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


# D3.2 Data Source Assessment Methodology v2.0

| | |
|---|---|
| **Project Reference No** | 645886 — YDS — H2020-INSO-2014-2015/H2020-INSO-2014 |
| **Deliverable** | D3.2 Data Source Assessment Methodology v2.0 |
| **Workpackage** | WP3: Data Layer |
| **Nature** | Report |
| **Dissemination Level** | Public |
| **Date** | 28/07/2016 |
| **Status** | Final v1.0 |
| **Editor(s)** | Uroš Milošević (TF), Bert Van Nuffelen (TF), Paul Massey (TF) |
| **Contributor(s)** | - |
| **Reviewer(s)** | Michalis Vafopoulos (NCSR-D), Niall O'Brolchain (NUIG) |
| **Document description** | This report represents the second version of the methodology for the assessment of data source quality and eligibility for use, under YDS requirements |

## Document Revision History

| Version | Date | Modifications Introduced | |
|---|---|---|---|
| | | Modification Reason | Modified by |
| **V0.1** | 01/06/2016 | ToC | TF |
| **V0.5** | 12/07/2016 | Updated methodology v1.0 | TF |
| **V0.6** | 22/07/2016 | Ready for review | TF |
| **V0.7** | 25/07/2016 | Peer review | NCSR-D |
| **V0.8** | 26/07/2016 | Integrated feedback | TF |
| **V0.85** | 26/07/2016 | Peer review | NUIG |
| **V0.9** | 27/07/2016 | Integrated feedback | TF |
| **V0.10** | 27/07/2016 | Final version | TF |
| **V1.0** | 28/07/2016 | Final version for submission to EC | ATC |

## Executive Summary

This document constitutes the second version of the YDS Data Source Assessment Methodology, taking into account the experiences and lessons learned since month six of the project. The data source assessment provides the methodological context in which the data source is being assessed. The presented flow covers the data source assessment from the moment there is a need identified for the introduction of a kind of data until the moment that the actual integration starts. It is a collaborative process which results in a technical motivation to how to add the data (the data harvesting assessment) and a roadmap towards the actual harvesting (the data source harvesting plan). This version builds on top of its predecessor by (1) elaborating further on the assessment criteria, (2) providing an overview of the data sources assessed and harvested since the publication of the first version of the methodology, (3) reflecting on experiences with respect to the assessed (and harvested) data sources so far, (4) providing an example of desired DCAT-AP output, as a direct outcome of assessment efforts, and (5) clarifying the goals of data source assessment in year two of the project, by shifting the focus onto data quality (i.e. added value), instead of quantity.

# Table of Contents

## List of Figures



## List of Tables



## List of Terms and Abbreviations

| Abbreviation | Definition |
|---|---|
| CSV | Comma Separated Values |
| DCAT | Data Catalog Vocabulary |
| DCAT-AP | DCAT Application Profile |
| DMP | Data Management Plan |
| DPU | Data Processing Unit |
| DSAM | Data Source Assessment Methodology |
| ETL | Extraction Translation Load |
| FTP | File Transfer Protocol |
| GZIP | GNU ZIP – compression format |
| JSON | JavaScript Object Notation |
| LOD | Linked Open Data |
| LOD2 | EC funded R&D project (http://stack.lod2.eu/blog/) |
| OGD | Open Government Data |
| PBO | Pilot Business Owner |
| PCO | Pilot Content Owner |
| PDW | Pilot Data Wrangler |
| QA | Quality Assessment |
| RDF | Resource Description Format |
| REST | Representational state transfer |
| SPARQL | SPARQL Protocol and Query Language |
| SVG | Simple Vector Graphics |
| URL | Uniform Resource Locator |
| WP | Work Package |
| XML | eXtensible Markup Language |
| YDS | Your Data Stories |

# 1 Introduction

## 1.1 Purpose and Scope

This Deliverable, D3.2, describes the second version of the methodology for assessing the usability or quality of YDS project data sources. This usability or quality assessment (QA) can be seen from different perspectives; including legal and applicability aspects, business added value, technical harvesting feasibility, quality requirements and so on. The aim of the methodology is still to help in resolving the trade-off in issues when choosing which of the data sources will best fit the end user application. Having said that, this update does not aim to discuss an entirely new approach to assessing potential data sources within the YDS framework, but to reflect on the experiences and lessons learned since the publication of the previous version, and integrate the knowledge acquired thus far.

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

The first version of the methodology, D3.1 [1], laid down the foundation for identifying viable data source and appropriate data harvesting plans. The updated version heavily relies on its predecessor, along with the information provided in D2.7 Data Management Plan [2], which required some of the same questions to be answered and still has a strong relationship to the methodology addressed in this deliverable. Additionally, D2.3 [6] describes the user requirements which impact the criteria which should be used in the data source assessments.

## 1.3 Methodology and Structure of the Deliverable

This is the second version of the data source assessment methodology and, as such, focusses on a refined version of the approach presented in D3.1. It is still a multi-step approach, which can be stopped as soon as a data source is deemed by currently unsuitable and which should result in a plan to harvest the data. The main assumption is that understanding the end-user needs for the data will drive the application and therefore the data sources which should be included. There are three YDS pilots being continuously developed, but the data source assessment methodology should not be only applicable to those pilots but generic to many application domains[1].

Once the potential data sources and datasets have been identified, assessing the quality becomes possible and this report looks at some of the initial questions which will need to be answered concerning the data source (but also provides the methodology for the auxiliary document needed around the data source).

[1] In fact, *pilot* should be considered in the broadest sense possible, and the methodology is not intended to be restricted to only the YDS pilot applications.

Section 2 provides a description of the assessment criteria or basic questions to be asked. These questions concern:

- What are the most usual data quality problems, in the Web of Data?
- How are these problems related with the assessment of dataset sources in YDS?
- How can we formulate data quality dimensions and measures to assess the quality of sources within the YDS project?
- What kind of tools are there for data quality assessment?
- How can the above tools relate or interoperate with the YDS harvesters?

Following that, in Section 3, an overview is given of the decision tree for each of the data sources and the roles which are expected to be involved in this decision process. Section 4 gives an overview of some of the initial assessments and the experiences since the publication of the first version of the methodology. Section 5 briefly discusses the status of the YDS pilot data assessments, and Section 6 concludes the deliverable.

## 1.4 Updates in the second version

We further strengthen our approach in this document by looking back at the experiences gained from our actual (detailed) assessment and harvesting efforts (D3.6 Data Harvesters [11]), based on the plans that followed from D3.1. More specifically, this deliverable builds on D3.1 by:

- Elaborating further on the assessment criteria,
- Providing an overview of the data sources assessed and harvested since the publication of the first version of the methodology,
- Reflecting on experiences with respect to the assessed (and harvested) data sources so far,
- Providing an example of desired DCAT-AP output, as a direct outcome of assessment efforts,
- Clarifying the goals of data source assessment in year two of the project, by shifting the focus onto data quality (i.e. added value), instead of quantity.

The above updates are integrated in the existing methodology so as to make a whole (not disrupting any existing/ongoing assessment), and the improved methodology backward compatible (i.e. making any re-evaluation of a data source result only in new, and not conflicting information).

# 2 Data Source Assessment Criteria

D2.7 [2] (Section 3) and D2.3 [6] "User Requirements" indicate data related questions which have to be answered by the Pilot Business Owner (PBO). Critical to assessing a data source and harvesting the associated datasets it is necessary to first find a potentially interesting data source[2]. At present, finding a relevant accessible dataset is often challenging, requiring lots of searching and hunting (even with data portals becoming available). Not all datasets are advertised as being available (for numerous reasons – Where to advertise? Who should be responsible for advertising that a dataset is available, etc.? How often should the dataset availability be advertised? Under what keywords/terms, etc. should it be advertised?). This is the initial work of the PBO (D2.7) who would need to define the application or pilot objectives along with the data source owners. The initial data focused questions concern such things as:

- Access points for retrieving the data,
- Machine Readability,
- Data source complexity,
- Durability of the data-source,
- Licensing,
- Contact points,
- Etc.

All these points are ones which need to be addressed when deciding on the "business value[3]" of the data-source. To aid in such an assessment, when there are a large number of data sources, it might be necessary for a Content Business Owner (CBO) to try to objectify the decision process and this point is addressed in Section 2.3, but the one addressed in the next section is the main question.

## 2.1 Machine Readability and Harvestability

This is single biggest issue for Linked Data and YDS applications – is the data in a format suitable for machine processing or not? If the format is not suitable it must be manually converted into a usable format and this will considerably slow down the harvesting processes (as well as vastly increasing the project costs). Data can be exchanged in many different formats between humans, but this is not the case when machines have to do the conversions. This section gives some general rules which can be used to determine how suitable it is for machine processing:

- Non-structured and other binary formats such as *PDF*, *word*, images, etc. are *not* suitable for machine processing and are not considered valid input formats for data harvesting,

[2] If there are no data sources, there is no application and so nothing of interest to present to an end user. The YDS pilots have in D2.3 identified a number of potential data sources which needed to be further assessed. The list was refined in the second half of year one of the project, with D3.6 providing a clearer picture of the data sources tackled thus far.

[3] Business value is used in the loose manner here, in that the end-user has to get something out of their use of the data source or associated application (if the end-user application is around economic targets, a data source around science fiction films, football or cartoons has zero value).

- Scrapping web-sites, or HTML input, for the information while sometimes possible is often expensive and error prone because
  - Web-sites are normally defined like PDF/Word documents for human processing rather than for harvesting (i.e. the concentration is on readability, eye-catching style/form, semiotics[4] inspired information presentation, etc.),
  - The look & feel of web-sites is often changed at no notice resulting in unexpected costs and risks to the harvesting operation.
- Excel (.xls) files are a very common way of exchanging tabular information, where:
  - If it is equal to CSV is may be considered as an acceptable input format,
  - If it contains macros and cell-references it is not acceptable input (because the harvester would not see the macros and cell-references as data).
- Desirable input formats are:
  - RDF,
  - CSV[5],
  - XML,
  - JSON[6].

YDS is a Linked Data project [2, Section 2], so the main issue is how to get the input data into the base RDF format (or the Extraction phase of the ETL processing using a tool such as UnifiedViews [4]). Essentially, this is the main question for all the YDS data sources – can an extractor be developed to make the translation and loading (into Virtuoso) possible?

In addition to the machine readable format question, there is the question of whether the contents of the data, even when in a suitable machine processing format, can be extracted without information loss (e.g. SVG would be machine readable, but what could be harvested from it?). In principle, any structured format can technically be converted to RDF using a simple naive conversion process. However, smarter conversion approaches usually mean less work in the data integration stage. This can include:

- Is pre-processing of the original source data required to obtain better quality ingestion?
- Which data elements are identifiers of concepts?
- Which data elements satisfy a data type?
- How many exceptions are there to the base conversion rule?

[4] Semiotics is the study of signs/symbols – many aspects of which are present on typical web-sites. E.g. green/red dots to indicate positive points and danger signals, ticks and crosses indicating the same, indications of a required top to bottom or left to right reading of the text, contextual ordering of items, etc. On websites, it further relates to the representation of information via links, buttons, scripts, etc.

[5] The more the complex the CSV, the more difficult it will be to define the harvesting rules (and to make sure that the subsequent harvesting attempts are error free if the data is updated).

[6] Such as that recovered from the CKAN API are acceptable since, once the convertor is defined, it can be reused on another CKAN based site (assuming the same API version).

This last point is crucial to understand, since values entered through forms in applications can have local usage defaults, but which are not documented anywhere in the model but will be visible in the data itself (e.g. height/weight values in a specific combination means something in a given context or operational unit – identifying these can mean a lot of data forensics).

## 2.2 Characteristics of Data Quality

In this section, a number of quality aspects are briefly described. These need to be considered for each of the data sources, but the relevance of the criteria depends on the individual business objectives[7]. Many of the questions result in capturing meta-data which is recorded in the DCAT-AP entries for each dataset. A sample DCAT-AP entry is provided in the Annex of this deliverable. Some of the quality criteria need to be clarified and contextualized for the intended end-user domain,since higher or more verified quality normally increases costs and slows down the system updates because of the additional checks. Some aspects are not only relevant for the pre-harvesting assessment, but continuously re-evaluated throughout the lifetime of the project. New insights result in improvements of the harvesting approaches and overall data quality.

### 2.2.1 Contextual dimensions

#### *2.2.1.1 Dataset completeness/coverage*

Completeness does not mean volume, but that within the defined range there are less gaps or holes in the data available. This refers to both overall dataset coverage (i.e. missing rows) or just individual entries (i.e. missing values).

#### *2.2.1.2 Amount-of-data*

This leaves an impact on the required processing capacity of the YDS platform and on the timeliness of the extraction and processing of the data. In year one, this was one of the key dimensions, as the goal was to establish a robust basis for the different pilot cases/

#### *2.2.1.3 Dataset relevancy*

This depends on the end-user application needs and is different for each of the pilots but it usually includes such additional things as volume, coverage of the domain, keywords, etc. It is often the starting point for any data quality related assessment. In year two, however, it is given a special emphasis, as it is combined with other dimensions to identify the added value of a given data source.

### 2.2.2 Trust dimensions

In order to have confidence when using a data source, there are elements of trust which need to be taken into account when assessing a data source. The expectations on the trust-ability of the dataset depend heavily on the business objectives and the assurances required by the end-users that the conclusions they will derive from the data sources can be relied on. This sections looks at the main ways these questions can be answered.

[7] Data suitable for one application area will not be suitable for another area.

#### *2.2.2.1 Dataset provenance*

It is important to know where the data has come from, how it was produced, manipulated and which organizations are responsible for it. Tables of data, where the method of accumulating the data changes in the middle are expected to have some indication of this change, thresholds for table containing values need to be indicated (i.e. this is the case in national statistics, for example, to prevent specific populations/people being referred to). DCAT-AP offers provenance potential at the granularity of the dataset and its distributions. If more detailed provenance is required, in particular at the data points themselves, the PROV-O vocabulary [5] can be considered.

#### *2.2.2.2 Dataset verifiability*

This is a part of the provenance issue where the origins of the data and the manipulations which the data has undergone are expected to be documented. Ultimately, there would be a full trace to say where each value came from. I.e. in case of aggregate data this means pointers to where the original data came from.

#### *2.2.2.3 Dataset Licensing*

Licensing of the data source is crucial since it provides the legal conditions under which it can be used. Also it impacts the data integration since incompatible licenses could mean that combining or enriching data sources would lead to restricted usable aggregations or even illegal.

### 2.2.3 Intrinsic dimensions

#### *2.2.3.1 Dataset accuracy*

This is linked to completeness of the dataset and the provenance of the dataset, but also encompasses such things as the processes and methods used to collect the data (intersecting here with trustability).

#### *2.2.3.2 Dataset interlinking*

Linking datasets together is one of the main ways of enhancing/enriching a given dataset. Interlinking datasets does, however, come with a data management impact. Whether that impact is significant or not will depend on the durability of the dataset (See Section 2.2.4.2).

#### *2.2.3.3 Dataset consistency*

The dataset value ranges should be defined, so the question then is whether the values are within the expected ranges and what values outside those ranges mean (could be a local application usage which needs to be verified). This could also take into account the coherency of the dataset and the conclusions which can be derived from combining the data in various ways. A common example in our year one harvesting efforts was that of inconsistent use of country codes/names (e.g. 'IRL', 'IRELAND' or 'IE' to refer to Ireland).

### 2.2.4 Accessibility dimensions

This is the physical assess to the data source or in this case internet based computer access to the dataset. In terms of harvesting efforts, APIs are usually preferred over data dumps.

#### *2.2.4.1 Dataset availability*

The dataset should be available for automatic processing. This does not, however, mean it should be available on-demand. It could be that a new dataset is made available on an FTP site as a tarred GZIP file. Or it could be once a month or year, but the data set availability should be predictable (so the harvesting can be scheduled in an ETL processor, such as UnifiedViews).

#### *2.2.4.2 Durability of the data source*

This question is the data source version of the questions discussed in Deliverable 2.7, Data Management Plan v1.0. If the data source is going to be supported over the long-term (e.g. DBpedia, Diavgeia, IATI, OECD, UN Comtrade, TED, etc.), it is likely that its quality will improve further over time. Assessing the durability of the data source also gives a measure of how valuable the data source could become, even if it is currently incomplete or limited in scope. At the other end, there are experimental data sources which could be interesting, but are very unlikely to be supported beyond the end of the project.

### 2.2.5 Representational dimensions

This section describes the usability of the dataset within the YDS pilots and the effort required to complete the mapping requirements.

#### *2.2.5.1 Dataset understandability and interpretability*

If the dataset contains only codes without an index or description of those codes then accessibility is very limited (note: the language of the models and descriptions are also important). The model has to be penetrable to reduce the effort of the data wrangler. Guessing at the meaning of the fields or structure of the data source reduces the likelihood of data mapping success (RDF predicates are assumed to be meaningful, but badly chosen predicate names have the same problems). Mappings should ideally enrich the data rather than lose data distinctions (e.g. address components should be separate rather than merged together into a single string value).

#### *2.2.5.2 Dataset dynamicity*

Most datasets are updated, others are snapshots of data states taken at a given point in time and, once archived, are not further updated (until the next snapshot). It is the intended usage of the dataset and the dynamicity of the data source that determines whether those intentions are compatible (this can also be defined as the lag between the data being available and the data being accessed).

#### *2.2.5.3 Age of data*

It used to be the case that datasets were updated, and distributed on DVDs, at regular intervals. People then purchased the update when they felt it was necessary (e.g. road maps, etc.). This meant often that the data users were relying on was out-of-date (often when purchased or received). This distribution method is not really used any more, but the requirement to know how old the data is remains. Using an online telephone number/address dataset requires a trust in the relevance of the data returned (using one from 20 years ago would have little relevance today, even a few years out-

of-date could reduce the reliability of the search results returned). However, a copy of the 1911[8] census is still of interest and will still be of interest in a 100 years' time.

## 2.3 Objectifying the data source assessments

When large numbers of possible data sources must be assessed, objectification of the parameters is desirable to support the decision making process. In order to be efficient and supportive the objectification must be semi-automatic, easy to answer by the people that assess and the outcome should be understandable but not trivial. Deploying a semi-automated decision support system is only helpful if the outcome of the process is helpful to the persons active in the process. Otherwise those actors will feel it more as an administrative burden, leading to lower quality work. Objectifying the assessment parameters represents a serious effort as it formalizes the experience done manually.

In [9], Amrapali et al. compare a number of quality assessment procedures described in the literature. They present a comprehensive overview of many dimensions according to which the actual data quality can be measured (objectively or subjectively). In the presented data source assessment these dimensions are touched during the process. Some of the presented dimensions however are beyond the question whether a source could be integrated in the YDS platform. They correspond to the quality assessment of the result of the aggregation process. Those dimensions are hence better addressed via a data quality monitor, than by a prior to harvest decision making process.

Based on our experience thus far, and the numerous different (natures of the) data sources (publishers, formats, APIs etc.) we have dealt with, we have come to a conclusion that building such an assessment process, without putting additional stress on the participants in the process would not be cost effective. It implies investing resources into investigating every possible scenario to build an omnipotent model – a goal which has proven elusive in initiatives dealing with similar issues, such as the EC's own SEMIC[9]. Instead of forcing a model, we still pay attention to each of the above described dimensions and provide general guidelines, but leave it to the person performing the assessment to evaluate a data source across those dimensions as they see most appropriate, given the nature of the source at hand. In Section 4.1, we give an overview of data sources chosen as viable in Y1, with only a high level score with respect to quality, and provide additional explanations in the accompanying text. An example of a more detailed breakdown, per dimension, is given in the Annex.

[8] http://www.ukcensusonline.com/census/1911.php
[9] https://joinup.ec.europa.eu/community/semic/description

# 3 High Level Overview of the Assessment Process

This section provides a high-level view of the basic data source selection process, while the previous Section 2 Data Source Assessment Criteria indicated some of the questions to be answered about the data source during the assessment. The flow of the assessment process is shown in "**Error! Reference source not found.**" with the aim of assessing the feasibility of (re)using the data sources. For each data source, the business relevancy in all its dimensions is assessed. When the outcome is positive the technical usage is assessed. It may be that technical integration of a relevant dataset is near to impossible because the distributed format cannot be easily handled or will not lead to a sustainable data pipeline. Those cases require a search for alternative input formats. The key roles that are involved are the content business owner and the data wrangler. The activities are described in more detail in this section. The subsequent section 4 lists a number of checkpoints and expected outcomes of the data source assessment process in more detail.

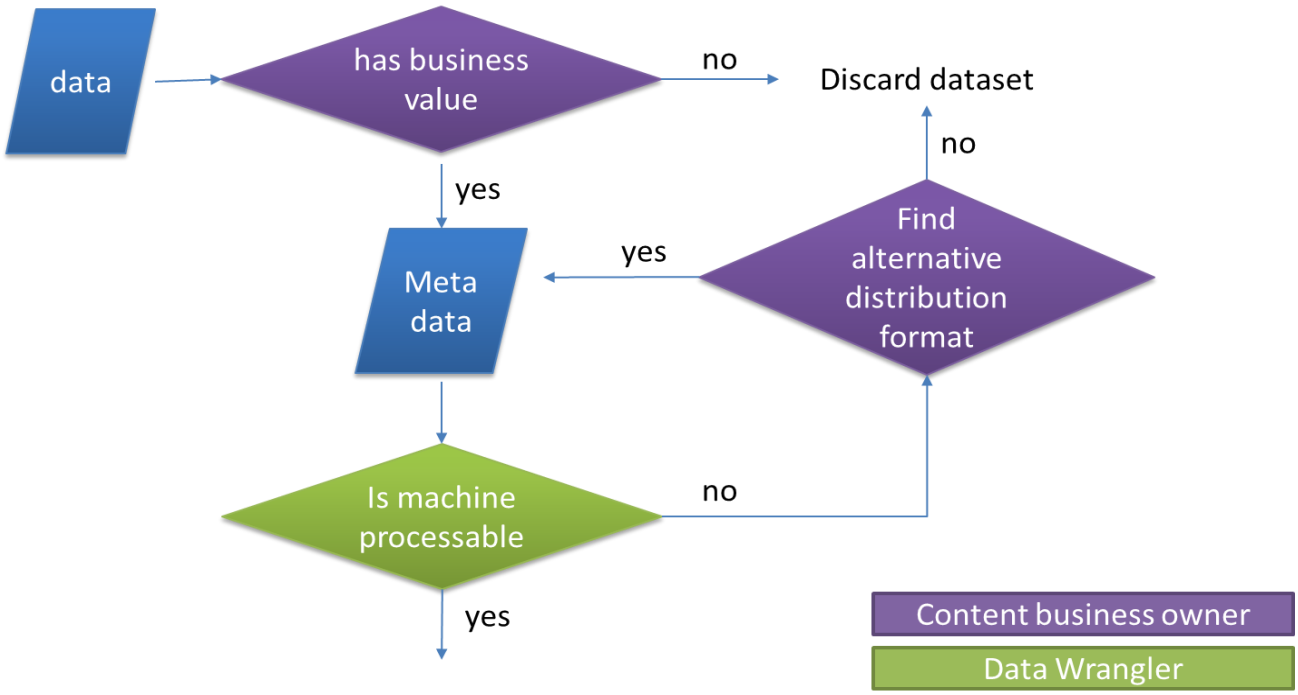


**Figure 1: YDS Data Source Assessment Flow**

In D2.7 [2], the YDS DMP is described which will also be created by the pilot business owner during the assessment of the user requirements (in D2.3 [6]). The DMP is applicable to each of the data sources and any newly created aggregated data. The focus of the DMP is on four main data parts of the data source life-cycle:

- Initial data source selection and recovery,
- Accessibility/use during the pilot/project life-time,

- Data source enrichments and aggregations,
- Data archiving following the project/pilot de-commissioning.

As such, the results of the data source assessment feeds directly into the creation of an instance of the DMP required for each of the YDS pilots and their sources. The DMP requires a consideration of the longer-term project objectives and not just the short term possibilities relating to the data source (i.e. is it technically possible). Considerations and questions outside of the purely technical, "can it be harvested question?" are part of "has business value" question and are the responsibility of the content business owner (See Section 3.1.1).

## 3.1 Relevant Roles involved in the data source assessment

### 3.1.1 Content business owner

For each of the content domains – each of the YDS pilots corresponds to at least one content domain – there is a content business owner (CBO), who is responsible for finding and assessing the necessary data sources to achieve the objectives of the CBO's content and application domain.[10]. This activity is an essential part of the user requirements analysis (See D2.3 [6] for more specific details), but the sort of non-technical requirements which need to be addressed are to:

- Define the domain's objectives and type of data sources needed,
- Identify and analyse the user requirements with the data sources in mind,
- Indicate the data source(s) relevancy to those objectives and requirements,
- Define the business value for the data sources and the end user applications
- Identify the appropriate licence(s),
- Defend the durability/volume/etc. of the data sources,
- …
- When business relevant, the information is reflected as meta-data of the source according to DCAT-AP vocabulary.

All these factors are also needed for instantiating the YDS DMP [2] and to aid the PDW (Section 3.1.2) in initially assessing the effort required in harvesting the data sources (which it is intended will be using UnifiedViews). When later following the initial assessment, this information is needed for actually performing the harvesting of the data from the data source and setting up the necessary schedules automatic harvesting operations.

### 3.1.2 Domain Content/Data Owner

The person responsible for the original data source and the original data sets is the domain content owner (DCO). They should know or have access to the descriptions of the original data sources. Contact is needed between them and the domain data wrangler to determine how to extract the content from the data source.

[10] Pilot in the very generic sense (could be in any application domain or in any data domain)

### 3.1.3 Domain Data Wrangler[11]

The role of the data wrangler (DDW) is to facilitate the movement of the data from the data source to the end-user application. This often means development of translation/transformation rules (e.g. SPARQL operations for use in a UnifiedViews DPU). However, in the initial stages in the data source assessment, the questions to be answered by the PDW is whether or not the data is available in a format suitable for machine processing. The basic flow is as follows:

- Is the data in a suitable format for machine processing (RDF, XML, etc.)?
  - Result: (not) acceptable data
- Is the content of the data, machine transformable without information loss?
  - Result: effort estimate for "raw" ingestion work
- Does the data model cover the content of the provided data
  - Result: effort estimate of data integration and harvesting work

A second portion of this role is to create a risk assessment or perform initial work required to harvest or (re)use the data (i.e. map the data to the YDS model). Such a harvesting assessment defines initial assessments of the cost/difficulty of performing the data mapping (understanding the source and target data models are a necessity here[12]). Additional to the data source mapping assessment is the question of whether it covers the expected user requirements for the data and the model? To determine this, the DDW:

- Starts a pre-data integration analysis to make sure that the core elements of the source data have a counterpart in the YDS data model,
  - In case this is absent information, investigate the possibilities of an extension of the data model.
- Identifies the scope of the data source and how the mapping between data sources can be performed (if needed).

The DDW has to find a balance between having compact data flows and the debugging potential for when harvesting goes wrong.

[11] This is seen here as being a specific role for *each* of the YDS pilots (even if the end data model is a common one, the source data models are often very distinct and require local communication with the CBO). They also have to be directly involved in the development of the end-user target data applications.

[12] Even the assessment can require trying to understand the basics of the model to be harvested (e.g. asking the data source owner about the meaning of column names).

# 4 Recording the data source assessment progress

Experience in the Open Data Support Project [3], has shown that getting access to data sources takes a considerable amount of time and effort. Equally, even when the sources are identified, progress on harvesting the data is at-variables-speeds. The main harvesting bottlenecks being the communication costs in requesting clarification of element/data item locations, definition of the value mappings, etc., all of which take a considerable amount (possibly hunting through code to understand a datum origin) and this is typically not a predictable response time operation. In the following, this process is broken down into three main steps:

- An initial data source assessment of the data source (which is intended to be quick),
- The creation of a harvesting assessment which will start to sketch the pipeline required to harvest the data, as well as the potential gaps.
- Finally, the creation of the harvesting plan which will contain all the information needed to create the pipeline, schedule it and map the required information.

This sort of step-like approach looks like it could be supported via a workflow application. However, as it stands today, the design and setup of such a workflow system would be of limited use. It would merely be an additional administrative overhead. Whenever the project reaches sufficient maturity and the complete harvesting process gets more crystallized, clear repetitive tasks can be chained together to assist the data source assessment and follow up. This section will therefore suggest a number of *checkpoints* which will describe the typical outcome of each of the assessment stages for each content domain, and thus in the first place for each YDS pilot.

## 4.1 Initial data source assessment and experiences

The initial data source assessment is intended to be a quick one, performed by the CBO and DDW, and is very likely to result in discarding a number of potential data sources as being unusable or having insufficient quality data. However, even for data sources which are deemed at present not suitable, the information should be maintained for future reference. Sometimes data sources will need to be reassessed at regular intervals because of enhancements to the data source or loss of the preferred data source (service might be stopped).

The basic steps which are needed here are:

1. Definition of the Pilot Objectives (CBO)
    - Which give the end-user expectations for the data,
    - Also, some story definitions detailing how the user will interact with the data.
2. Data source(s) identification (CBO)
    - Licensing details, contact points, durability, etc.
3. Data source(s) accessibility and convertibility testing (DDW)

- Technical verification of data source accessibility[13]
  - Access the URL/REST interface (persistent URI required),
  - Identification of the UnifiedViews DPU to be used for the extraction,
  - Recover data source contents, etc.,
  - Input format is acceptable?
- Provide an assessment of the cost of converting the data to RDF
  - Quality of the documentation available,
  - Outline of harvesting sequencing,
  - Volume of data available,
  - Quality of data available.
- Data source coverage is used to assess the data source data fields with the required data fields for the application (it must be noted that, sometimes, several data sources need to be consolidated to get the data views required by the pilot application[14]).

4. Definition of target validation rules (DDW)[15]
5. Determination of data source(s) viability (CBO), which at this points should mean that all the necessary information is available for the:
   - Creation of the Data Source Harvesting assessment (Section 4.2)
   - Partial information for the creation of the data source harvesting plan (Section 4.3)

It should be noted, that the above does not indicate an exact way for all the YDS pilots to assess their data sources; this is up to the CBO/DDW to define, since it can vary for each pilot (See Section 2.3). For example, if the costs of accessing the data are too high then the associated data entry costs could mean that the application at present isn't cost effective (but spending time to scrape a high-value web-site could be very cost effective if the free alternatives aren't as comprehensive). Deciding on these points is domain and application specific and so the approach here is a map towards harvesting the data source, which can be stopped at various points if the data is deemed to be unusable at present.

In Deliverable D2.1 [7], the first initial assessment was formalized by answering the questions in the open data certification process as specified by the ODI [8]. For each pilot key data sources have been identified. Their assessment was summarized in D3.1 Data Source Assessment Methodology v1.0.

A more careful inspection of the listed data sources showed the importance of the assessment methodology presented in D3.1. For instance, the document listed the Atlas of Economic Complexity[16]

[13] From ODS, the easiest way to validate this is to try to create the outline unified views pipeline (first thing is to find out how to extract the data from the target – developing new UnifiedViews DPU's will increase the cost of the harvesting).

[14] Which raises the question of how the data should be consolidated (which is means of identity/reference shared between the data sources – e.g. country code or URI or UUID, etc.)

[15] In the case of a shared target data model, these validation rules could be defined once, but this depends on the specific pilot business objectives. Too strict validation rules could exclude data which, while not suitable in one of the pilot data applications, might be acceptable in another (mandatory vs. optional predicate expectations could conflict, etc.)

[16] http://atlas.cid.harvard.edu

as a potential source of international trade data for Pilot 2. The Atlas is a powerful interactive tool that enables users to visualize a country's total trade, track how these dynamics change over time and explore growth opportunities for more than a hundred countries worldwide. However, as also revealed in D3.6 Data Harvesters v1.0, the solution hosted by Harvard University did not offer machine-readable data. The one hosted by the Massachusetts Institute of Technology[17] did (as CSVs), which made us abandon the former.

Furthermore, a closer look at the contextual dimensions, namely, the dataset verifiability, later revealed that a number of sources did not represent the actual point of origin for certain data. For international trade data, the original source of the data, for both instances of the Atlas, was the United Nations Comtrade Database[18]. Moreover, in the case of Official Development Assistance data, the source behind AidData.org, was the International Aid Transparency Initiative[19] (IATI). Further investigation lead us all the way to the Dutch Ministry of Foreign Affairs. However, the Dutch government, just like any other government in the world that provides its data to the IATI, is obliged to comply with the reporting standard set by IATI. A standard vocabulary, an open license, trusted publishers and an open API easily made IATI a viable data source, replacing AidData.org.

### 4.1.1 Selected sources

The tables below represent the list of sources tackled so far, reflecting on our experiences, per pilot, up until M18. As described earlier, in Section 2.3, objectifying each of the assessment criteria is anything but trivial, so we indicate the overall level of provided support using the values given in Table 1: Level of support, and provide additional explanations in the accompanying paragraph, where needed. An example of a more detailed breakdown is provided in the Annex.

**Table 1: Level of support**

| | |
|---|---|
| ✓ | supported |
| ✓ | partially supported |
| × | not supported |

As described in D3.6, pilot one covered the published administrative decisions of the Greek Transparency Portal and the official portal for distributing the progress report of the NSRF (National Strategic Reference Framework). Even though machine readability and harvestability (R & H) were not at the desired level in both cases, as the latter was not available in one of the preferred formats, the positive assessment across all the data quality dimensions (Section 2.2) proved building a custom harvester worth the effort. At the time of writing, though, the Greek government published an API, which will ease future harvesting efforts.

[17] http://atlas.media.mit.edu
[18] http://comtrade.un.org/
[19] http://www.aidtransparency.net/

Table 2: Pilot 1 data sources

| Data source | Machine R & H | Quality |
|---|---|---|
| http://opendata.diavgeia.gov.gr | ✓ (API, JSON) | ✓ |
| http://anaptyxi.gov.gr | ✓ (HTML) & ✓ (API, JSON) | ✓ |
| https://www.cityofathens.gr/khe/proypologismos | ✓ (API, JSON) | ✓ |
| http://e-prices.gr | ✓ (API, JSON) | ✓ |
| http://www.fuelprices.gr | ✓ (API, JSON) | ✓ |

As mentioned earlier, pilot two initially relied on one of the two Atlases of Economic Complexity as a source of international trade data. Even though the UN Comtrade Database API limits the number of records to be retrieved per call, as the original source of the data in question, it scored higher with respect to data quality dimensions, especially durability (Section 2.2.4.2), dynamicity (Section 2.2.5.2) and age (Section 2.2.5.3), making it the preferred choice (please consult the Annex for the complete overview).

Table 3: Pilot 2 data sources

| Data source | Machine R & H | Quality |
|---|---|---|
| http://atlas.media.mit.edu | ✓ (API, CSV) | ✓ |
| http://comtrade.un.org | ✓ (API, CSV) | ✓ |
| http://www.aidtransparency.net | ✓(API, CSV) | ✓ |

With cross-Europe Financial Comparability as the ultimate goal of the third pilot in mind, after a detailed assessment of the potential pilot three data sources, the Tender Electronic Daily[20] came out on top as a highly relevant dataset for the use case. Further investigation lead to a mirror database provided by the Open Spending project[21] offering a highly accessible, easy to transform data dump.

Table 4: Pilot 3 data sources

| Data source | Machine R & H | Quality |
|---|---|---|
| http://ted.openspending.org | ✓ (CSV) | ✓ |
| http://data.gov.au/dataset/historical-australian-government-contract-data | ✓ (XLSX) | ✓ |
| http://ec.europa.eu/budget/fts/index_en.htm | ✓ (CSV, XML) | ✓ |

### 4.1.2 Quantity vs. quality

It is worth noting that YDS is not a Big Data project. Therefore, it is not our goal to amass as much data as we can, so as to cover the pilot cases. Instead, after having established the basis for the three

[20] http://ted.europa.eu
[21] http://www.openspending.org

pilots using the above listed datasets, we now focus on maximizing the *added value*. In terms of contextual and intrinsic data quality dimensions, we look at *dataset relevancy* (Section 2.2.1.3) and *interlinking* (Section 2.2.3.2).

This inevitable and expected change of perspective resulted in a new list of potential data sources. We, once again, provide only an initial evaluation. In other words, the list provided below (Table 5) is not final; it is expected to evolve as a direct outcome of further data source assessment. It should also be noted that in the second part of year two, the plan is to consider social-to-semantic integrations, but the requirements for such integrations are yet to be defined.

**Table 5: Potential data sources**

| Pilot | Data source | Machine processable data ready for use | Data collection |
|---|---|---|---|
| **1** | http://www.hsppa.gr/index.php/m-gnomes/m-oles-oi-gnomes | Yes (CSV) | Yes |
| **2** | https://aiddata.rvo.nl | Yes (CSV) | Yes |
| **2** | http://aiddata.org/maps | Yes(CSV) | Yes |
| **2** | https://www.rijksoverheid.nl/opendata/ontwikkelingssamenwerking | Yes (CSV, XML) | Yes |
| **2** | http://data.worldbank.org/products/wdi | Yes (CSV) | Yes |
| **3** | http://data.cso.ie | Yes (RDF) | Yes |
| **All** | http://dbpedia.org | Yes (RDF) | Yes |
| **All** | http://ec.europa.eu/eurostat/data/database | Yes (TSV[22], XML etc.) | Yes |

## 4.2 Data Source Harvesting Assessment

The initial data source assessment steps gather a potentially large amount of data together and this is used to create "Data Source Harvesting Assessment". This is a formalization of the available information (so that the assessments can also be compared with the pilot objectives). Information which should be available:

1. Pilot Data Quality objectives and end-user data interaction descriptions,
2. DCAT-AP Meta data concerning the source,
3. Individual instance of the DMP for that pilot data-source(s) [2],
4. Bibliography of data model descriptions,
5. Outline of the harvesting stages (by DDW)
    a. Identification of the extractors (e.g. UnifiedViews DPUs) to be used,

[22] TSV is often considered an acceptable alternative to CSV.

   b. Identification of missing DPUs (which will have to be developed),
   c. Outline of the harvesting pipeline structure.
   d. Assess the quality of the data which can be obtained (See Section 2.2)
6. Costing/risk assessments for creation of the harvesting plan (by DDW).

Once this information is available, the selection of the data source to harvest can be made and the creation of the harvesting plan can be started.

For each of the potential sources the above questions have to be elaborated. From the experience gained in Y1, the following conclusions can be drawn:

- Many of the data sources are themselves catalogues of datasets. The assessment on the data has to be done for each dataset that is part of that catalogue. Since it cannot be guaranteed that each dataset is of identical structure, the assessment can cost a substantial amount of time. The CBO and the DDW have to decide which datasets to assess by priority.
  - It is worth noting that not all datasets in a catalog are expected to be accompanied by the same license. Moreover, special attention needs to be paid to datasets that integrate data from other datasets.
- The most frequent returning data format is CSV. The format is machine processable, but usually represents a large amount of semantic assessment work.
  - In relation to this, sometimes tracing the data back to its point of origin (i.e. original publisher) can lead to corresponding vocabularies and code lists that provide the necessary semantics.
- Some of the data sources are only available in HTML. If the HTML is structured enough, and such an approach is expected to prove worthwhile (as in the case of the NSRF data), scraping can be used to extract data. However, this is a very vulnerable data extraction methodology. It is advised to first look for alternative accesses to the data.

## 4.3 Data Source Harvesting Plan

Following the creation of the "data source harvesting assessment" there are a number of follow-up tasks which the PDW has to perform to create the "Data Source Harvesting Plan":

1. Definition of the ETL (e.g. UnifiedViews) Pipeline,
2. Description of Required DPUs, which would require for each DPU:
   - Description of the DPU (input and output/RDF vocabularies),
   - Development of the data source DPU (if not available).
3. Development of the mapping rules (SPARQL transformations),
4. Validation of mapping results (see final points above),
5. Definition of the harvesting schedule,
6. Assessment of the harvested data.

The initial steps (Section 4.1) are the responsibility of the CBO, although the costing and technical risk assessments in doing the extraction will need input from the DDW. These steps are essential to setup everything needed for the data wrangler to complete the data source harvesting plan and to implement the actual harvesting of the data source. The estimates on the costs are essential to determining the practicality of using the data source[23] for the pilot application.

As the summary of the initial data source assessment shows, many of the identified data sources represent data catalogues. An important parameter in the cost estimates is the amount of datasets that has to be retrieved from those catalogues. In some cases, this can be done with a universal conversion process applicable to all datasets in the harvested catalogue. However, it may be that the catalogue contains a very diverse set of data (with respect to both structure and licensing, as mentioned earlier) and then individual pipelines have to be setup.

[23] The pilot business objectives are the final determinants though on the acceptability of the cost of the data source harvesting.

# 5 YDS Pilot data assessments status

In D2.1 [7] User Characteristics and Usage Scenarios v1.0, sample scenarios have been sketched, and then further elaborated in D2.2 User Characteristics and Usage Scenarios v2.0 [12]. They refer to some data sources, for which the above data source assessment is briefly demonstrated. These user-level data source assessments represent the CBO aspects of the initial assessment steps described in Section 2.3 and were followed by the PDW assessment to create the harvesting assessment:

- A technical assessment of the viability of recovering the information from each of the data sources (which should be done by the PDW). The main decision criteria were thoset given in Section 2.1 concerning the machine readability of the data, but there were other questions to be addressed, as further elaborated in this very document,
- Outlines of the pipeline to be used (and missing components)
- Assessment of the cost of the data recovery (manpower, time, etc.),
- Alternative sources if the cost of information recovery was deemed prohibitive.

Once the harvesting assessment is created for each individual, harvestable data source, the PDW uses it to create the harvesting plan (Section 4). The data source harvesting plan is the final assessment on the usability and data QA of the data source (which typically results in the items described in Section 4.2; including the DMP, DCAT-AP entry, bibliography, etc.,). The plan contains everything needed to actually start the definition of the harvesting pipeline and the harvesting of the data. The harvesting plan would then need to be executed (by the PDW).

The end results of the data source quality and harvesting assessments, along with the harvesting plans and descriptions of respective ETL pipelines for the three pilots are given in D3.6 Data Harvesters v1.0. Nevertheless, further alterations and improvements are always possible, as seen in the case of the international trade data (pilot 2).

It is also worth noting that in order to support certain data quality dimensions, such as that of understandability and interpretability (Section 2.2.5.1), additional efforts might be needed. Providing a necessary code lists in a machine readable and interoperable format, while remaining in line with the existing data model, might imply an assessment and a harvesting plan of its own, as certain metadata, such as that provided by concept schemes / taxonomies can be interpreted as separate data sets/sources. The IATI data alone is supported by 9 different code lists, maintained by 3 different publishers (IATI, OECD, ISO).

# 6 Conclusions and Future Work

This deliverable has outlined how data sources should be assessed with the intention of creating a data source harvesting plan for the usable data sources. It is a refinement of the first version of the methodology, based on the real world experiences and applications encountered since month six of the project. The original methodology is improved with respect to the assessment criteria, both by elaborating further on the individual aspects, and by highlighting the most important criteria as the project progresses (i.e. by shifting the focus from data quantity onto data quality). Moreover, an overview of the outcome of the year one assessments is given, along with an initial assessment of the potential sources for the future.

For those data sources which are chosen as viable, the expected outcome is to follow through and create a data source harvesting plan (as indicated in Section 4). The approach described is one which would need to be followed for using UnifiedViews as the harvesting engine, but whatever the harvesting tool, the same inputs would be required. Since this is an assessment of the data source with respect to the end-user applications, the roles involved in the assessment are those closest to the pilot application areas (i.e. the CBO and PDW). One of the key outputs of year one assessments is represented in terms of the relevant DCAT-AP descriptions, as shown in the Annex of this deliverable.

This is the final version of the data source assessment methodology and, as such, provides the most complete overview of the crucial QA aspects, as seen from different perspectives, be it legal and applicability aspects, business added value, technical harvesting feasibility, or quality requirements, ensuring reusability beyond the lifetime and scope of this project.

# 7 References

[1] Deliverable 3.1 - Data Source Assessment Methodology v1.0

[2] Deliverable 2.7 - Data Management Plan v1.0

[3] Open Data Support [http://open-datasupport.eu/]

[4] UnifiedViews, [http://www.unifiedviews.eu/]

[5] An Overview of the PROV Family of Documents, [http://www.w3.org/TR/prov-overview/]

[6] Deliverable 2.3 - User Requirements v1.0

[7] Deliverable 2.1 - User Characteristics and Usage Scenarios v1.0

[8] https://certificates.theodi.org/

[9] Zaveri, Amrapali, Rula, Anisa, Maurino, Andrea, Pietrobon, Ricardo, Lehmann, Jens and Auer, Sören. "Quality Assessment for Linked Data: A Survey." *Semantic Web Journal* (2015)

[10] Data Quality Assessment Framework [http://dsbb.imf.org/images/pdfs/dqrs_factsheet.pdf]

[11] Deliverable 3.6 - Data Harvesters v1.0

[12] Deliverable 2.2.- User Characteristics and Usage Scenarios v2.0

# 8 Annex

## 8.1 Sample DCAT-AP description

```
<http://linkedeconomy.org/resource/Dataset/ODA-NL>
        a dcat:Dataset ;
        dct:title “Official Development Assistance (ODA) of the Netherlands” ;
        dct:description “Official development assistance (ODA) is a term coined by the Development
        Assistance Committee (DAC) of the Organisation for Economic Co-operation and Development
        (OECD) to measure aid. It is widely used as an indicator of international aid flow. The data set
        covers Dutch government funded ODA activities in developing countries worldwide and is
        derived from the data provided by the International Aid Transparency Initiative (IATI)..” ;
        dct:publisher <http://linkedeconomy.org/resource/Organization/YourDataStories> ;
        dcat:distribution [
                a dcat:Distribution ;
                dct:description “The YourDataStories (YDS) SPARQL endpoint provides access to all YDS
                datasets. The ODA dataset is represented in the form of the
                <http://yourdatastories.eu/data/ODA> graph.” ;
                dct:license [
                        a dct:LicenseDocument ;
                        rdfs:label “IATI Standard (http://iatistandard.org/)” ;
                        rdfs:comment “Copyright © Development Initiatives, on behalf of the IATI
                        Secretariat Released under the Creative Commons attribution license
                        http://creativecommons.org/licenses/by/4.0/” ;
                ] ;
                dcat:accessURL <http://143.233.226.61:8890/sparql> ;
                dct:format <http://publications.europa.eu/resource/authority/file-type/SPARQLQ> ;
                dct:modified "2015-12-22T11:01:31+0000" ;
        ] ;
        dcat:keyword “ODA”, “development”, “public spending”, “Netherlands” ;
        dcat:contactPoint <http://linkedeconomy.org/resource/Organization/YourDataStories> ;
        dct:theme <http://cv.iptc.org/newscodes/subjectcode/04008007> .
```

## 8.2 Data quality assessment example

**Table 6: Pilot 2 data source quality assessment**

| Source | Dataset completeness/coverage | Amount-of-data | Dataset relevancy | Dataset provenance | Dataset verifiability | Dataset Licensing | Dataset accuracy | Dataset interlinking | Dataset consistency | Dataset availability | Durability of the data source | Dataset understandability and interpretability | Dataset dynamicity | Age of data |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| http://atlas.media.mit.edu | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| http://comtrade.un.org | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| http://www.aidtransparency.net | ✓[24] | ✓ | ✓ | ✓ | ✓ | ✓ | ✓[25] | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

[24] Pieces of data are sometimes omitted by the publisher.

[25] Errors at the publisher's end can occasionally hamper not only accuracy, but also interpretability. See http://discuss.iatistandard.org/t/ref-attribute-for-implementing-organizations-in-dutch-oda-data/418