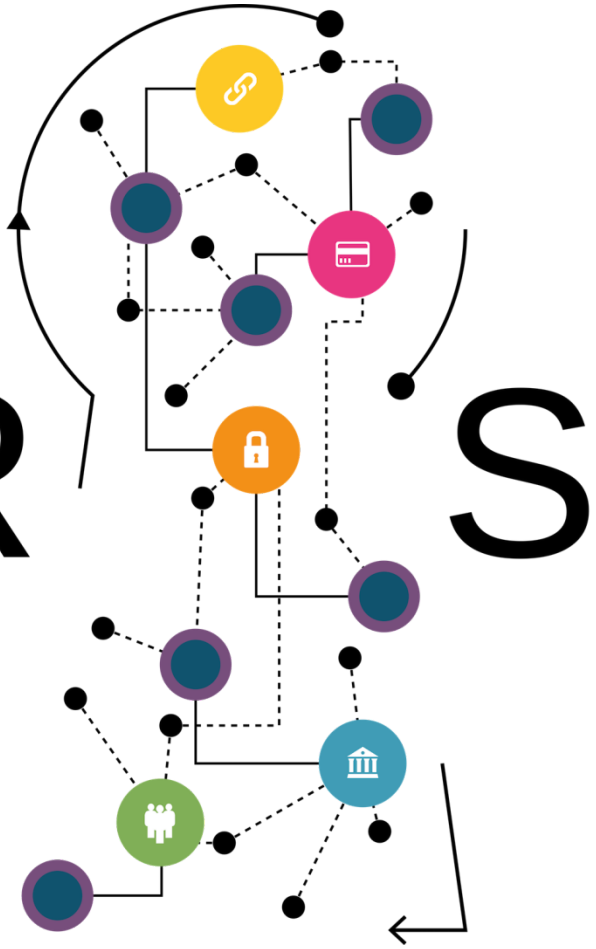
PETRAS



PETRAS

INTERNET OF THINGS
RESEARCH HUB

Enhanced interface for PEDASI-IoT data sources

16 May 2019 (version1.1)

Author(s): Mark Schueler, Steven Crouch, Adrian Cox
James Graham, Wendy Hall

Details of document preparation and issue:

Version no.	Prepared	Checked	Reviewed	Approved	Issue date	Issue status
1.1	Adrian Cox				16.5.19	Final

PETRAS

Enhanced interface for PEDASI-IoT data sources

CONTENTS

Enhanced interface for PEDASI-IoT data sources

1. INTRODUCTION

This project aimed to improve features with respect to PEDASI's interfaces with external data sources (particularly the HyperCat API), beyond those set out in the initial specification for the PEDASI development. The project has explored current programming options to determine desirable and specific feature enhancements that could be prioritised according to their feasibility within the existing PEDASI workplan and the project's timeframe.
The work has enabled enhancements to indicators of data quality and provenance, as well as improved data handling within PEDASI, with the aim to provide greater assurance to users working with the PEDASI infrastructure, along with improving the level of trust in the data processed by the platform. Other key data sources will also be explored for potential inclusion to provide a broader range of demonstrator applications and data synthesis to be explored during the next phase of development.

The outputs documented in this report will be advanced within recommendations from the PEDASI development project through their inclusion in enhanced developer guidance, operational policies, software revisions, further collaboration, and development recommendations for future versions of PEDASI.

Whilst this report's primary focus was research to explore options for the identification of indicators of quality along with methods for their evaluation two additional deliverables focussed on features to enhance security, these included D2.2 Volumetric security & rate limiting, D3.1 Data Validation & Detuning. These are summarised at Appendix 8.2 and 8.3.

2. IDENTIFYING DATA QUALITY INDICATORS

Data services and attendant analyses are necessarily constrained and impacted by the quality of data they provide. PEDASI offers a common data brokerage service and applications interface for applications and researchers to securely search, access, and retrieve disparate IoT data and metadata from diverse sources. The ultimate value of this service relies upon data trust, privacy, governance, and quality and the influencing factors these have on each other. This report provides a summary review of previous related metadata considerations which informed PEDASI's precursor, the IoT Observatory, as well as a survey of current data quality (DQ) indicators across the IoT domain and beyond. It concludes with recommendations for applying heuristics to PEDASI's implementation, describing in turn standards-based, theoretic, and practical approaches to automating this functionality.

As a nascent data brokerage/platform, PEDASI has few datasets and applications on offer at this time. For those that are available, it provides the ability for data and application providers to capture metadata describing the service(s). The earlier IoT Observatory offered access to dozens of datasets and several applications – also providing structured space for provision of comprehensive metadata. Best practices suggest that such provision is constructive of DQ [5]. On recent review, these metadata fields were found to be mostly vacant. Indeed, metadata absence presents a significant impediment to DQ assurance [2, 5].

PEDASI adds value by enabling recombination of heterogeneous datasets in a secure and accountable manner, which provides feedback to such public initiatives and helps to improve DQ and trust in a virtuous cycle. You have to manage the quality of the data coming out of such as smart cities, the provenance of that data, and the metadata around that [1].

PEDASI currently supports globally defined metadata fields that can apply to data sources. With appropriate development and additional research, extensions to this model will support the management of DQ indicator questions for 'Data Providers', and enable PEDASI Central Administrators to specify tiered assessment frameworks that classify these questions and their answers within specific DQ 'maturity levels' within an assessment framework.

These enhancements will enable PEDASI to automatically evaluate DQ for a given data source against this framework and present an assessment report that includes recommendations to Data Providers to improve DQ. This could include development of assessment scoring for displayed data sources within the PEDASI service for users and re-users.

3. METADATA SCHEMA STANDARDS

Effective and intelligent IoT data management requires high-level analytic information in addition to accurate and reliable data. For those considered data re-users, including application providers, they would benefit from as much contextual information, or metadata, as possible to enrich and inform the lower-level data represented. Earlier work on the IoT Observatory covered these needs in detail [4]. Earlier research made the case that IoT data may be adversely impacted by heterogeneity, inaccuracy, real-time volume, and lack of semantics. The IoT Observatory was designed to meet these challenges and to achieve "conceptual interoperability" by providing the means to capture low-level 'access' metadata as well as high-level 'semantic' metadata.

The earlier work first considered application layer protocols, such as HTTP REST, CoAP, MQTT, MQTT-SN, XMPP, AMQP, DDS, and Websocket. While useful in understanding system level requirements and constraints (and informing contents and formats of data), these are not in themselves constructive of DQ, and may be set aside from consideration as such.

Similarly, common data encodings help to facilitate interoperability, but do not impact on DQ (again, in themselves; subjective implementations can offer DQ improvements – see below in cataloguing technologies) any more than application layer protocols. Encodings described include such text-based models as XML, CSV, and JSON, as well as binary encodings CBOR, Protocol Buffers, and Cap'n Proto.

Finally, specific metadata and ontologies for IoT were considered, including sensor/device ontologies, domain-specific ontologies, and cataloguing technologies. The first section here described sensor/device ontologies including Sensor-ML, OntoSensor, SSN, SWAMO, CESN Ontology, A3ME, SCO, SAREF, and BSI Publicly Available Specifications (too close to hardware). "In summary, existing ontologies for

describing sensors and devices focus on different aspects associated with the hardware and software and the data produced by the sensors, such as sensor identities, locations, functions, capabilities, configurations, discovery mechanisms, data access and sharing mechanisms and data descriptions" [4]. The second covered such domain-specific ontologies as COBRA-ONT, WGS84, Kim et al., Okeyo et al., Brick Schema, and several temporal ontologies. The third and final section offered brief descriptions of the Hypercat, DCAT, and Schema.org, cataloguing technologies. Being more prescriptive than the lower-level protocols, the latter, in particular, begin to offer some contextual dimensions which help with DQ assurance, such as ease-of-sharing and, through the use of JSON documents, greater flexibility in describing datasets in detail [4].

The redefined IoT Observatory, delivered in the PEDASI project, supports RESTful APIs that are quite similar to Hypercat, as well as supporting DCAT and Schema.org markups for greater metadata detail of each shared dataset and application. It further uses a lightweight vocabulary called IoTO which helps with dataset discovery, search, and access. Moving forward, PEDASI supports Hypercat in its initial implementation, as well as RESTful APIs and JSON documents. These then provide the foundation for rich detailed metadata and, with successful application of the practical approaches detailed in Section 5, reliable DQ improvements.

4. THEORETIC APPROACHES

Many DQ assessment standards have been articulated by organisations and researchers seeking to improve DQ practices. This section reviews three standards which offer theoretic guidance but little practical implementation advice. As such, they provide several perspectives on what features of data may be used to assess its quality, but not how this may be achieved.

4.1 IMF's data quality assessment framework

The International Monetary Fund (IMF) described five dimensions of DQ in addition to a set of prerequisites for statistical analysis [6]. Shown in Table 1, it further articulates associated DQ indicators as sets of policies, but stops short of any description of automation for use with big data.

Quality Dimensions	Elements
0. Prerequisites of quality	0.1 Legal and institutional environment —The environment is supportive of statistics.
	0.2 Resources—Resources are commensurate with needs of statistical programs.
	0.3 Relevance—Statistics cover relevant information on the subject field.
	0.4 Other quality management—Quality is a cornerstone of statistical work
1. Assurances of integrity *The principle of objectivity in the collection, processing, and dissemination of statistics is firmly adhered to.*	1.1 Professionalism—Statistical policies and practices are guided by professional principles.
	1.2 Transparency—Statistical policies and

	practices are transparent.
	1.3 Ethical standards—Policies and practices are guided by ethical standards.
2. Methodological soundness *The methodological basis for the statistics follows internationally accepted standards, guidelines, or good practices.*	2.1 Concepts and definitions—Concepts and definitions used are in accord with internationally accepted statistical frameworks.
	2.2 Scope—The scope is in accord with internationally accepted standards, guidelines, or good practices.
	2.3 Classification/ sectorization—Classification and sectorization systems are in accord with internationally accepted standards, guidelines, or good practices.
	2.4 Basis for recording—Flows and stocks are valued and recorded according to internationally accepted standards, guidelines, or good practices.

Table 1: IMF DQ dimensions [6]

Quality Dimensions	Elements
3. Accuracy and reliability Source data and statistical techniques are sound and statistical outputs sufficiently portray reality	3.1 Source data—Source data available provide an adequate basis to compile statistics.
	3.2 Assessment of source data—Source data are regularly assessed.
	3.3 Statistical techniques—Statistical techniques employed conform to sound statistical procedures.
	3.4 Assessment and validation of intermediate data and statistical outputs—Intermediate results and statistical outputs are regularly assessed and validated.
	3.5 Revision studies—Revisions, as a gauge of reliability, are tracked and mined for the information they may provide.
4. Serviceability *Statistics, with adequate periodicity and timeliness, are consistent and follow a predictable revisions policy.*	4.1 Periodicity and timeliness—Periodicity and timeliness follow internationally accepted dissemination standards.
	4.2 Consistency—Statistics are consistent within the dataset, over time, and with major datasets.
	4.3 Revision policy and practice—Data revisions follow a regular and publicized procedure.
5. Accessibility *Data and metadata are easily available and*	5.1 Data accessibility—Statistics are presented in a clear and understandable manner, forms of

assistance to users is adequate.	dissemination are adequate, and statistics are made available on an impartial basis.
	5.2 Metadata accessibility—Up-to-date and pertinent metadata are made available.
	5.3 Assistance to users—Prompt and knowledgeable support service is available.

Table 1 (continued): IMF DQ dimensions [6]

4.2 DAMA UK's approach to data quality

Prepared by the Data Management Association (DAMA) UK Working Group on "Data Quality Dimensions", six dimensions of DQ are recommended for use when assessing or describing DQ [7]:

1. Completeness
2. Uniqueness
3. Timeliness
4. Validity
5. Accuracy
6. Consistency

Each of these dimensions is characterised across further dimensions of:

- Definition
- Reference
- Measure
- Scope
- Unit of Measure
- Type of Measure
- Related dimension
- Optionality
- Applicability
- Example(s)
- Pseudo code

Application of these dimensions is described as a mechanistic process which might be:

1. Identify which data items need to be assessed for DQ, typically this will be data items deemed as critical to business operations and associated management reporting.
2. Assess which DQ dimensions to use and their associated weighting.
3. For each DQ dimension, define values or ranges representing good and bad quality data. Please note, that as a data set may support multiple requirements, a number of different DQ assessments may need to be performed.
4. Apply the assessment criteria to the data items.
5. Review the results and determine if DQ is acceptable or not.
6. Where appropriate take corrective actions e.g. clean the data and improve data handling processes to prevent future recurrences.
7. Repeat the above on a periodic basis to monitor trends in DQ.

While this appears to comprehensively address a DQ process, it too stops well short of describing or offering support for scalable automated assessment.

4.3 Data quality assessment of big data

A third theoretic approach considers seven dimensions of DQ which are described in some detail but again not systematised [8]:

1. "*Accuracy*, correctness, validity and precision focus on the adherence of data to a given reality of interest.
2. *Completeness*, pertinence and relevance refer to the capability of representing all and only the relevant aspects of the reality of interest.
3. *Redundancy*, minimality, compactness and conciseness refer to the capability of representing the reality of interest with the minimal use of informative resources.
4. *Readability*, comprehensibility, clarity and simplicity refer to ease of understanding of data by users.
5. *Accessibility* and availability are related to the ability of the user to access data from his or her culture, physical status/functions, and technologies available.
6. *Consistency*, cohesion and coherence refer to the capability of data to comply without contradictions to all properties of the reality of interest, as specified in terms of integrity constraints, data edits, business rules and other formalisms.
7. *Trust*, including believability, reliability and reputation, catching how much data derived from an authoritative source."

This approach concludes with recognition that DQ for BD, "need for the discovery of methods and techniques for the traditional life cycle of data: that for DQ corresponds to a) collection, b) quality assessment, and c) improvement; while for BD corresponds to a) collection, b) fusion, c) analysis, d) processing, and e) usage" [8](emphasis added).

5. PRACTICAL APPROACHES

Current practical DQ assessment practices appear in multiple domains. These include finance and economy, environment, health, energy, education, transportation, employment, infrastructure, and population, not to mention IoT and smart cities. Some of these are more evolved than others and operate at different levels of data granularity.

This section describes two approaches which operate at two such distinct levels: raw data and metadata. While multiple DQ assessment strategies for raw data have been proposed [9, 12, 13, 14, 15], most articulate an extensive list of DQ dimensions (cf., Table 2) and offer high-level procedural implementation advice, but again stop short of describing clear processes for automation.

Dimensions	Definitions
Accuracy	The extent to which data is correct and reliable
Completeness	The extent to which data is not missing and is of sufficient breadth and depth for the task at hand
Consistency / representational consistency	The extent to which data is presented in the same format
Time-related dimensions	The extent to which data is sufficiently up-to-date for the task at hand
Interpretability	The extent to which data is in appropriate languages, symbols, and units
Ease of understanding / understandability	The extent to which data is easily comprehended
Believability	The extent to which data is true and credible
Reputation	The extent to which data is highly regarded in terms of its source or content
Objectivity	The extent to which data is unbiased, unprejudiced, and impartial
Relevancy / relevance	The extent to which data is applicable and helpful for the task at hand
Accessibility	The extent to which data is available, or easily and quickly retrievable
Security / access security	The extent to which access to data is restricted appropriately to maintain its security
Value-added	The extent to which data is beneficial and provides advantages from its use
Concise representation	The extent to which data is compactly represented
Appropriate amount of data/ amount of data	The extent to which the volume of data is appropriate for the task at hand

Table 2: Data quality dimensions; adapted from [14, 15]

5.1 Quality assessment process for big data

This section describes a high-level dynamic assessment process for raw data which starts from a macro level to scope its value proposition. Articulated in Data Science Journal [9], it recognises the emerging big data challenges represented as the 4Vs: Volume, Velocity, Variety, and Value [16]:

Volume	massive amounts of data, with TB or greater being created and stored.
Velocity	these amounts of data are being created with unprecedented speed.
Variety	this includes broad diversity of both structured and unstructured data.
Value	this represents low-value density – the larger the data scale, the less valuable the data.

As such, DQ assessment seeks to surface high-quality, accurate data from massive, variable, and complicated data sets. This faces multiple challenges including: a) myriad data sources producing diverse and complex data structures makes data integration difficult; b) the massive data volume makes it difficult to process in a timely manner; c) data can change very rapidly, making its “shelf life” or working value

quite brief, increasing pressure on its rapid use; d) no broadly endorsed DQ standards and methods are yet available (ISO 8000 identifies DQ principles and processes but stops short of methods or procedures to achieve them [17]). This approach first identifies DQ dimensions and elements, quite similar to those in Table 2, shown in Figure 1:

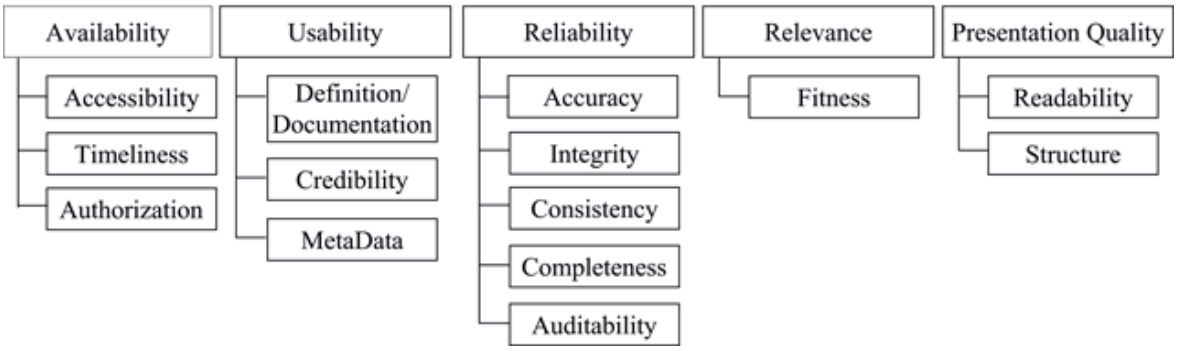


Figure 1: DQ dimensions and elements [9]

A set of DQ indicators are then added to this hierarchy, as shown in Table 3:

Dimensions	Elements	Indicators
1) Availability	1) Accessibility	Whether a data access interface is provided
		Data can be easily made public or easy to purchase
	2) Timeliness	Within a given time, whether the data arrive on time
		Whether data are regularly updated
		Whether the time interval from data collection and processing to release meets requirements
2) Usability	1) Credibility	Data come from specialized organizations of a country, field, or industry
		Experts or specialists regularly audit and check the correctness of the data content
		Data exist in the range of known or acceptable values
3) Reliability	1) Accuracy	Data provided are accurate
		Data representation (or value) well reflects the true state of the source information
		Information (data) representation will not cause ambiguity
	2) Consistency	After data have been processed, their concepts, value domains, and formats still match as before processing
		During a certain time, data remain consistent and verifiable
		Data and the data from other data sources are consistent or verifiable

Table 3: DQ dimensions, elements, and indicators [9]

	3) Integrity	Data format is clear and meets the criteria
		Data are consistent with structural integrity
		Data are consistent with content integrity
	4) Completeness	Whether the deficiency of a component will impact use of the data for data with multi-components
		Whether the deficiency of a component will impact data accuracy and integrity
4) Relevance	1) Fitness	The data collected do not completely match the theme, but they expound one aspect
		Most datasets retrieved are within the retrieval theme users need

		Information theme provides matches with users' retrieval theme
5) Presentation Quality	1) Readability	Data (content, format, etc.) are clear and understandable
		It is easy to judge that the data provided meet needs
		Data description, classification, and coding content satisfy specification and are easy to understand

Table 3 (continued): DQ dimensions, elements, and indicators [9]

In order to proceed with DQ assessment in this model, owing to heterogeneity and subjectivity of raw data across domains and technologies, we must use a process which includes a dynamic feedback loop (cf., Figure 2). It starts from identifying the goals of data collecting, addresses the instant DQ dimensions and elements, determines subjective indicators, and sets an evaluation baseline. Data is then collected, cleaned as necessary, quality assessed, and a determination is made whether the baseline has been satisfied. On failure, the process returns to data collection; on success it proceeds to output creation. If the original goals have been met, the DQ assessment is output; in either case, the findings are furnished back to the baseline evaluation for adjustment and validation.

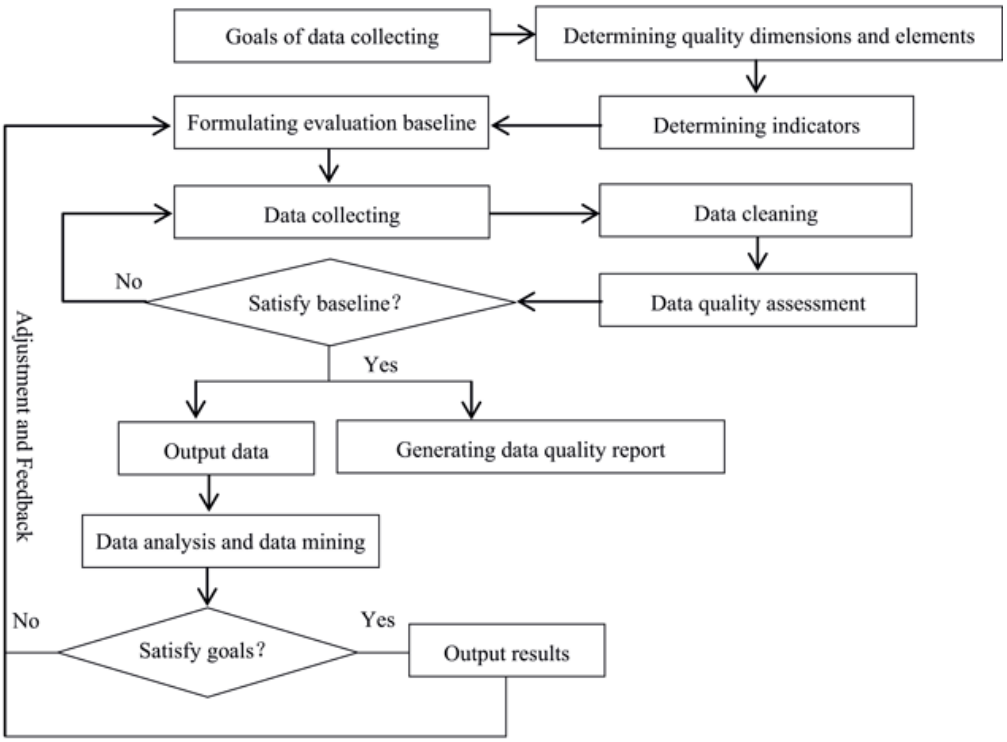


Figure 2: DQ assessment process [9]

Owing to the complexities of big data and the challenges seen above in the 4 Vs, application of such a process to raw data quickly becomes costly and difficult. While this may provide a useful model at some point going forward, it is unlikely to be a practical approach for PEDASI at this time.

5.2 Metadata quality assessment using Analytic Hierarchy Process

The Analytic Hierarchy Process (AHP) was developed by Thomas L. Saaty in the 1970s as a structured technique for making complex decisions [10, 11]. It has been applied in many domains, including recently in evaluating metadata quality in open data portals [5]. Here it is used to integrate DQ dimensions and end-user preferences. It achieves this by using a set of criteria related to data openness and transparency (cf., Figure 3) to evaluate quality metrics across a matrix in combination with a set of DQ dimensions and sub-dimensions.

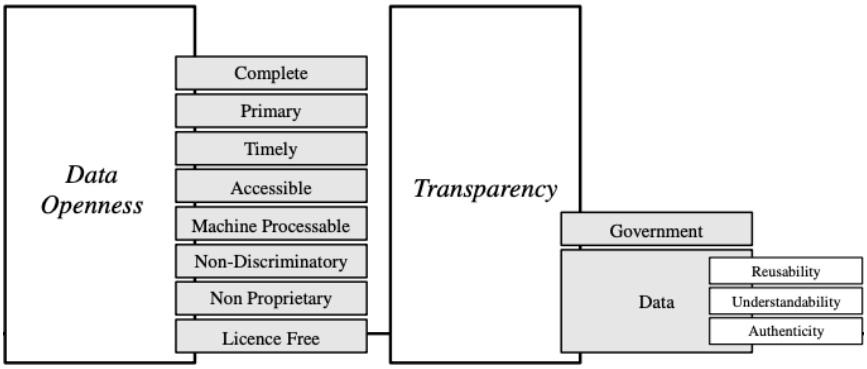


Figure 3: Data openness and transparency indicators; adapted from [5]

These DQ dimensions and sub-dimensions (cf., Table 4) may be applied to metadata with subjective weightings set according to end-user preferences to enable automated and scalable assessments. By combining the above (greatly simplified) criteria with quality dimensions and weighted metric values in a matrix (cf., Table 5), a comparative assessment of DQ can be achieved.

Dimensions	Sub-dimensions		Description	Metric
Existence (Qe)	Access	Qe(acc)	The extent to which access information for resources is provided	%
	Discovery	Qe(dis)	The extent to which information helping to discover/search datasets is provided	%
	Contact	Qe(con)	The extent to which information helping to contact the dataset owner is provided	%
	Rights	Qe(rig)	The extent to which information about the dataset's or resource's license is provided	%
	Preservation	Qe(pre)	The extent to which information about the resource's format, size or update frequency is provided	%
	Date	Qe(dat)	The extent to which information about the creation and modification dates of metadata and resources is provided	%
	Temporal	Qe(tem)	The extent to which temporal information is provided	%
	Spatial	Qe(spa)	The extent to which spatial information is provided	%
Conformance (Qc)	AccessURL	Qc(acc)	The extent to which the values of access properties (HTTP, URLs) are valid	%
	ContactEmail	Qc(ema)	The extent to which the email contact properties are valid	%
	ContactURL	Qc(ext)	The extent to which the URL/HTTP contact properties are valid	%

Table 4: Quality dimensions derived from DCAT [5]

	DateFormat	Qc(dat)	The extent to which the date information is specified using a valid date format	%
	License	Qc(lic)	The extent to which the license maps to the list of licenses given at Open Knowledge International (2017)	%

	FileFormat	Qc(fil)	The extent to which the file format or media type is registered by IANA (1988)	%
Retrievability (Qr)	Dataset	Qr(dat)	The extent to which the described dataset can be retrieved by an agent	%
	Resource	Qr(res)	The extent to which the described resource can be retrieved by an agent	%
Accuracy (Qa)	FormatAccr	Qa(for)	The extent to which the specified file format is accurate	%
	SizeAccr	Qa(siz)	The extent to which the specified file size is accurate	%
Open data (Qo)	OpenFormat	Qo(for)	The extent to which the file format relies on an open standard	%
	MachineRead	Qo(mac)	The extent to which the file format can be considered as machine readable	%
	OpenLicense	Qo(lic)	The extent to which the used license complies with the open definition	%

Table 4 (continued): Quality dimensions derived from DCAT [5]

The procedure for such DQ assessment begins with determination of the presence or absence of metadata attributes demonstrated in Table 5 in combination with the quality dimensions above, with relative values being generated for each quality metric. The matrix uses a two-level scale (+, ++) to indicate whether the metrics slightly or strongly influences the criteria of interest (the exemplar in Table 5 being derived from the e-Government Openness Index proposed by [18]). Assessment would be required to determine appropriate metric weights to measure DQ within PEDASI.

This model provides a practical (and perhaps supplementary) approach to the DQ framework described in D1.1 Specification for PEDASI data quality indicator support, as the matrix of weighted valuations offers a model for realising automation of the framework versioning and provenance capabilities.

The model's authors further recommend greater engagement with data providers [5] by:

- Providing a schema/ontology/model for their metadata that maps to standards such as DCAT or DCAT-AP (DCAT Application Profile for data portals in Europe);
- Deriving metadata values directly from the data in an automated way (e.g., file size, format, availability);
- Restricting certain metadata values to a predefined list of options (e.g., for license descriptions, field formats);
- Checking/validating the conformance of certain metadata values (e.g., URLs, emails).

Key criteria		Existence (Q*e*)								Conformance (Q*c*)						Retriev. (Q*r*)		Accuracy (Q*a*)		Open Data (Q*o*)		
		Qe (acc)	Qe (dis)	Qe (con)	Qe (rig)	Qe (pre)	Qe (dat)	Qe (tem)	Qe (spa)	Qc (acc)	Qc (ema)	Qc (ext)	Qc (dat)	Qc (lic)	Qc (fil)	Qr (dat)	Qr (res)	Qa (for)	Qa (siz)	Qo (for)	Qo (mac)	Qo (lic)
Data openness	Complete	++	++	++	++	++	++	++	++								++	+	+		++	
	Primary						++						++		+			+		++	++	++
	Timely					++	++	++											+			
	Accessible	++			+					++						++	++	+		+	+	+
	Mac.Process.					+									+			+	+		++	++
	Non-discrim.				++									++								++
	Non-prop.					+									+			+	+	++	++	
	License free				++									++								++

Transpar -ency	Reusability														+			+	+	++	++	+
	Understand.		++																			
	Authenticity			++							++	++										

Table 5 Decision matrix evaluating key criteria against quality dimensions; adapted from [5]

6. CONCLUSION AND NEXT STEPS

Evaluating metadata using AHP offers an achievable and reliable approach to improving the quality of data surfaced by PEDASI over the near to medium term. Support for capturing such metadata and data quality frameworks used to assess DQ against that metadata - as outlined in D1.1 - already exists within PEDASI. A key next step will be to render an initial set of questions based on this AHP approach. Initial answers to these questions can be elicited initially from manual input from Data Providers within PEDASI, and in time, mechanisms can be developed to determine at least some of these responses automatically. Indeed, the framework for implementing validation and other similar checks through PEDASI's (as depicted in D3.1) has already been implemented, and provides a solid foundation for automating DQ checks where it makes sense to do so.

Of course, the rendering of such a framework and its implementation within PEDASI will need to be developed, validated and tested. This to a large extent will be explored in more detail through support for a 'Special Interest Group' within the PETRAS Programme with the aim to inform each of these activities. Obtaining and using suitable test cases from across PETRAS2 will also help to inform and guide these efforts with input from both industrial partners and wider academia focussed on IoT data related projects. It is very evident from early research findings there is considerable development required in order to demonstrate programmatic methods for demonstrating data quality e.g. through API provided metadata and profiling of data owners/providers.

It should be noted that PEDASI does not (currently, at least) *generate* data. It rather provides a brokerage or platform for data interoperation from disparate data producers to data consumers. As such, the onus remains with data producers to guarantee their own DQ through policies and processes upstream of PEDASI. However, a primary potential benefit of this this work is to inform and assist those producers on practices and methods to improve their DQ and related processes to improve their data, and at a minimum, make it fit for purpose.

7. REFERENCES

[1] Schueler, M., Cox, A., Yuan, S., Crouch, S., Graham, J. and Hall, W. (2019). "From Observatory to Laboratory: A Pathway to Data Evolution". *Living in the Internet of Things: Realising the socioeconomic benefits of an interconnected world*. 1–2 May 2019.

[2] Marnewick, Carl, (2017). "The reality of adherence to best practices for information system initiatives", *International Journal of Managing Projects in Business*, Vol. 10 Issue: 1, pp.167-184. https://doi.org/10.1108/IJMPB-05-2016-0045

[3] Batini, C., Cappiello, C., Francalanci, C., and Maurino, A. (2009). "Methodologies for data quality assessment and improvement". *ACM Comput. Surv. 41*, 3, Article 16 (July 2009), 52 pages. http://doi.acm.org/10.1145/1541880.1541883

[4] Wang, X., Tiropanis, T. and Schueler, M. (2018). "Metadata Standards, Best Practice and Recommendations for Sharing IoT Datasets". *PETRAS Internet of Things Research Hub Deliverables*, 20 July 2018 (version 2.0).

[5] Kubler, S., Robert, J., Neumaier, S., Umbrich, J., and Le Traon, Y. (2017). "Comparison of metadata quality in open data portals using the Analytic Hierarchy Process". *Government Information Quarterly*. https://doi.org/10.1016/j.giq.2017.11.003

[6] IMF Statistics Department. (2003). "Data Quality Assessment Framework". Factsheet. URL: https://unstats.un.org/unsd/dnss/docs-nqaf/IMF-dqrs_factsheet.pdf

[7] DAMA UK Working Group. (2013). "The Six Primary Dimensions for Data Quality Assessment". White Paper. October 2013. URL: https://www.whitepapers.em360tech.com/wp-content/files_mf/1407250286DAMAUKDQDimensionsWhitePaperR37.pdf

[8] Batini, C., Rula, A., Scannapieco, M., and Viscusi, G. (2015). "From Data Quality to Big Data Quality". *Journal of Database Management*. Volume 26 Issue 1, January 2015. Pages 60-82. https://doi.org/10.4018/JDM.2015010103

[9] Cai, L. and Zhu, Y. (2015). "The Challenges of Data Quality and Data Quality Assessment in the Big Data Era". *Data Science Journal*, 14, p.2. http://doi.org/10.5334/dsj-2015-002

[10] Saaty, T. L. (1977). "A scaling method for priorities in hierarchical structures". *Journal of mathematical psychology*, 15(3), 234–281. https://doi.org/10.1016/0022-2496(77)90033-5

[11] Saaty, T. L. (1980). *The Analytic Hierarchy Process*. New York: McGraw-Hill.

[12] Cichy, C. and Rass, S. (2019) "An Overview of Data Quality Frameworks". IEEE Access, volume 7 (February 2019). https://doi.org/10.1109/ACCESS.2019.2899751

[13] Sidi, F., Panahy, P., Affendey, L., Jabar, M., Ibrahim, H. And Mustapha, A. (2012) "Data Quality:A Survey of Data Quality Dimensions". *2012 International Conference on Information Retrieval & Knowledge Management*, 13-15 March 2012. https://doi.org/10.1109/InfRKM.2012.6204995

[14] Scannapieco, M., Missier, P. and Batini, C. (2005) Data Quality at a Glance". *Datenbankspektrum*, Vol. 14, 2005. URL: http://dc-pubs.dbs.uni-leipzig.de/node/799

[15] Pipino, L., Lee, Y. and Wang, R. W. (2002) "Data Quality Assessment". *Communications of the ACM*, April 2002/Vol. 45, No. 4ve. https://doi.org/10.1145/505248.506010

[16] Katal, A., Wazid, M., and Goudar, R. (2013) "Big Data: Issues, Challenges, Tools and Good Practices". *Procedures of the 2013 Sixth International Conference on Contemporary Computing*, Noida: IEEE, pp 404–409.

[17] International Organization for Standardization. (2016) "ISO 8000-61 Data Quality – Part 61: Data quality management: Process reference model". First edition 2016-11-15.

[18] Veljković, N., Bogdanović-Dinić, S., and Stoimenov, L. (2014). "Benchmarking open government: An open data perspective". *Government Information Quarterly*, 31(2), 278–290.

8. APPENDIX

8.1 D1.1 Specification for a Data Quality Framework

8.2 D2.2 Specification for Volumetric Security Rate Limiting

8.3 D3.1 Specification for data validation & detuning

D1.1 Specification for PEDASI data quality indicator support

1 Overview

PEDASI currently supports globally defined metadata fields that can apply to data sources. This model will be extended to support the management of data quality indicator questions for Data Providers, and support implemented for PEDASI Central Administrators to specify tiered assessment frameworks that classify these questions and their answers within specific data quality 'maturity levels' within an assessment framework. The design and implementation will consider the need for the framework to be extensible allowing for future granularity within the quality framework e.g. has metadata schema allowing assessment of completeness.

These enhancements will enable PEDASI to automatically evaluate data quality for a given data source against this framework and present an assessment report that includes recommendations to Data Providers to improve data quality. Assessment scoring will also be displayed for data sources within the PEDASI service for users.

2 Data quality support specification

2.1 Enhance PEDASI metadata support

Firstly, the PEDASI metadata model will be reorganised and extended beyond the implicit metadata field/value system to support the following concrete categories of metadata:

- Operational: core metadata which are specified by Data Providers and used by PEDASI to support its functional capabilities as part of standard operation of the system, e.g. licence, metadata schema, application profile characteristics (possibly includes too much detail?), data format.
- Annotation: arbitrary metadata, aimed at providing further value to the description of the data source, which Data Providers can use to add metadata specific to their service, e.g. API data query parameters

In order to sufficiently support quantitative and qualitative answers to specific questions, PEDASI's support for metadata types will also be extended beyond strings to cope with integers (for specifying answers to quantitative questions) and JSON data (as strings, for specifying things like JSON schemas, useful for application developers who want to know and validate data source responses to queries made via PEDASI's Applications API).

A new PEDASI Django View interface will be created for Data Providers to provide answers to these data quality questions. Throughout the development cycle explore opportunities to surface content programmatically understanding current technology capabilities and how these will affect efficiency, resilience and service dependencies e.g. accessed published standards.

These enhancements will enable questions regarding data quality of data sources to be added by PEDASI Central Administrators, to be filled in by Data Providers.

2.2 Implement support for data quality assessment criteria

Answers to data quality indicator questions will be assessed against a tiered assessment framework, which classifies each question within separate 'maturity levels', as evolved, within the framework. By indicating compliance with all given criteria for a maturity level, a data source has effectively reached that level.

Also important is support for framework versioning, aligning with development of PEDASI (and published via GitHub, so as assessment frameworks are developed and evolve over time, they are able to supercede previous versions whilst retaining records of previous versions for criteria provenance. This mechanism will also inherently support the use of completely different assessment frameworks, which has the potential for data sources to be assessed and scored against multiple different quality assessment frameworks as industry standards emerge over time. See Figure 1 for an example rendering of this model.

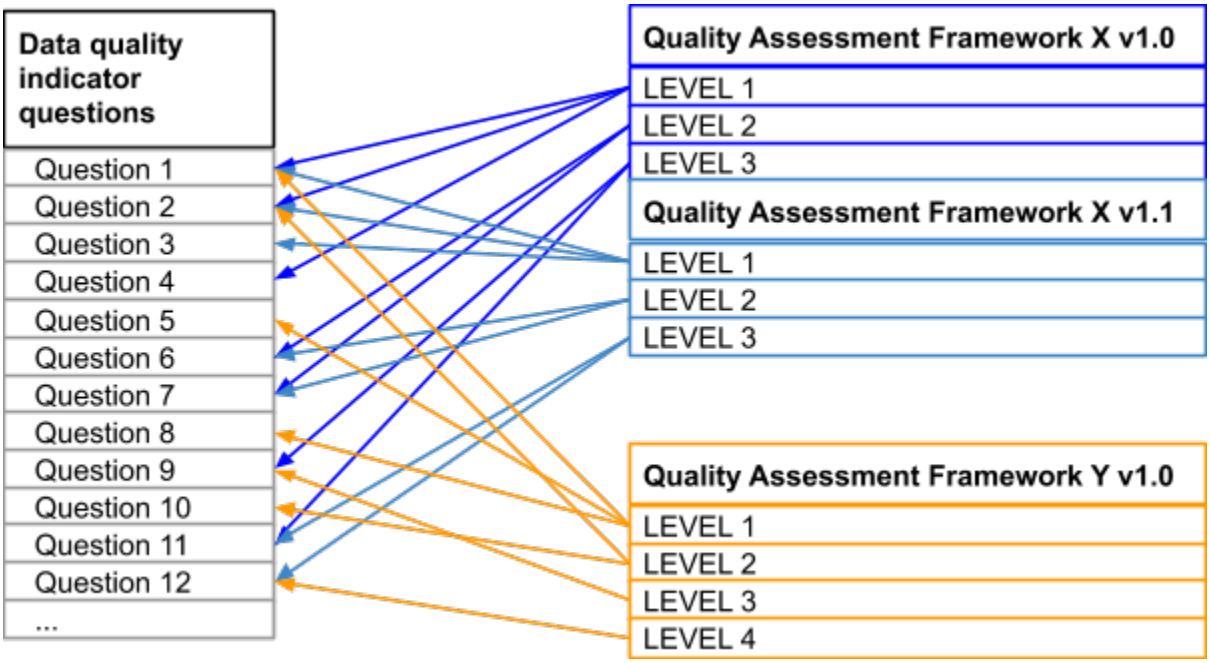


Figure 1. Example usage of the proposed way of mapping data quality indicator questions to versioned quality assessment frameworks. In reality, there will likely be far more questions defined and used in any framework

The assessment framework will be managed through the PEDASI administrator interface by the Central Administrator, mapping each maturity level to a specific set of data quality questions and associated with a specific assessment framework version.

Initially, Data Providers who provide affirmative answers to all questions within a given framework level will allow them to be classified at that level. However, this will be implemented within the context of a weighted system of assessment, with each question being given a specifiable (but initially equal)

weighting. This will allow future expansion of assessments to provide more complex criteria as required, taking into account weighted answers to questions within non-linear scoring systems when assessing the maturity level for a data source. For example, a criteria framework could have a specific question which is critically important, having a much higher weighting than other questions, and if affirmatively answered contributes significantly to reaching a given threshold score of data quality for a given quality level.

2.3 Implement data quality assessment and reporting

An assessment of a data source's maturity level based on their answers to questions are automatically calculated by PEDASI. This calculation is based on their answers to the data quality questions as they apply against the assessment framework.

A separate Django View interface for Data Providers will be created to present an automatically generated [status] report based on their maturity level and answers to framework questions - possibly represented as a status page for PEDASI brokered data sources. This will include recommendations for improvement to successive levels for criteria to which they are not yet compliant.

Within the wider PEDASI interface, the calculated maturity level for a data source will be represented as a number of stars according to that level, and Data Provider answers to assessment framework questions will be presented in the data source details page.

3 Required policy input

The following will be required in terms of policy to feed into the technical implementation:

1. What is the initial set of data quality indicator questions to ask Data Providers? (See Appendix A)
    - The ODI has published the Open Data Maturity Model[1], which is certainly worth considering for this at some level. It deals with processes in regard to how organisations deal with data (which may or may not be the core point), but has lots of other relevant aspects too described within each level.
    - Also perhaps worth considering this report[2] (there are many other similar sources) which deal with 6 dimensions of the actual quality of data, e.g. completeness, uniqueness, timeliness, validity, accuracy, and consistency. This offers a basic status initially with the potential for granularity over time based on validation against emerging metadata or application profile standards.
2. What is the first Quality Assessment Framework, which provides a mapping for each framework maturity level to a set of data quality indicator questions?
    - What is an overall summary 'term' or 'data property' that represents what it means to achieve a level, e.g. for level 1, it could be "Data is publicly available", with level 1

[1] https://theodi.org/article/open-data-maturity-model-2/
[2] https://www.whitepapers.em360tech.com/wp-content/files_mf/1407250286DAMAUKDQDimensionsWhitePaperR37.pdf

mapping to a series of questions that deal with data availability, on topics such as licencing, the level of public access via their API, etc.

4 Implementation

Support for data quality levels has been implemented into PEDASI as described above.

The implementation deviates from the specification slightly in two areas, in order to achieve a working product more quickly. These deviations are:

- Metadata items do not distinguish between data types (e.g. string, integer, JSON). Since entry is via a text box, it was simpler to accept the text representation of the data and transform on use if necessary. Since a particular quality metric is based upon the presence or absence of a metadata item (not its content), this does not negatively affect this piece of work.
- The metadata item entry page has been merged with the recommendations page into a single page which displays each metric at each quality level. This page makes it clear to data providers which metrics would be necessary to progress to a higher quality level since the levels are listed in order.

Appendix A: Data Quality Indicators: How Should we Specify them?

For each question, e.g.:

- The question text.
- The expected data type of the answer, e.g.
    - Yes or no, e.g. "Is your data source populated from where it is generated via a fully automated process?"
    - Selectable from a list, e.g. "How is your data source populated with data?", with possible answers "Manually, Partially automated, Fully automated"
    - General text, such that providing an answer leads to an affirmative response, e.g. "What is your data retention policy?". Note where Data Providers can provide links to specific information, this type of question is preferred where possible over "Do you have a data policy?", since it provides more useful information that also fulfils the criteria.
- The validation condition to assess if the answer successfully fulfils the criteria underlying the question, e.g. in terms of the questions above:
    - "Yes"
    - "Fully automated"
    - (That it is non-empty)
- Some textual recommendation for improvement for each question, where they don't successfully pass the criteria, together with any links to online resources which provides guidance.
- Does your data/metadata conform to a schema or application profile standard (possibly a prompt or advice in the user guide initially (hardened as standards maturity is achieved)

D2.2 Specification for volumetric security in PEDASI

1 Overview

Volumetric security is a method in data security which aims to protect against an attacker acquiring a complete dataset over multiple requests, or assembling a composite dataset which provides more information than the attacker is permitted to have. It is in many ways similar to rate limiting, but must take note of the volume of data being transmitted rather than just the number of requests. It applies both to requests sent to a single data source, but also requests across multiple data sources.

PEDASI currently logs data accesses by users and applications using the PROV-DM specification, but does not currently record the data itself. In addition to this long-term logging, PEDASI will maintain a short-term record of the content of every response in order to measure the volume of data being accessed by each user and through each data source. These response logs will be used to produce a report of the volume of data being accessed by users and identified transgressions. How this report is acted upon will be dictated by policy decisions taken as a further part of this project.

2 Volumetric Security Specification

2.1 Storage of data accesses

In order to provide volumetric security reports, PEDASI will be required to inspect the responses being returned from requests to data sources, both internal and external. This will require that PEDASI be able to accept a schema which describes the structure of the response (e.g. JSON Schema) and makes it possible to identify what constitutes a single record.

Since parsing the response to count the records is likely to be slow relative to the current response time, it it likely that the parsing will have to happen out of process of PEDASI. To achieve this, it is proposed to temporarily store every response so that they may be read by another process. Responses will be identified by a hash of their content so it is not necessary to store responses which are duplicates. This may also form the basis of response caching in a future piece of work.

The existing PROV tracking will be extended to include the content of the request received by PEDASI and a link to the content of the response. This means that the new volumetric reporting functionality would not need to maintain a redundant set of records.

It is possible that using the existing PROV records for this may prove to be too rate-limiting, in which case a database table will be created to store each request that PEDASI receives, containing: the user who performed the request, the content of the request and a link to the content of the response. This would have some redundancy with the information stored in the PROV records, but would be more efficiently queryable.

2.2 Volumetric reporting

The reporting component will be run as a separate process, which may be ultimately run on a separate host. This process will take all PROV records (or in the contingency case, the record access table) within a certain time period e.g. the last 24 hours, and inspect the volume of data which was returned to the user as a result of each request. The stored responses will be parsed using the schema provided by the data provider so that the number of data records can be counted.

Initially, the number of rows requested by each active user against each data source will be reported. A number of aggregate metrics may be created later, depending on policy decisions around the purpose and consequences of the reporting, e.g. to support identification of transgressions.

3 Policy

This deliverable requires several policy decisions to be made:

1. What metrics should be reported in the volumetric report?
2. What levels of transgression exist? (e.g. severity / frequency)
3. What are the consequences for different levels of transgression?

4 Implementation

There is significant overlap in the implementation work required here with D3.1[3] (validation and detuning) in that both require the presence of a data pipeline and a degree of caching infrastructure. Both this pipeline and the necessary caching have been implemented, but the volumetric reporting is not yet present.

[3] https://docs.google.com/document/d/1WOgCdlKVR-2Efvmy80RYfY2nerDNxTg4eFthaDsuq20

1 Data Validation

1.1 Overview

As requests for data from data sources are made either via PEDASI's web interface or via its Application API, there is a need for PEDASI to automatically validate data being returned from data sources to ensure it is syntactically well-formed and conforms to a structure expected from that data source. This will form a foundation for additional complex features, such as advanced detuning methods and volumetric security, where schema-level descriptions of data are required to operate.

To support validation, the originating Data Provider will optionally be able to supply a data schema for that data source, either when it is created or otherwise during its lifecycle. Where a schema is provided the data quality indicator assessment and rating for that data source will be correspondingly increased. This will be achieved within PEDASI by extending its architecture to process inbound data from data sources within a pipeline, with data validation forming one step within this pipeline.

Data schema validation will initially be included for two of the most common high-level data structures typically returned from API-based data sources, JSON and CSV. Future support for other data types will be implemented as required by new data sources, where data schema validation is viable and desired for that data type, aligning with wider discussion on API surfaced standards.

1.2 Data Validation Specification

PEDASI's core architecture will be extended to include an inbound data processing pipeline consisting of a number of sequential tasks applied to incoming data (i.e. from data sources in response to a request). This pipeline capability will form a foundation to optionally extend PEDASI's capabilities for inbound data checking and even data transformation tasks, evolving as data owner policy and data roles permit (i.e. in terms of acting as a data processor or a data controller, since any data transformation may consequently elevate obligations for data owners which needs to be considered). This would potentially allow data sources to have their own pipeline configuration, built from a pool of data validation and transformation pipeline 'connectors'. See Figure 1.

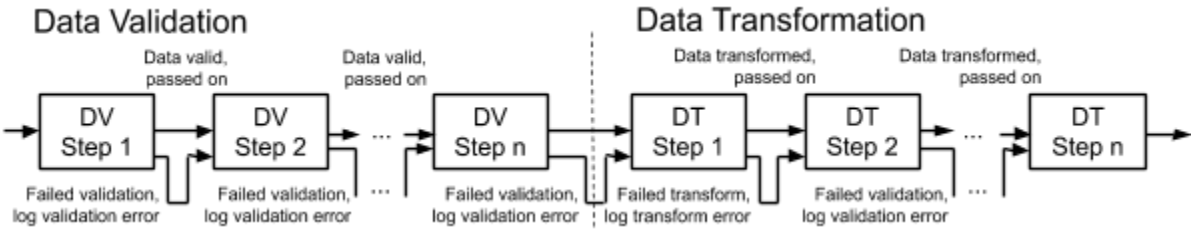


Figure 1. Inbound data processing pipeline architecture

Data schema validation support will be developed as a DV step within this architectural extension. Data Providers will be able to optionally specify a data schema for a data source, either for JSON or CSV,

which if present is then used to validate all incoming data. The schema format will support the capability to specify multiple valid potential data schemas as necessary, depending on what represents valid return data from a data source.

There are a number of ways to handle a data validation failure. At a minimum, any transgression will be logged within a Data Provider-specific log, but could also deliver the received data to the requestor (with an optional warning), or just deliver an error code instead. The selected method will be to deliver the received data with a validation warning that is also logged. PEDASI's Applications API will be revised to accommodate and return these warnings for application developers, similarly for the API Explorer these warnings will be displayed.

In time, this pipeline architecture will enable support for more complex features such as complex semantic data checking (as a DV step), and pseudo-anonymisation (as a DT step).

1.3 Recommendations for Data Validation Implementation

PEDASI's data connector architecture will be extended to encompass a configurable data pipeline. In addition to current data request-style connectors that request and retrieve data from a data source, other data processing connectors that can perform arbitrary tasks will be implementable that conform to the same connector interface.

The PEDASI global metadata for data sources will be extended in two ways:

- Additional metadata fields to hold the configuration of a number of implemented data processing connectors (e.g. validation, and any others that are implemented in the future) for execution in a given sequence on inbound data.
- An additional data source metadata field for Data Providers to specify a JSON or CSV schema. This could be optional at the global level, or connectors could have their own configurable metadata fields with this as one of them for this data validation connector.

Based on this configuration, PEDASI will then manage the execution of this pipeline for any given request to that data source.

The pipeline metadata for a data source will be configured through PEDASI's web interface, offering the choice of adding, removing, editing, and reordering pipeline stages, each associated with an available connector. This will be achieved by replacing the 'Plugin name' field (used to just specify the data connector) in the 'Edit data source' view with an inbound pipeline editor.

When a data source is first created, a default pipeline will simply contain the data connector for that API endpoint type as a single DV step.

Based on this architecture, a new data validation connector will be implemented to validate incoming data from a data source request against this schema where it is specified. In addition to the standard Python

JSON Schema and Fast JSON Schema packages, Frictionless Data provides a set of open source Python packages[4] for CSV and JSON validation which will also be explored for suitability.

Data validation support within this revised architecture will be implemented for the IoT UK Nation Database as an initial use case, for which a PEDASI data connector already exists.

1.4 Policy

This aspect requires the following policy support:

1. As part of our policy for supporting data sources for Data Providers, guidance should be provided for specifying a data schema within PEDASI's data source metadata.
2. Data connectors could be specific to a “Trust Framework” application therefore operating under a defined policy framework clarifying complex arrangements between data management roles.

2 Detuning

2.1 Overview

To mitigate against inefficiencies in users and applications requesting data from data sources that hasn't changed, and to provide some level of obfuscation in determining the exact state of data from a data source at any given precise time, a method of 'data detuning' will be employed within PEDASI.

There are a number of approaches. Firstly, there is a basic solution that provides a data access interval as a recommendation for application developers that isn't enforced, and there are also two key development paths for more comprehensive detuning solutions that effectively (but not obviously) enforce data access intervals.

2.2 Detuning Specification

A first basic solution is to allow Data Providers to specify a recommended data request interval for accesses to a data source that can chosen to be honoured by an application developer. This wouldn't be enforced. For obfuscation purposes, the Data Provider can specify an interval that is appropriately larger than any actual interval at which any of its constituent data is updated, such that it effectively masks those real update intervals.

Beyond this basic approach, there are two more complex solutions that employ local data source request caching as a mechanism to provide a detuning capability, as well as the obvious performance benefits of using a cache. With this approach:

- Data source access requests and responses are stored locally on disk with access time data, hence providing the details necessary for caching the responses to such requests.
- Detuning becomes an extension to this, whereby desired time access intervals designate the cache expire time for a given request/response pair. Upon expiry, the cache response value for that request is refreshed via an actual API request call to the data source (which may or may not be different). Detuning is realised

[4] https://frictionlessdata.io/software/

2.2.1 Detuning using a Fixed Interval

The first more complex solution has two alternatives based on the basic initial method described above:

1. Fixed Detuning and Cache (FDAC): the data source's request interval as described above is used to determine cache expiry for all data requests.
2. Granular Fixed Detuning and Cache (GFDAC): a more complex extension to FDAC acknowledges that the data source may have different update interval characteristics for subsets of the data it provides. In this case, the data source associates different update intervals to different 'sub-schemas' of the data it provides. An additional check is made when any response is received to determine if that response matches one of these sub-schemas, and if so, apply its corresponding time interval as the cache expiration value for that request/response.

2.2.2 Optimised Detuning using an Adaptive Interval

The second complex solution provides an adaptive caching strategy, which we'll call Adaptive Frequency Detuning and Cache (AFDAC), which stores and adapts the caching interval for requests over time. This approach makes use of a data source-level request interval as the default minimal expiration for any request, and actively monitors responses to determine if they have changed. Essentially:

- If the response data has not changed, the interval for that cached request/response is increased for caching expiration purposes.
- If the response data has changed, the interval for that cached request/response is reduced for caching expiration purposes, but not below the minimum default expiration for that data source (hence maintaining a requisite level of obfuscation for when data actually changes on the data source).

The amount by which the interval is increased or reduced could be governed linearly or logarithmically.

A more advanced version of this second solution, named Granular Adaptive Frequency Detuning and Cache (GAFDAC), would employ an approach similar to the 'sub-schema' approach described by GFDAC. This solution would maintain separate intervals for identifiable response types according to these sub-schemas. This would allow for more granular adaptive optimisation for subsets of data source data.

2.3 Recommendations for Detuning Implementation

The first basic solution would be implemented by adding a new metadata field to data sources that would be specified by the Data Provider on the data sources' 'Edit data source' view and publicly visible.

For the two more advanced solution types:

- Python caching libraries will be explored for suitability, particularly the python-diskcache file caching library[5] which is very well-established for use in production, well documented, flexible, and has a comprehensive API for customising cache behaviour.
- For the granular approaches GFDAC and GAFDAC, an additional metadata structure will be provided (instead of a single field) to specify sub-schemas with their associated intervals, updating the 'Edit data source' view accordingly to allow adding, removing, and updating of sub-schema/interval pairs for a data source. Additional metadata fields would be used to hold the current interval value, with all this additional metadata for a data source only visible to its Data Provider. This would make use of data validation technologies as described in the validation section.
- When a data source is deleted by a Data Provider, all cache data related to that data source is deleted from the cache to ensure any residual data from that data source cannot be retrieved.

2.4 Policy

This aspect requires the following policy support:

1. As part of our policy for supporting data sources for Data Providers, guidance should be provided for specifying suitable data schema and desired time intervals (depending on which approach is selected) within PEDASI's data source metadata.

3 Implementation

The data validation and pipeline component of this specification was implemented as described above, with the caveat that the external API call has not yet been converted into a pipeline component. There were two models which would have been valid for constructing the pipeline, each with distinct advantages and disadvantages.

The model adopted was an iterative approach, whereby data is processed by each pipeline stage in turn. Stages are able to prematurely terminate the pipeline in cases such as a validation error, or returning data from a cache instead of with a call to an external API. This approach was selected due to being the simpler of the two to document.

The alternative was a recursive approach whereby each pipeline stage would request data from the next, ending with a call to the external API and data being passed back up the chain. As data was being passed back, the transformations would be applied.

Significant work towards frequency detuning has been implemented, with pipeline components for cache storage and retrieval. These components added to a pipeline, provide a proof-of-concept level implementation of FDAC, while the changes necessary to support AFDAC would be relatively small.

[5] https://pypi.org/project/diskcache/