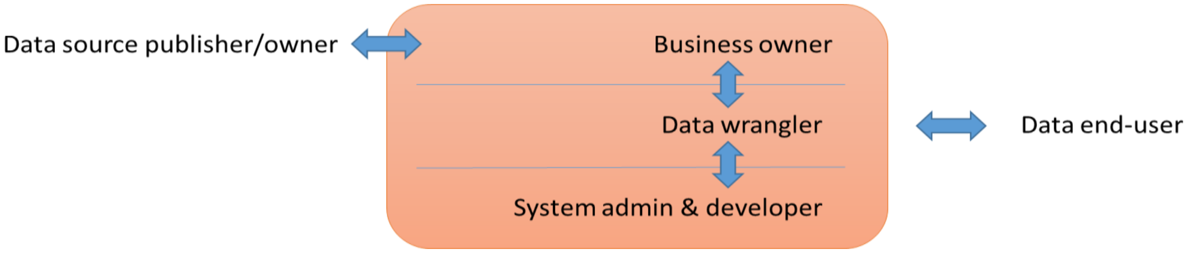

Figure 1: DMP Role Interactions

Data end-user(s):

The data end-users make use of the aggregated datasets to create their own story. In Deliverable D2.1 the main data end-user types for the YDS platform are identified: media & data journalists, auditors, web developers, and suppliers of business opportunities in public procurement, the civil society and public institutions. The data end-users are the main drivers of the YDS platform content: their need for data is the key driver for the content of the YDS platform.

Data source publisher/owner(s):

Represent the organization(s) which will provide the data to be integrated into the YDS platform. For many data sources, especially those that are published as Open Data by the public bodies, the interaction between YDS and the data source publisher/owner will be limited to technical access to the data (a download of a file, a registration to obtain an API key). To ensure a quality service level to the data end-users it is required to setup a more intense collaboration with the key data sources. This is, however, expected to happen only when the YDS platform matures.

Content business owner:

Is the person responsible for the content business objectives. The content business owner makes sure that the necessary data sources are found in a usable form and that the desired aggregations are being defined so as to realize the aggregated enriched content for the supported YDS stories. For each content domain a business owner is required.

Data wrangler [10,11]:

This person acts as a facilitator in that they interact at with all stakeholders but at the level of the integration of the source data into the platform. The data wrangler massages the data using the YDS platform to realize the desired content. They must understand both the business terminology used in the source data model(s) and the YDS target mode, understand the end user objectives and ensuring that the mapping between the models is semantically correct. The data wrangler is assisted by the YDS system administrator and YDS platform developers to tackle the technical challenges, but their central concern is the mapping of the data.

System administrator and platform developer:

Are responsible for the building and support of the YDS platform in a domain agnostic way.

2.2 The generic YDS data value chain

The complex process of a data value chain can be described using the following stages:

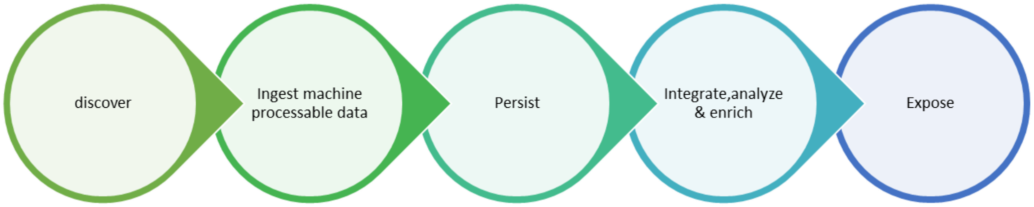

Figure 2: Data value chain stages

- **Discover**: In today's digitized world, there are many sources of data that help solve business problems that are both internal and external to organizations. Data sources need to be located and evaluated for cost, coverage, and quality. For YDS the evaluation of the data sources is part of the data source assessment methodology (See Deliverable XXX). The description and management of the resulting dataset meta-data is one of the main best practices used in the Linked Data community.
- **Ingest & making the data machine processable**: The ingest pipeline is fundamental to enabling the reliable operation of entire data platforms. There are diverse file formats and network connections to consider, as well as considerations around frequency and volume. In order to facilitate the value creation stage (Integrate, analyze & enrich) the data has to be turned into machine processable format. In the YDS case, this is RDF [1].
- **Persist**: Cost-effective distributed storage offers many options for persisting data. The choice of format or database technology is often influenced by the nature of other stages in the value chain, especially analysis.
- **Integrate, analyze & enrich**: Much of the value in data can be found from combining a variety of data sources to find new insights. Integration is a nontrivial step which requires domain knowledge and technical knowhow. Exactly by using a Linked Data approach with a shared

ontology the integration process is facilitated in YDS. Where-as the other stages have a high potential of automation to a level where humans are not anymore involved, this stage is driven by human interest in the data. New insights and better data interconnectivity are created and managed by a growing number of data analytical tools and platforms.

- **Expose**: The results of analytics and data that are exposed to the organization in a way that makes them useful for value creation represents the final step in deriving value from data.

The structure of the stages is based on the vision of the IBM Big Data & Analytics group on the data value chain [21].

When contributing a new data source to the YDS platform the stages are roughly followed from left to right. In practice the activities are, however, more distributed in order to keep the platform and the data it provides in the desired state. Indeed, data that is not actively nursed becomes quickly outdated. More and more imperfections will show up, to the point that data end-users consider the data not valuable anymore. Taking care of the YDS platform content, is hence a constant activity. From a technical perspective this work is supported by the tooling available during the Integrate, Analyse and Enrich phase. It is similar work as creating newly added value, but then with the objective to improve the overall data quality (coherency, completeness, etc.).

A further point to consider is that data based applications also have a tendency to generate new requirements based on insights which are gained when studying the data (this forms a loop which will continue as understanding of the data increases[2] and this is shown in Figure 3: Linked Data ETL Process). This will depend heavily on what the data is intended to allow or what it is intended to be used for (search for understanding, support of a particular story, tracking of an ongoing situation, etc.).

In the following sections, the above data value chain stages are made more concrete.

Discover

The **content business owners** are the main actors in this stage. Using the data source assessment methodology, relevant data sources for their content domain are being selected to be integrated.

An important outcome of the data source assessment is the creation of the meta-data description of the selected datasets. In section 4, the meta-data vocabulary that is going to be used is described (DCAT-AP). The expectation raised in creating the meta-data is that the data sources will be well described (what is the data, which are the usage conditions, what are the access rights, etc.), but experience has shown that collecting this information represents a non-trivial effort because it is often not directly available.

Ingest machine processable data

The selected datasets are being prepared *data technical* so that they can be ingested in the YDS platform. The data wrangler will hook up the right data input stream, for instance a static file, a data feed or an API, into the YDS platform. During this work the data is prepared for machine processing. Especially for static files such as CSV's often additional contextual information is required to be added

[2] GNUplot (http://www.manning.com/janert/) has a nice example of where the data once visualized, resulted in a correction of a false initial analysis of the data.

in order to make the semantics explicit. Without this preparation the conversion to RDF results in a technical reflection of the input, yielding more complex transformation rules in the Integrate, analyze and enrich stage.

Persist

Persistence of the data is de-facto an activity that happens throughout the whole data management process. However, when contributing a new data source to the platform, the first moment data persistence is explicitly handled is when the first steps have been taken to ingesting data into the YDS platform.

Since the YDS platform is about integrating, analyzing and enriching data from different sources *external* to the YDS partners, persistence of the source information is not only an internal activity. It requires interaction between the content business owner and the data source publisher/owner to guarantee that during the life time of the applications build on top of the data the source data stays available. Only carefully following up and the continuous interaction with the data source publishers/owners will create a trustable situation. Technically, this is reflected in the management of the source data meta-data activity.

Despite sufficient attention and follow up, it will occur that data sources become obsolete, are temporary not available (e.g. due maintenance) or completely disappear (e.g. the organization dissolves). Many of these cases are addressable to a certain extent by implementing data persistence strategies such as:

- *Keeping local copies*: explicit activity of copying data from one location to another. The most frequent case is copying the data from the governmental data portal to the YDS platform.
- *Caching*: a technical strategy which main intention is to enhance data locality so that the processing is smoother. It may also act as a cushion to reduce the effects of temporary data unavailability.

From the perspective of the YDS data user, *archiving & high available data storage* strategies are required to address the availability of the outcome of the YDS platform. This usually goes hand in hand with a related, yet orthogonal activity, namely the application of a dataset versioning strategy. Introducing dataset versioning provides clear boundaries were along data archiving has to be applied.

Integrate, analyze and enrich

In this stage, the actual value creation is done. The integration of data sources, their analysis and the analysis of the aggregated data and the overall content enrichment is realized by a wide variety of activities. In [4], the Linked Data life cycle is described: a comprehensive overview of all possible activities applicable to Linked Data. The Linked Data life cycle is shown in Figure 3: Linked Data ETL Process. (Note: Some activities of the Linked Data life cycle are also part of other phases like ingestion, persistence and expose.)

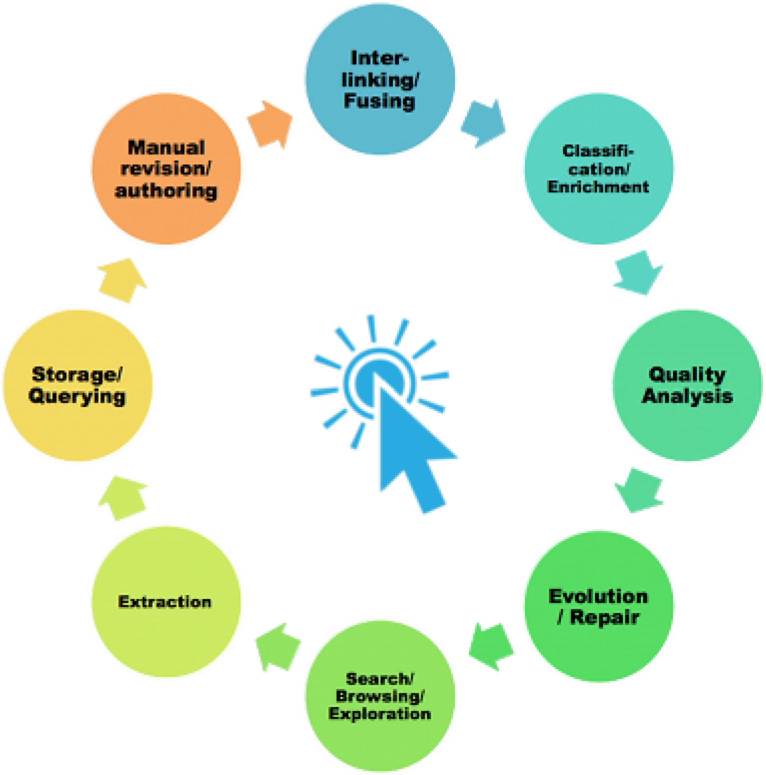
Inter-linking/ Fusing
Classifi-cation/ Enrichment
Quality Analysis
Evolution / Repair
Search/ Browsing/ Exploration
Extraction
Storage/ Querying
Manual revision/ authoring

Figure 3: Linked Data ETL Process

Start reading from the left bottom stage called “Extraction” and going clock-wise.

As most data is not natively available as RDF extraction tooling will provide the necessary means to turn other formats into RDF. The resulting RDF is then stored in an RDF storage system, available to be queried using SPARQL. Native RDF authoring tools and Semantic Wiki’s allow then the data to be manually updated to adjust to the desired situation. The interlinking and data fusion tools are unique tools in the world of data management: Linked Data (or a data format with similar capabilities as RDF) are the enablers of this process in which data elements are interlinked with each other without losing their own identity. It is the interlinking and the ability of using entities from other public Linked Data sources that creates the web of data. The web of data is a distributed knowledge graphs across organizations which is in contrast to the setup of a large data warehouses. The following 3 stages are about further improving the data: when data is interlinked with other external sources new knowledge can be derived and thus new enrichments may appear. Data is off-course not a solid entity but it evolves over time: therefore quality control and evolution is monitored. To conclude the tour the data is published. RDF is primarily a data publication format. This is indicated by the vast amount of tooling that provides the search, browsing and exploration of Linked Data.

Expose

The last stage is about the interaction with the YDS data users. The YDS platform is a Linked Data platform, and hence the outcome of the data integration, analyzes and enrichments will be made available according to the common practices for Linked Open Data:

- A meta-data description about the exposed datasets
- A SPARQL endpoint containing the meta-data

- A SPARQL endpoint containing the resulting datasets
- A public Linked Data interface for those entities which are dereferenceable.

Additionally the YDS platform supports dedicated public API interfaces to support application development (such as visualizations). The specifications of these are to be defined.

2.3 Best Practices

The YDS platform is a Linked Data platform and in this section, the relevant best practices for publishing Linked Data are described [12, 13]. The 10 steps described in [13] are an alternative formulation of these stages in the context of publishing a standalone dataset. Nevertheless, these steps formulate major actions in the creation of Linked Data content for the YDS platform concisely (and that is why they are quoted here):

1. *STEP #1 PREPARE STAKEHOLDERS:*
   Prepare stakeholders by explaining the process of creating and maintaining Linked Open Data.
2. *STEP #2 SELECT A DATASET:*
   Select a dataset that provides benefit to others for reuse.
3. *STEP #3 MODEL THE DATA:*
   Modeling Linked Data involves representing data objects and how they are related in an application-independent way.
4. *STEP #4 SPECIFY AN APPROPRIATE LICENSE:*
   Specify an appropriate open data license. Data reuse is more likely to occur when there is a clear statement about the origin, ownership and terms related to the use of the published data.
5. *STEP #5 GOOD URIs FOR LINKED DATA:*
   The core of Linked Data is a well-considered URI naming strategy and implementation plan, based on HTTP URIs. Consideration for naming objects, multilingual support, data change over time and persistence strategy are the building blocks for useful Linked Data.
6. *STEP #6 USE STANDARD VOCABULARIES:*
   Describe objects with previously defined vocabularies whenever possible. Extend standard vocabularies where necessary, and create vocabularies (only when required) that follow best practices whenever possible.
7. *STEP #7 CONVERT DATA:*
   Convert data to a Linked Data representation. This is typically done by script or other automated processes.
8. *STEP #8 PROVIDE MACHINE ACCESS TO DATA:*
   Provide various ways for search engines and other automated processes to access data using standard Web mechanisms.
9. *STEP #9 ANNOUNCE NEW DATA SETS:*
   Remember to announce new data sets on an authoritative domain. Importantly, remember that as a Linked Open Data publisher, an implicit social contract is in effect.
10. *STEP #10 RECOGNIZE THE SOCIAL CONTRACT:*
    Recognize your responsibility in maintaining data once it is published. Ensure that the dataset(s) remain available where your organization says it will be and is maintained over time.

3 Basic data Information

Each YDS pilot handles content within the Linked Open Economy domain. The following information will need to be recorded by the ***content business owner*** of each pilot. These questions, similar to that found in [5] will provide the starting point for using the data sources. The aim being to find any data usage issues, earlier rather than later[3]. This basic data information, information about the data or meta-data will require managing and will be further discussed in section 4.

3.1 Data source acquisition

- How will the data be acquired?
- When and where will the data be acquired?
- What are the licenses required to access and used the data (See Section ⍰)?
- If existing data is to be used, what are their origins (provenance information)?
- What documentation is available for the data source models, etc.?
- For how long will the data be available?
- What are the contact points for accessing the data?
- Is there help available for the data-wrangler in understanding the available data?

3.2 Data harvesting and collection

- How will the data collected be combined with existing data?
- What is the relationship between the data collected and existing data?
- What are the tools and/or software that will be used?
- How will the data collection procedures/harvesting be documented?

3.3 Post collection data processes

- How is the data to be processed?
- Basic information about software used,
- Are there any significant algorithms or data transformations used (or to be used)?

3.4 Data formats

- Describe the file formats that will be used, justify those formats,
- Describe the naming conventions used to identify the files (persistent, date based, etc.)

3.5 Data quality assurance

- Identify the quality assurance & quality control measures that will be taken during sample collection, analysis, and processing[4],
- What will be the data validation requirements? Are there any already in place?

[3] For example, a major impact on the cost of harvesting the data is the format of the data. If the data is only available in a non-machine readable unstructured format such as PDF or Word, then the pilot could become infeasible.

[4] relates to WP3 Task 3.1 but basic information will also be provided here.

3.6 Short term DM

- How will the data be managed in the short-term? Consider the following:
    - Version control for files,
    - Backing up data,
    - Security & protection of data and data products,
    - Who will be responsible for management (Data ownership)?

3.7 Long term DM

- See Section 6 for more details

4 Meta-data management

The data collected and aggregated in the YDS platform can also be distributed to the public or be used in another aggregation process. A coherent set of data is called a dataset. Distributing the dataset requires describing the dataset using meta-data properties.

Within Europe an application profile of the W3C standard DCAT [15] called DCAT-AP [16] is being used to manage data catalogues. With this standard dataset descriptions in Europe can be exchanged in a coherent and harmonized context. At the moment of writing, i.e. June 2015, DCAT-AP is undergoing a revision to better fit the European needs.

In addition to this motivation, YDS has extensive in-house knowledge and experience: the YDS partners, NUIG and TenForce are organizations that played key roles in the establishing and success of the standards. NUIG actively supported the creation of DCAT as being the co-editor of the standardization process and it has continued sharing its expertise in the development of the DCAT application profile. TenForce, lead and was/is participating in several projects that contributed to the technological application of the standard DCAT and the creation of DCAT-AP: LOD2, the European Open Data Portal, Open Data Support (in which TenForce has established the first implementation of DCAT-AP). Recently TenForce supported the revision of the DCAT-AP process and is it responsible for the first study on creating a variant for statistical data STAT DCAT-AP.

Building upon DCAT-AP will integrate the YDS platform in the European (Open) Data Portal ecosystem. Data being made available through the YDS platform is being picked up and distributed to the whole of Europe. On the other hand the European (Open) Data Portal ecosystem can provide access to data that has not yet being identified as relevant. For instance the Open Data Support project data catalogue [17] offers access to more than 80000 dataset descriptions of more than 15 European Union member states.

The core entities are Dataset and Distribution. The Dataset describes the data and its usage conditions. Each Dataset has one or more Distributions, the actual physical forms of the Dataset. A collection of Datasets is managed by a Data Catalogue. The details are shown in Figure 4: DCAT-AP Overview.

As the DCAT-AP vocabulary is a Linked Data vocabulary, it fits naturally the technological choices of the YDS platform. It is expected that the DCAT-AP vocabulary covers the majority of the YDS data cataloguing needs. In case of gaps or more specific needs, the YDS platform will further enhance and detail the DCAT-AP vocabulary to fit its needs. One such aspect that requires further elaboration is the management of licensing, rights and payments. In the ongoing revision of DCAT-AP some additional properties are being added covering these aspects, but it has to be expected that those are not sufficient for YDS.

The adoption of DCAT-AP creates also the availability of tooling. There is EDCAT [18] and API layer to manage data catalogues, a web interface [19] and the ODIP platform [17] that harvests open data portals (based on an earlier version of UnifiedViews [20], the central component of the YDS platform).

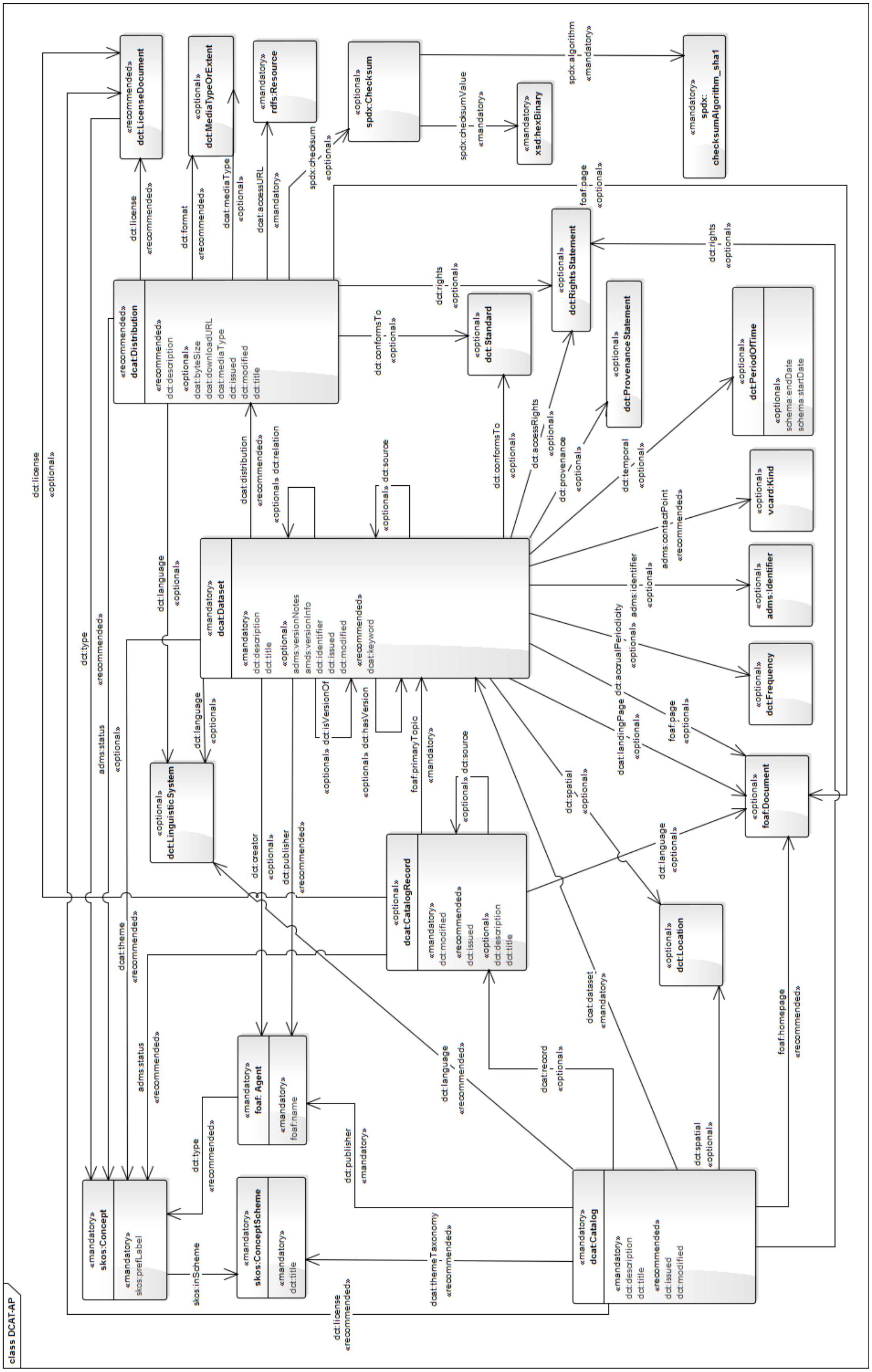

Figure 4: DCAT-AP Overview

5 Access, sharing and re-use policies

For a data platform, such as YDS, the access, usage and dissemination conditions of the used source data determine the possible access, usage and dissemination conditions of the newly created aggregated data. Despite the sizeable amount of public open data that is available and that will be imported, it is likely to occur that there will be source data which is subject to restrictions. When combining open data with restricted data, it cannot be taken for granted that the resulting new data is open (or restricted). In such mixed licensing situations, decisions will need to be made by the content business owner and the data source owners concerning the accessibility of the merged data. For example, it may be decided that some aggregated data is only accessible for a selected audience (subscription based, registration based, payment required or not …).

This context poses not only a business challenge, but also a technological challenge. Some common practices when moving data from one source to another may not be acceptable anymore. For example: if one data source A describes the overall spending of a government by project and another data source B describes the governmental projects and their contractors. The aggregated data A+B provides thus insight in how the budget was spend by the contractors. Merging the data into one aggregation usually makes it impossible to determine from where the individual data elements came from. This is not problematic when the aggregated data is subject to the same or more restrictive access, usage and dissemination conditions as the source data themselves.

More complex and problematic is the situation where the aggregations are being distributed throughout channels to audiences that do not satisfy the conditions stipulated by one of the sources. To prevent incorrect usage, managing the access, usage and dissemination conditions of the newly created aggregations is important. That information will form the cornerstone of the correct implementation of the required access, usage and dissemination policies.

As shown above this aspect of the data management is a non-trivial work. Today it is part of ongoing discussions. See the outcomes of the LAPSI project [14]. Therefore YDS will apply the following strategy:

- The content business owner ensures that for each data source the access, sharing and reuse policy information is known.
- The content business owner decides whether the outcome of the integration & aggregation process is open (in all meanings = public, reusable, free of charge) or non-public (some restrictions apply).
- The data wranglers and system developers setup a data aggregation flow and data publication exposure according to the specification by the content business owner.
- The dataset meta-data of the created outcome is always public. This ensures transparency of the knowledge that is gathered within the YDS platform. The openness of the meta-data repository yields transparency.

The openness of the meta-data repository may conflict (see [6]) with the notion of "protection of sources" (See [7]), the right that is granted to journalists to keep their sources anonymous. With a centralistic approach this dilemma is non-trivial. A distributed approach such as that depicted in Figure 5: Data Accessibility shows, however, a possible resolution. The public open instance of the YDS platform will publish the public data, a local instance at the journalist's office will use the data from the public instance as one of the data sources. The journalists can then augment the public data with confidential data within their safe environment. The collected insights can then be turned into a story, ready to be published.

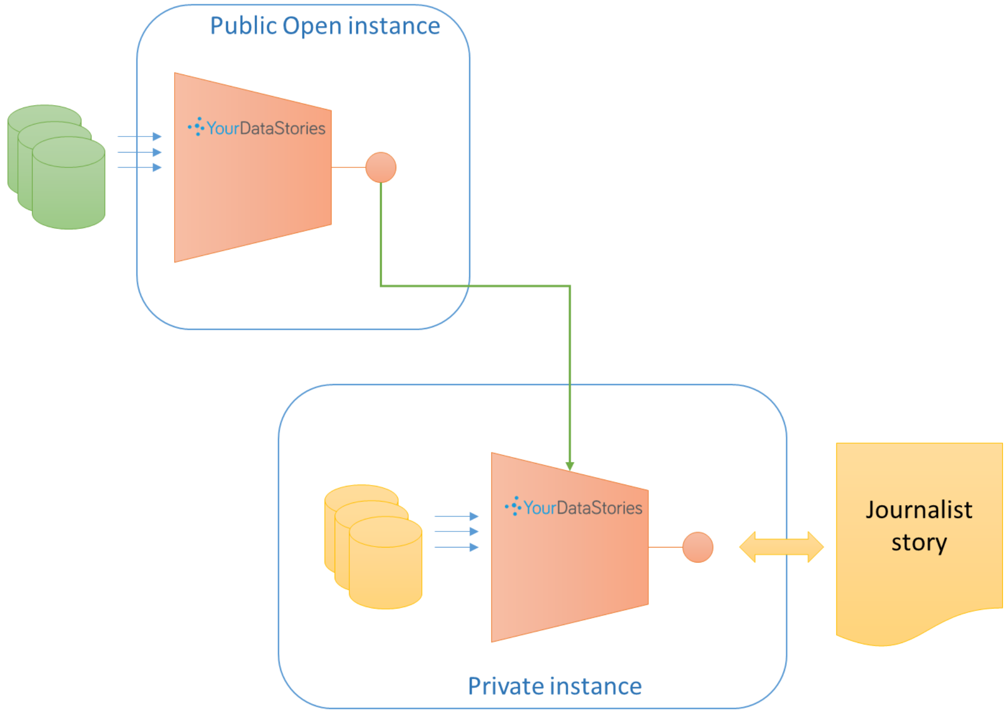

Figure 5: Data Accessibility

The technological foundations of the YDS platform, i.e. Linked Data, ensure that the above scenario is supported out of the box without any additional work.

As the above situations already indicate, the situations that might occur may be very complex. Therefore YDS will start with a simpler more uniform initial setup of only open data that is free for reuse. Since the YDS specify to create for each dataset a DCAT-AP entry, the base usage conditions get registered. It will enable to identify a complex situation of which some are sketched above. The effect and decisions to resolve the case will be recorded and added as notes to the relevant DCAT-AP entries. In doing this, the DCAT-AP record for a dataset becomes the key reference point of the dataset decision making.

6 Long term data management and storage

The questions to be addressed concerning long-term storage are not new: environmental datasets, medical testing datasets, component test results relating to safety will all have to be stored for a long time (the definition of long-term being defined as part of a legal requirement, others will simply be seen as being expected, e.g. datasets relating to academic published results). These issues are complicated for when the data is made available over the internet, in that the data could be merged with other data coming from other sources, so the definition of meaningful long-term becomes problematic. So, each content business owner needs to consider:

- What is the volume of the data to be maintained?
- What is considered long-term (2-3 years, 10 years, etc.)?
- Identification of archive for long-term preservation of YDS data.
- Which datasets will need to be preserved in the archive?
- What about relevant dependent datasets? Snapshots of external datasets?
- Preserved datasets will need to be updated and this means a data preservation policy and process will need to be defined (and operational).

A central consideration for any long-term DMP is the cost of preserving that data and what will happen after the completion of the project? Preservation costs may be considerable depending on the exploitation of the project after its finalization. Examples include:

- Personnel time for data preparation, management, documentation, and preservation,
- Hardware and/or software needed for data management, backing up, security, documentation, and preservation,
- Costs associated with submitting the data to an archive,
- Costs of maintaining the physical backup copies (disks age and need to be replaced).

7 Conclusions

Applying & setting up a data management platform requires not only the selection of the right technological components but also the application of a number of best practice data management guidelines [12, 13] and given in Section 2.3. Those best practices guide the users to the best ways to the creation of data better ready to become a sustainable data source in the web of data. Two of these best practices have led to a concentration on two focal areas that require initial attention for the YDS data stories. These initial focus points being:

- Dataset meta-data management both for both the sources and the published datasets, and
- Data access considerations, sharing possibilities and re-use policies and licenses.

In all this, the DCAT-AP dataset descriptions are a key requirement. Having the dataset descriptions in machine readable format creates potential on effective traceability, status monitoring and sharing with the YDS target audiences. Each DCAT-AP entry will act as the individual DMP for the dataset it describes.

The high level principles of the YDS project DMP have been presented from data source discovery up to publishing of the aggregated content. The best practices for publishing Linked Data – which is followed by YDS – describe a data management plan for publication and use of high quality data published by governments around the world using Linked Data. Via these best practices the experiences of the Linked Open Data community are taken into account in the project.

The technological foundations of the YDS platform separate very cleanly data semantics, data representation and software development. Linked Data makes the platform more flexible to implement at a later point in time the technological and data provenance support which is required by the pilots as basic support. This ability is unique in the data management technology space. Here and there throughout the report some tooling is mentioned, but it has to be noted that the actual software is irrelevant for the discussion in this report.

Given the current initial status of the YDS pilots and the fact that for each pilot the more concrete DMP will be different (because of the data source types, the access licenses, etc.) more detailed & precise guidelines will require further analysis of the common situations as they are identified. This will be on-going work which will initially be on a case by case basis, which will be combined into a YDS DMP best practices guide for the various pilots.

8 References

[1] Resource Description Format, http://www.w3.org/RDF/ &
http://www.w3.org/standards/techs/rdf#w3c_all

[2] IEEE Standard 802.11-1997, Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications, 26 June 1997.

[3] P.Brenner, A technical tutorial on the IEEE 802.11 protocol, Breezecom Wireless Communications edition, 1997.

[4] LOD2 Book Reference - the Linked Data Stack [http://stack.linkeddata.org], LOD2 book.

[5] UK DMP checklist [http://ukdataservice.ac.uk/manage-data/plan/checklist.aspx]

[6] Ethics in data journalism: protection of sources, leaks and war, XX, [http://onlinejournalismblog.com/2013/09/19/ethics-in-data-journalism-protection-of-sources-leaks-and-war/]

[7] Protection of Sources, https://en.wikipedia.org/wiki/Protection_of_sources

[8] Data Management Plan, http://en.wikipedia.org/wiki/Data_management_plan

[9] Accenture Technology Vision, 2014 [http://www.accenture.com/microsites/it-technology-trends-2014/Pages/data-supply-chain.aspx]

[10] Data wrangling [https://en.wikipedia.org/wiki/Data_wrangling]

[11] Data wrangling – Job description [http://blog.okfn.org/2015/03/19/were-hiring-at-open-knowledge-project-managers-developers-and-data-wranglers/]

[12] [https://dvcs.w3.org/hg/gld/raw-file/default/bp/index.html]

[13] [http://w3c.github.io/dwbp/bp.html#bp-summary]

[14] European Thematic Network on Legal Aspects of Public Sector Information [http://www.lapsi-project.eu/]

[14a] [http://www.slideshare.net/OpenDataSupport/licence-your-data-metadata]

[15] DCAT Vocabulary [http://www.w3.org/TR/vocab-dcat/]

[16] DCAT-AP [https://joinup.ec.europa.eu/asset/dcat_application_profile/description]

[17] Open Data Support [http://data.opendatasupport.eu]

[18] EDCAT, [http://edcat.tenforce.com]

[19] DCAT Editor - web interface [https://dcat-editor.com/manager/#view=start]

[20] UnifiedViews [http://www.unifiedviews.eu/]

[21] Understanding the Data Value Chain, http://www.ibmbigdatahub.com/blog/understanding-data-value-chain

[22] http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf