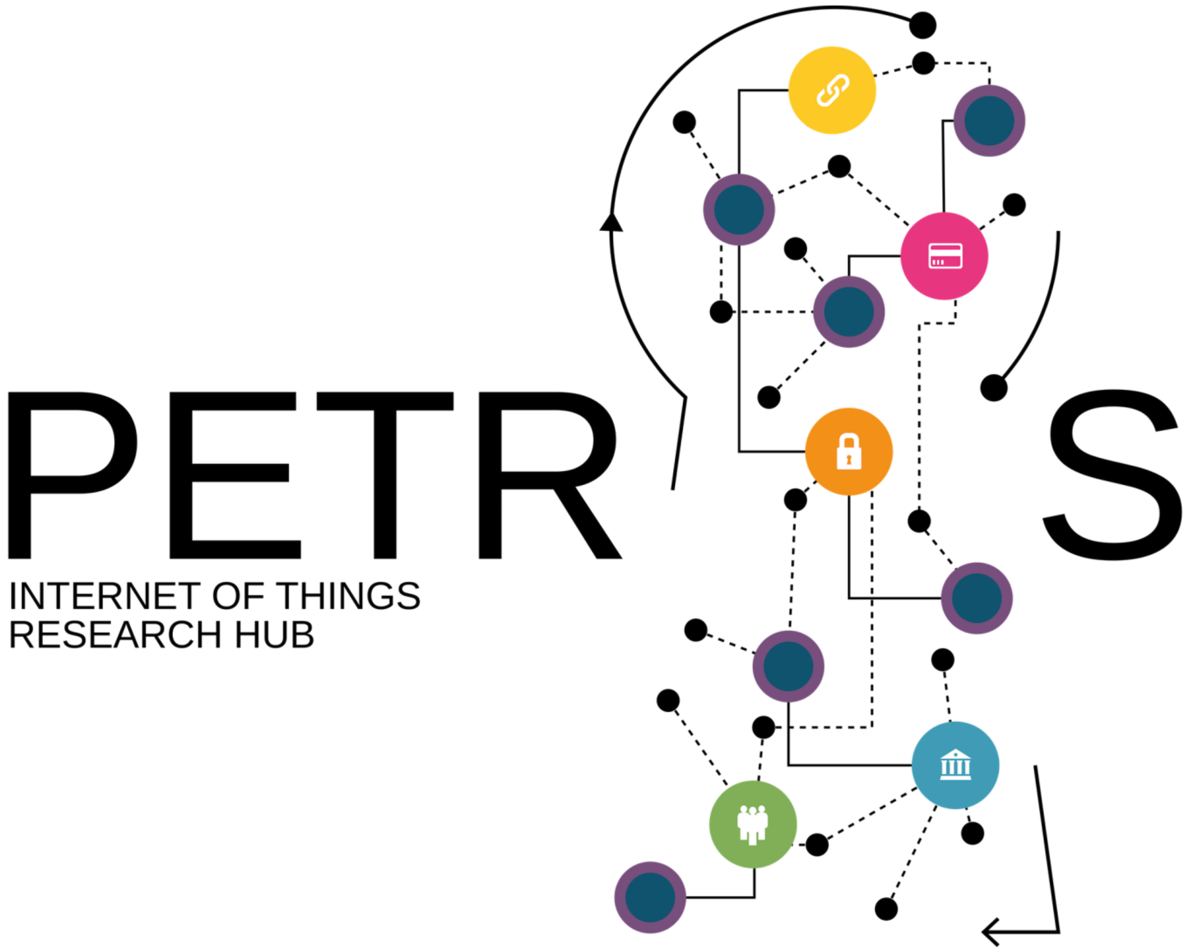


PETRAS-IoT Data Management and Sharing Infrastructure: An Evolution of IoT Observatory (PEDASI)

25 March 2019 (version1.3)

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

Project/Stream title: Harnessing economic value, Privacy & Trust,
Safety & Security

Details of document preparation and issue:

Version no.	Prepared	Checked	Reviewed	Approved	Issue date	Issue status
1.1	Adrian Cox					Draft
1.2	Steve Crouch / James Graham					Draft
1.3	Adrian Cox				15/3/19	Final

PETRAS-IoT Data Management and Sharing Infrastructure: An Evolution of IoT Observatory (PEDASI)

CONTENTS

List of Figures

PETRAS-IoT Data Management and Sharing Infrastructure: An Evolution of IoT Observatory (PEDASI)

1. INTRODUCTION/SUMMARY

1.1 PEDASI – The Concept

In recent years our world has experienced massive and ever-increasing distribution and uptake across a range of sophisticated sensors, “smart devices,” and cyber physical social machines which create a similarly ever-increasing volume of related data. These datasets are distributed across multiple domains and in a variety of formats -- e.g. healthcare, smart cities, smart home, networked mobile and wearable devices, connected industry -- this data may be shared and recombined to offer both opportunities for socioeconomic improvements in our homes, workplaces, and cities as well as risks to privacy, trust, and ethical behaviour in our communities.

This report summarises the development of the PETRAS-IoT Data Management and Sharing Infrastructure (PEDASI) IoT Observatory to a Minimum Viable Product (MVP) and its positioning in relation to cyber-physical and sociotechnical domains, recognising and supporting the evolving understandings and requirements for data trust, privacy, governance, provenance, and quality and the influencing factors these have on each other.

It further describes:

- The PEDASI architecture which provides user, data, and application APIs for securely accessing and combining disparate data sources through a common medium for research purposes while supporting the above requirements. This in turn offers decentralised, accountable edge computing with its advantages of bandwidth and server resource reduction.
- A case study combining and analysing data from BT CityVerve Manchester[1], See.Sense[2] route monitoring cycle lights, and environmental and Met Office data sources. This study explores bicycle routes in Manchester against their actual use in real time by overlaying data from sensors embedded in bicycle lights. It further overlays data regarding roadworks and reported hazards, as well as real-time air quality and weather conditions.
- Future work plans that include potential collaboration with the City of Southampton on similar cycling and transportation-related research, as well as multiple data reporting and analytic projects of interest to the City. We also anticipate further collaboration with BRE, GCHQ, and multiple PETRAS projects.

The approach to the development of PEDASI provides researchers and members of the community with a decentralised, secure platform for accessing a broad range of high-quality data to inform analysis and decision-making, while providing access control, and provenance of all usage and actions performed within its ecosystem.

[1] https://cityverve.org.uk/

[2] https://seesense.cc/

1.2 Evolution of the IoT Observatory

The early IoT Observatory instance iotobservatory.io focussed on a commercial application with the aim to deliver an industry pilot. Architecturally the IoT Observatory provides a registry of resources in the form of applications which consume data outside the observatory. Early in the development it became apparent that the scope would be limited to develop this platform sufficiently to deliver a robust version to a minimum viable product such that it would deliver features addressing the requirements around ethics, trust and data linkages. A revised specification was developed to address these requirements and the PEDASI IoT Observatory project agreed with the PETRAS Board.

2. USE CASE “DELIVERING A DATA DRIVEN SERVICE”

2.1 Platform Feature Requirements

The aim of PEDASI is to deliver a cloud based data brokerage platform enabling the discovery and retrieval of diverse data sources from across the web to enable the provision of data driven services both for research and application development.

The key features the project set out to deliver within the MVP are:

- Searchable catalogue of supported data sources registered by data owners
- Extensible connector interface to external data sources that currently supports HyperCat and IoTUK Nation Database data sources
- Dataset discovery and access via a web interface or via an Applications API
- Queryable and extensible metadata associated with datasets
- Adoption of W3C PROV-DM specification to track and record dataset creation, update, and access within internal datastore
- Internally hosted support for read/write NoSQL datastores
- Functions as a reverse proxy to data sources, returning data from requests exactly as supplied by the data source

Delivery of these features will provide PEDASI with the foundation to support future research into aspects such as Peer-to-Peer networking of multiple PEDASI instances within data trusts, delivering data quality indicators e.g. through analysis of metadata schema and aggregation of diverse data sources in secure environment.

2.2 Development Principles

The following non-functional requirements, centred around sustainability, were adopted throughout development for the MVP:

- Reproducibility and reusability: to ensure that third parties, such as research groups involved in data research and establishing data trust relationships, are able to deploy and maintain their own PEDASI services. This has been accomplished via use of the GitHub source code management system to publicly host the platform's source code, support for automated deployment within development and production environments, and an internal data source hosting model that can ingest and host datasets from common data formats such as Comma-Separated Value (CSV) files and JavaScript Object Notation (JSON).
- Interoperability: an integral part of the product is to ensure it can provide interfaces to external and internal data sources providing data, and applications consuming data. This has been accomplished via well-defined internal and external data Application Programming Interfaces (APIs) and Application API.
- Extensibility: to ensure future connectivity with other third-party data sources and applications, its capabilities need to be inherently extensible. This has been accomplished via architectural extensibility points within the platform that build on these defined APIs, including a connector architecture that enables interfaces to new data sources to be developed and deployed, an automated test suite for testing functional correctness. The hosting of the platform's source code within GitHub enables contributions to the central platform to be contributed from external developers.
- Maintainability: to facilitate low update and issue resolution overhead for future development, the platform needs to be maintainable. This has been accomplished via a well-defined architecture, reuse of community-established technologies, an automated test suite to validate maintenance changes, and well-structured, readable, appropriately commented source code. Again, the use of GitHub enables issue fixes to the core platform to be externally contributed.
- Documented: documentation is a key requirement to support third-party deployment and development. This has been accomplished via the creation of task-oriented, publicly available guides for users, administrators (of the deployed platform, data sources and applications), and application developers.

2.3 Implementation Technical Choices

The MVP product specification outlined a set of platform technical choices that provided excellent matches for these design principles and an agile software development process:

- Operating system - Ubuntu 18.04 LTS: a very popular Linux distribution with vendor update support until April 2023, making it a solid choice for a potentially long-running server platform.
- Data persistence
    - MongoDB: supports storage of JSON documents, ideal for PROV-DM records rendered in PROV-JSON.
    - MySQL Community Server: required for Django's data persistence.
- Web development and hosting:
    - NGINX: an open source, scalable, high-performance and memory-efficient HTTP server to host the PEDASI platform.

  - Python: for server and test suite development the open source Python language was adopted, which is is relatively simple to learn and seeing significant uptake in academia. Version 3 of Python was adopted given Python 2 support will end in 2020.
  - Python Django: an industry-standard, well-established and supported open source web development framework designed to rapidly create prototype and production web applications that operate at scale. Very well documented and supported, and developed using Python.
- System interfaces
  - REST interface for APIs: the industry-standard format for web service APIs. Simple and efficient, it's ideally suited to the straightforward operations required by PEDASI's APIs.
  - PROV library: developed at the University of Southampton, an actively developed open source Python library that supports the PROV-DM specification for provenance records.
  - HyperCat: natively support was developed for this prototype, given existing Python language bindings only supported Python version 2

2.4 PEDASI architecture

The PEDASI architecture contains three key layers, with APIs facilitating the interface between each layer:

1. Data Provider: these are external data sources owned by Data Providers, which interface with PEDASI Core via a Data Connector API, each connector tailored to servicing requests for a given type of data source. Internal data sources are also developed against this connector interface for simplicity.
2. PEDASI Core: the core system that houses the web interface, data/metadata catalogue, and provenance tracking and recording capabilities. It orchestrates requests and responses to/from the web interface and applications to external and internal data sources.
3. Applications Provider: applications that interface with PEDASI Core and make data/metadata requests from data sources, via the RESTful Applications API.

Throughout development, consideration has been given to interoperability with data sources including other observatory platforms and the future potential for peer-to-peer networking of multiple instances of PEDASI. Such capability will be fundamental in supporting the establishment of data trusts, although will form a more detailed deliverable in a future version of PEDASI.

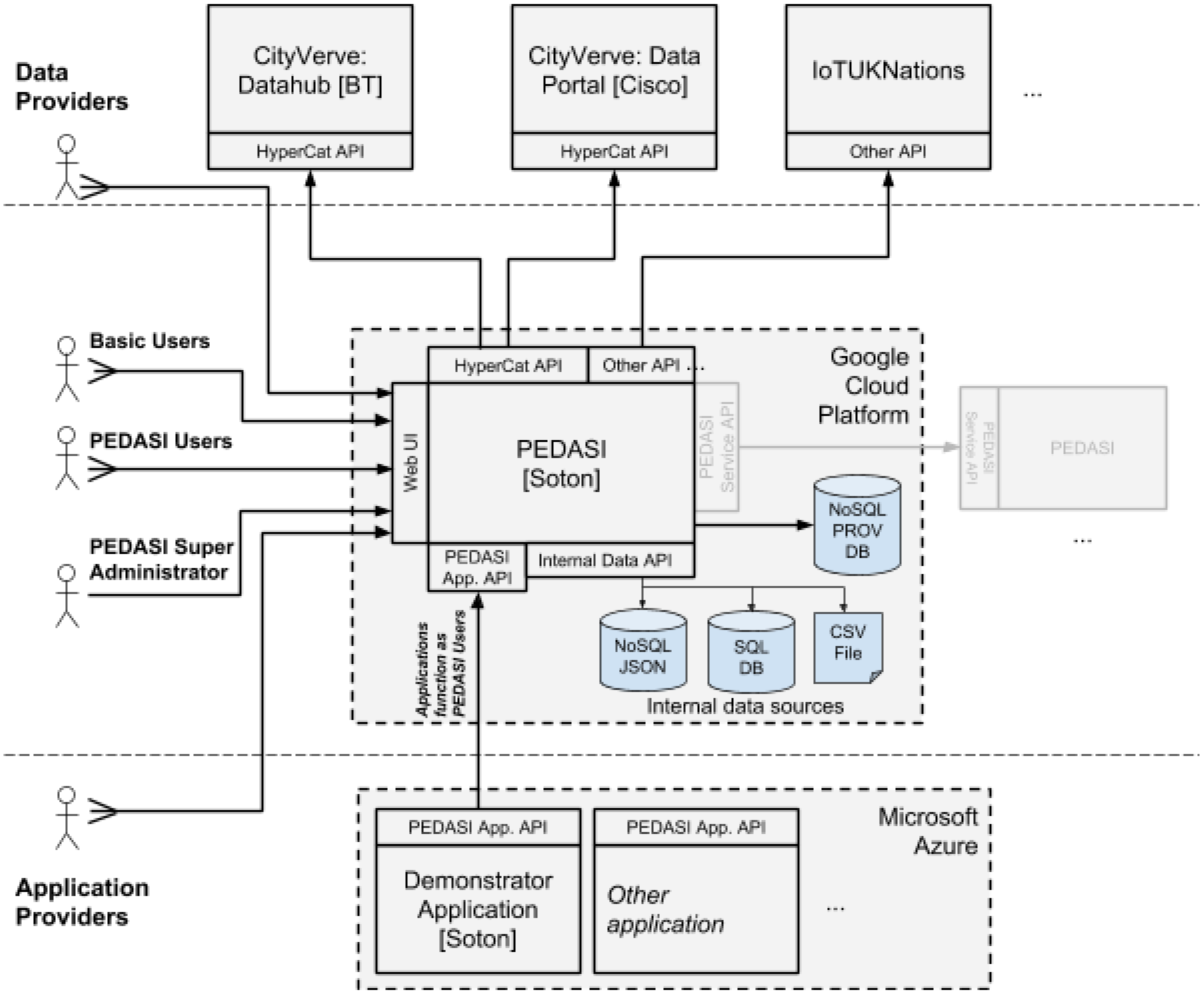


Figure 1.1 PEDASI structure

2.5 MVP Development

Following a core platform development cycle, an agile development process was adopted that enabled evolving platform requirements (such as those emerging from application development) to be taken into consideration where they added justifiable value and did not impact delivery of the MVP.

A suite of tests was co-evolved with the core components, which allowed the development team to ensure that components functioned as intended. As part of the development process support for regular automated deployment within the Google Cloud Platform (the PEDASI service hosting provider), enabled new functionality and other platform modifications to become available, testable, and usable rapidly.

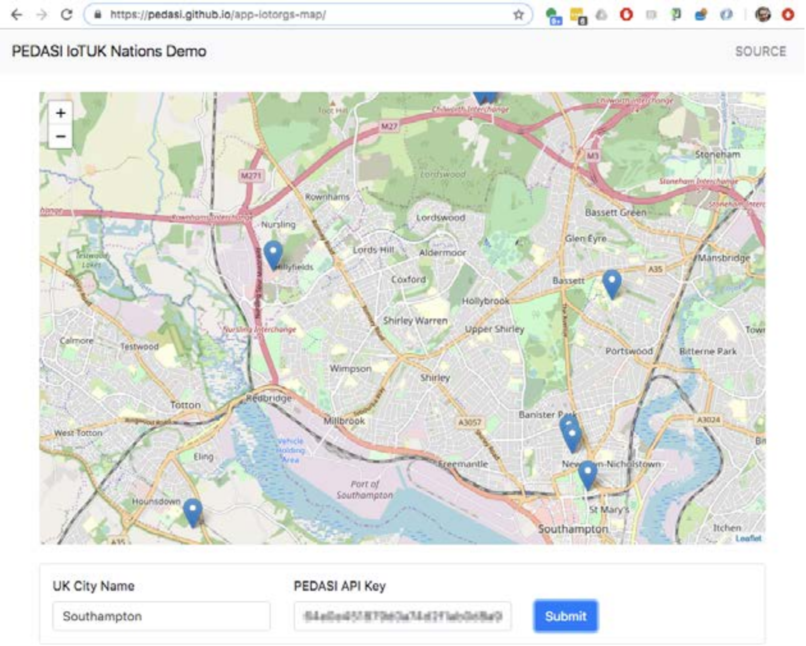
User, administrator (for central administrators and data/application providers), and developer documentation was written to support each user role, which detail how to use, deploy, and configure PEDASI, and how to develop third-party PEDASI applications using its Applications API.

The first publicly available MVP version of the software was released on 28 February on GitHub[3] [4] as open source under the permissive MIT licence, along with public documentation available on readthedocs.org[5].

2.6 Demonstrator – Application development

The development of two end user web applications provided the project team with complete workflows of typical PEDASI user scenarios. This development allowed an overview of the PEDASI IoT Observatory environment to be developed explaining how the architecture works and further provided detailed demonstrations of how it facilitates access to a range of datasets, both secure and open, to deliver a web application.

The first demonstrator application is a simple web application written in JavaScript designed to provide a lightweight example for developers aiming to use the PEDASI Applications API within their own applications. The application uses this API to query and retrieve data from the UK Nation Database for a given location, translates the organisation addresses into geolocation coordinates using a PEDASI internal geolocation lookup service, and finally uses the OpenStreetMap API to render a localised map of the organisations based around that location.



https://pedasi.github.io/app-iotorgs-map/

PEDASI IoTUK Nations Demo

UK City Name: Southampton

PEDASI API Key: [redacted]

Figure 1.2 PEDASI application of UKIoTNation geolocation with postcode look-up

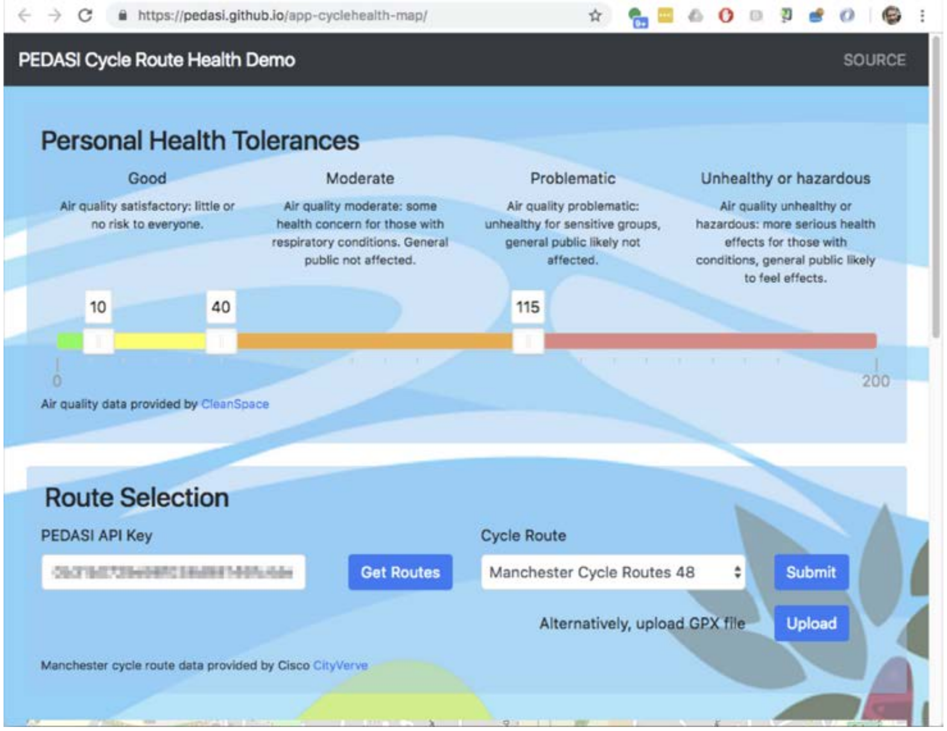
The second demonstrator application uses data accessed via PEDASI and is designed to focus on journey based decision making (cycling and/or walking) with a regard to route based air quality. The application

[3] https://github.com/PEDASI/PEDASI

[4] https://github.com/PEDASI/PEDASI/releases/tag/0.1.1

[5] https://pedasi.readthedocs.io/en/master/

brings together Manchester Smart City cycling routes from the CityVerve project, with an air quality 'layer' provided by CleanSpace[6], a mobile air quality application developed by Sensyne Health, whose data is partially derived from Government reported air quality and city based air quality monitors. Also written in JavaScript, it similarly makes use of the PEDASI Applications API to query and retrieve data (via PEDASI) from these data services, and the OpenStreetMap API to render a map of a cycling route with an overlay of the air quality data along that route.



https://pedasi.github.io/app-cyclehealth-map/

PEDASI Cycle Route Health Demo

Personal Health Tolerances

Good: Air quality satisfactory: little or no risk to everyone.

Moderate: Air quality moderate: some health concern for those with respiratory conditions. General public not affected.

Problematic: Air quality problematic: unhealthy for sensitive groups, general public likely not affected.

Unhealthy or hazardous: Air quality unhealthy or hazardous: more serious health effects for those with conditions, general public likely to feel effects.

Air quality data provided by CleanSpace

Route Selection

PEDASI API Key: [redacted]

Cycle Route: Manchester Cycle Routes 48

Alternatively, upload GPX file

Manchester cycle route data provided by Cisco CityVerve

Figure 1.3 Route based air quality application query screen

Both applications are appropriately documented with deployment guides and commented source code to highlight how the applications operate, and both are available as exemplars from PEDASI's GitHub organisation[7] under the MIT open source licence[8].

2.7 Establishing Trust, Privacy and Quality

Fundamental to the user's perception of "trust" are considerations on quality, privacy (GDPR), governance, metadata and associated standards.

As highlighted in 2.1 the role of the PROV feature is to provide data owners with enhanced trust through the tracking of data source usage through PEDASI. This PROV tracking records updates to the data sources record within PEDASI and each access of the data through PEDASI. Users with access to the

[6] https://www.sensynehealth.com/cleanspace
[7] https://github.com/PEDASI
[8] https://opensource.org/licenses/MIT

PROV records may use them to discover patterns in the usage of the data source, such as relationships with other data sources when two are commonly accessed in combination by the same set of users.

Although not a focus for the development of the MVP, particular attention has been given to the implications for General Data Protection Regulation (GDPR) and how PEDASI may address these regulations from the perspectives of the product and its users. Discussions are currently ongoing to determine the extent to which PEDASI as a Data Processor is subject to these regulations and what it needs to do to ensure compliance. Options are being explored to establish the legal feasibility of delegating these regulatory responsibilities to Data Providers, via support for opt-in policies for Data Providers when registering data sources within PEDASI.

It is recognised that an indication of the license applied to a dataset has significant implications for the future use of the dataset. Such licences may include more permissive licences such as CC0 or OGL to proprietary licences which may contain any restrictions the data owner deems necessary. Some licenses require the application of brand within the data re-use, or may prevent the data from being used for certain purposes, such requirements could have significant implications as to the scope of use. In a first step to addressing this issue PEDASI requires Data Providers to indicate license applied to the dataset. In the future the identification of license programmatically e.g. if embedded in the metadata may be an option.

Improving the application of standardised metadata will be a fundamental consideration for data quality, a) when querying metadata to understand data content b) in the aggregation of common datasets from different sources e.g. Local authority registers of brownfield sites with the register of contaminated land. On the face of it both should be straightforward land records which could be visualised on a map solution. However, with no standard metadata schema the geo coordinates could be recorded in different formats or there could be semantic issues on what constitutes a “site” and so on.

It is these data management challenges that mean PEDASI will only ever be as good as the data made available via the service. The challenge is therefore for data owners and or providers to implement appropriate processes at the data source to improve data they intend to publish as well as ensuring that the data is well described within PEDASI i.e. providing basic curation e.g. license identification and appropriate metadata (schema if known). It is therefore important for PEDASI to account for the natural lifecycle of a data source; from creation, through updates both to the data and its underlying structure, to its decay and eventual failure. To this end the project team have identified some of the key issues of policy for data providers which will be developed into appropriate full guidance documentation in the future.

Fundamental to the end user experience will be metadata quality as this provides a core element of the data source within the browsing and data usage experience. The primary way in which data will be accessed is via the PEDASI API for researchers who have experience in using third-party data. However, the current system does provide a basic metadata query interface providing the basic user the ability to query metadata and explore the structure of data sources. This raises the question of metadata quality, data owners will need consider revised procedures to ensure improved data quality – possibly beyond their own primary users. In the Higher Education sector a number of basic metadata schema are in use, often underpinned by policy obligations e.g. within funder reporting obligations such as use of Dublin Core Metadata Initiative[9] for research outputs. Across industry the use of standard schema is understood to be a far greater challenge with very limited adoption of standard schema with initiatives such as Schema.org[10] having limited impact to date. Establishing the use of metadata schema as an indicator for data quality will therefore have significant implications for system architecture and current data management processes.

What the current development has highlighted are the future data management challenges for data owners as we demand greater trust and data quality within secondary and tertiary data usage.

3. ENGAGING WITH THE PETRAS COMMUNITY AND BEYOND

3.1 Summary

A range of engagement activities were undertaken during the project to inform the development of the MVP and better understand its positioning within the data sharing/observatory landscape. This included project development workshops, industry and public engagement events and networking opportunities.

As the PEDASI project commenced later in the PETRAS Programme, alignment with development on potential related projects was evidently an issue, with many projects either completed or nearing completion therefore severely limiting engagement with relevant programme partners. This is something the project team have now factored into future plans for sector engagement to maximise the benefits from related IoT development projects.

3.2 Software Design Workshop 5th September

This workshop brought together the PEDASI project partners and the PETRAS programme team members to present and refine the initial PEDASI MVP proposal. This included valuable discussions on MVP requirements, the proposed system design and implementation, security considerations, and potential applications as future demonstrators. Outcomes from the workshop helped inform consideration for security features such as anonymisation, role-based access control, and differential privacy, and led to the development of the health cycling application demonstrator. This meeting also formed the basis for follow-on security discussions with GCHQ.

3.3 Industry Demonstrator and "Datathon" Workshop 5th February

Event attendees represented a good cross-section of interested parties, including; PEDASI industry partners, external industry, and members of the developer community.

Demonstration of the PEDASI data brokerage Minimum Viable Product (MVP) provided the community with a good insight into the development achieved during the PEDASI Project, highlighting the progress

[9] http://dublincore.org/documents/dces/
[10] https://schema.org/

made with the user interface, in particular in considering a wider user base with the “API query builder”, and the external Applications API. The demonstrator application, a route based air quality app, highlighted functionality achievable through the access facilitated to internal datasets, alongside, external data sources provided via a partner project (CityVerve) and commercial data provider.

Parallel development sessions explored future research and data synthesis opportunities alongside a session considering opportunities for future application developments. The sessions discussed the potential for delivering application based user interfaces for exploring metadata and opportunities to explore interoperability across other project data repositories. A number of future data providers were identified demonstrating potential to explore the wider interoperability of the Data Provider API interface, presenting further opportunities to support future research proposals both as a data driven service though also from a user perspective exploring PEDASI’s role as broker for of secure quality data.

3.4 TATE Modern exhibition “Living with the Internet of Things” 8th Feb 2019

Contribution to the event focussed around a relatively informal 15 minute presentation and demonstration of the web application driven by data accessed via PEDASI, followed by discussion. The event was well attended for such a session, with over 25 attendees, and the audience were particularly engaged during discussions.

Given the widely reported issues around technology and privacy in recent years, we found that the audience were well-informed and aware of many implications and concerns regarding the collection and use of personal data, through technologies such as the Internet of Things. The involvement of industry in wider Internet of Things projects such as Smart Cities raises a particular concern with the ownership of such data, and how it should be handled and used in collaboration with public projects.

3.5 Industry engagement

Throughout development and testing, both Cisco and BT, PEDASI partners and involved in the CityVerve Smart Cities project, provided valuable technical assistance for their HyperCat API-based data sources. As the CityVerve project completed in March Cisco withdrew support for the service at that time. However, the partnership with Cisco and BT enabling access to their CityVerve data through PEDASI has greatly informed our own data source interface design when working with large-scale catalogue APIs like HyperCat. In addition, BT is currently upgrading the instance used for CityVerve for a new project, and have requested technical access details to ensure we are able to continue using this instance.

Throughout the project we have also worked with GCHQ to determine desirable and specific feature enhancements focussed on data security and assessment of quality. These were prioritised according to their feasibility within the PEDASI workplan and the project's timeframe. Identified feature enhancements would provide indicators of data quality and provenance, as well as improved data handling within PEDASI, providing greater assurance to agencies working with the PEDASI infrastructure, along with an improved level of trust in the data processed by the platform. Other key data sources will also be explored for potential inclusion. These features will be achieved through the creation of developer guidance, operational policies, software revisions, further collaboration, and development recommendations for future versions of PEDASI.

What our early work developing the MVP demonstrator has highlighted are the potential opportunities to support data driven services either analytical based or mobile apps. Early conversations with Southampton City Council and the health technology company Sensyne Health has informed our thinking in the service design and specification for the PEDASI e.g. around ease of use scope of interoperability for wider provision of data sources, in particular datasource API specification.

In particular early interest has been shown in the role PEDASI could play in the aggregation of smart devices/sensors in a secure cloud based platform. With interest in capturing data informing the social insight e.g. the impact on social wellbeing of IoT in the home/business devices, PEDASI provides a mechanism for handling a diverse range of data sources through a singular Applications API and web interface.

3.6 What has our initial engagement told us?

Applications providing a user friendly interface layer e.g. an interface that would allow basic data analysis and visualisation to be performed within the web browser, will be essential if the widest possible user base is to be achieved as it is recognised not all users will be skilled in writing API queries. Further development of the "Query builder tool" could provide some of the necessary functionality. It is becoming apparent the greater success is likely to come from application developers or those requiring a stable and sustainable aggregation platform to surface data to an audience through an application layer.

It is the challenge of delivering sustainability which is most critical to future system development. Sustainability has two primary facets 1) Financial security and 2) Delivering service resilience. Financial security will be a fundamental challenge as a business case for PEDASI is developed, but a number of options to explore in terms of viability, suitability, and scalability include traditional web-based advertising mechanisms (and how these could be adopted appropriately with an analysis to understand user numbers and sectors), a Data Provider sponsorship model (where industry-based Data Providers optionally pay for advertising across the site), and a user/researcher subscription model for accessing secure data sources (where Data Providers supply access to significant value-add datasets). An improved understanding of the requirements affecting service resilience will be established as we explore wider data contribution through the data provider API, therefore gaining an understanding as to extent of API standardisation across data repositories and published datasets.

4. POSITIONING PEDASI AS A DATA BROKERAGE SERVICE

4.1 The observatory landscape

Although not explored in the development of the MVP it became apparent there would be further research required to explore the future positioning of the PEDASI within the observatory and data repository landscape. In particular this would consider primary users, identifying sector and skills and the interoperability issues across this growing landscape.

4.2 The user perspective

PEDASI users can be broadly split into two groups 1) Researchers looking for data to inform ongoing research i.e. those users of PEDASI who are the end users of the data, and 2) Application developers who intend to make use of the data as part of an application to be provided to users other than themselves i.e. those users of PEDASI who are not the end users of the data.

1. Research users: The MVP has established early viability as a data query platform requiring a reasonable skill level in writing API queries. The scope of this user group is not yet fully understood and further activity will need to consider the breadth of users and associated skills in undertaking simple data/metadata aggregations.
2. Application developers: This group will need an increased awareness of the service and associated user policies. This will include brand strategy and business plan informing the positioning of the service in the application development sector. Additionally as for research users the "User policy" will need to highlight license considerations and GDPR for secondary and tertiary use through applications.

A third group might be "Industry users" which might include businesses capturing a wide range of IoT sensor data from diverse sources, large market research organisations exploring data insights or web/mobile application service providers. This group may emerge as the business model and functionality evolves, for example secure data brokering between groups of IoT devices e.g. sensors and an industry facing platform providing sector analytics of these devices.

4.3 Market position

The MVP delivered during this project presents defined users with a basic functionality for the service providing a brokerage of data from a diverse range of data sources. Future work will further define user roles identifying profiles and establishing a broader understanding of what the functions of the service offer and how these users wish to engage with the service as it is evolved.

Further work will explore implications on service delivery based on user understanding/expectations of "Trust" e.g. indicators of data quality, license obligations and data privacy. This will include exploring the current scope for data owners to provide such information when adding a data source and the future potential to surface license and other application profile information programmatically through API queries.

5. IMPACT AND EXPLOITATION

What the MVP development project has highlighted is the clear role PEDASI could deliver contributing to the enhancement of existing data driven services. Current work evolving the demonstrator application has highlighted potential in this space as data brokerage. Outcomes from the air quality demonstrator has highlighted the potential, with further development, to establish an application that could contribute to the decision process for medics and or members of the public with respiratory related health conditions through the provision of detailed route based air quality indicators, e.g. if using a correlation with Google traffic volumes and other route based datasets.

Related activity during the development of the MVP has highlighted the potential to contribute to work in metadata standardisation and associated data management challenges that are necessary if improved secondary and tertiary use of data is establish a level of trust through informed indicators of quality. Informing this potential is the (novel) application of PROV providing data owners with informed usage of data. With further investigation this could be a significant factor informing data usage and application, which in turn could inform the owner of potential quality improvements. Future work will continue to advance the thinking around managing data quality to improve the potential to commoditise the secondary and tertiary use of data and improve trust in data driven services or data synthesis e.g. informing social and economic insights.

PEDASI's design as a single point of entry to multiple data sources awards the benefit of a singular interface to access potentially very diverse datasets, particularly since PEDASI is data-domain and data-type agnostic and delivers data acting as a reverse proxy to those datasets. This reduces the burden on application developers, since they need only familiarise themselves with one data access interface. Similarly, for those exploring multiple datasets through PEDASI's web interface (e.g. via the query builder), only a singular web access point is required.

Since PEDASI acts in both the roles of data producer and data consumer, making use of data access APIs and presenting its own API, it has the potential to be inserted anywhere within an infrastructural data pipeline where either (or both) of these roles exist. For example, exploring the potential for PEDASI to act as a Data Provider itself within the Ocean Protocol[11] data ecosystem, developing a PEDASI application (that uses the Applications API) to act as an Ocean marketplace application with its data sources as available assets. The sustainable and extensible development philosophy behind PEDASI lends itself well to contributing to such ecosystems that require secure, monitored, and robust infrastructures.

Working with Tsinghua University we are exploring the potential use of PEDASI in establishing data trusts, where multiple instances of PEDASI would be networked via peer-to-peer or hierarchical topologies. In such deployments, a single PEDASI instance represents a data access hub within an administrative, regional, and/or security domain and networked to other PEDASI instances to service and potentially forward data access requests within the network. This will allow us to explore the research and technical challenges of such networks, and questions around the alignments of data and user policies.

6. CONCLUSIONS & RECOMMENDATIONS

6.1 System architecture at scale

The software release of the MVP has established the very basic concept which has informed early thinking on potential positioning and scope of use. In the general case, within a wider infrastructure ecosystem, essentially any service that provides data via an API could contribute its data sources to

[11] https://oceanprotocol.com/

PEDASI (as Data Providers), and anything that accesses data by consuming it via an API can make use of external data sources through PEDASI. With both of these service facets there is a clear need for well-defined and documented interfaces at the provider and consumer ends to achieve the optimum potential of the service. PEDASI's well-defined Applications API accomplishes the consumer perspective, with the need to publish the data connector API in the future to satisfy the provider perspective.

When considering the future design of PEDASI to offer scalable support to multiple instances, there is the need to determine an approach that will not place an undue burden of systems maintenance across these instances. In particular, whilst PEDASI needs to support the customisation of singular instances, taking into account stakeholder branding and locally required functionality within its operating environment, managing these customisations centrally is not a scalable approach. Since PEDASI's native Applications API provides access to its core functionality, this will be adapted and extended as necessary to support customised, branded, and functionally amended web applications that will act as localised access portals to PEDASI's capabilities. These web applications will be based upon a simple modifiable and extendable template that can be customised to such localised needs, and registered as applications within PEDASI using its pre-existing applications registry.

6.2 PEDASI as a service provider

Discussions with PETRAS colleagues and wider industry highlighted an emerging role for PEDASI within the concept for “Data Trust Frameworks”. PEDASI presents a unique facility driving data capability through its capacity as a data brokerage service for diverse data sources managed in a secure legal trustworthy framework, a concept which conceptually underpins the establishment of data trusts.

Early discussions have been held with regards to multiple instances of PEDASI, which would support options for networked deployments, e.g. potential for “hub and spoke” and/or "peer-to-peer" deployment topologies. Preliminary work to stress test PEDASI - using 128 simulated simultaneous client requests - has already been successfully achieved, with one PEDASI instance operating as a Data Provider to another instance of PEDASI in a basic "hub and spoke" model. Such PEDASI arrangements would support the establishment of collaborative data trusts that aim to transcend administrative, security, and regional boundaries for the purposes of discovering, sharing, and making use of 'dispersed' data sources. Within such complex topologies, there will be the need to understand and establish documented procedures for managing system updates across the potential many instances of functional data trusts.

6.3 Driving demand for data quality

Fundamental to future success of PEDASI as a data driven service provider will be user trust, informed by indicators for quality (e.g. metadata schema and acknowledged license details allowing ease of secondary use without the need to clarify with the data owner), as well as provenance and data integrity (i.e. providing assurances that data is delivered to users unaltered by any malicious or faulty intermediary and has followed a verifiable trail from its original generating source). Support for these features, and the assurances they provide, will help PEDASI to adopt the role of a secure and trusted data broker that provides a foundation for establishing data trusts.

As this report acknowledges the use of existing standards i.e. surfacing metadata schema and application profiles is not without its challenges, in the first instance encouraging wider adoption. This represents a significant future challenge for data management within industry, where the downstream benefits for improved data management will need to be clearly illustrated.

6.4 Informing the future business model

As explained in section 4. there is a clear need to establish a comprehensive business case with clearly outlined market position, demonstrating scalability of the service and model for sustainability.

Maintenance of the core PEDASI service and services accessing the data it supports via applications is a fundamental aspect of PEDASI, with consideration required for development maintenance of the core system in terms of its implementation, APIs, and software releases, as well as administrative management of active PEDASI instances such as User access permissions, internal data curation and issues of broken links to data sources. Early consideration has been given to such challenges during the MVP development with early specification design considering the use of automated email alerts to data and application owners.

There are very few data management obligations within the service as PEDASI delegates responsibility for maintaining a data source back to the data provider, giving them complete control over how their data is accessed and represented. However, with such functionality there is the responsibility for the data provider to address any issues arising as part of the natural data lifecycle. Clear guidance is therefore required to ensure data providers understand the data management requirements for the representation of their data within PEDASI at all stages of this lifecycle. This has begun to be addressed within this project and will form the basis of a future piece of work.

6.5 The skills challenge

As early project engagement has highlighted there is currently a clear need for users to have a basic understanding of using APIs. To date, PEDASI has focused on the prototypical ‘data scientist’ as a core user type to inform system design. However, future work towards establishing the positioning of PEDASI in the widest possible user community will also require a better understanding of future user requirements and associated skills at a broader scale. This will have a significant impact on interface features and design. Wider influential factors such as the drive for improved quality in data management will have an impact on future skills requirements both for the primary users of data e.g. data owners as well as implications for secondary users who will require a greater understanding of the application of data schema and policies e.g. security and licensing. Such factors may further drive the need to consider the user interface design for PEDASI providing a more informed user support for generating PEDASI queries to data sources, which has already been partially addressed with the development of an API query builder.

We therefore have significant questions to consider in future developments such as “What will future skills look like?”, “Are user skills shifting at the same pace as technology development?” and “Who might be our primary future user base?”.

6.6 The future for PEDASI

Whilst PEDASI is currently a data brokerage based around a centralised service, future developments will aim to extend PEDASI to support a decentralised architecture. This will enable multiple instances of PEDASI to be networked within a peer-to-peer topology, forwarding and servicing requests within a sociotechnical trust framework. This will provide a conceptual, operational, and infrastructural foundation to support the establishment of data trusts across data provider stakeholders that transcend regional, administrative, and security boundaries.

As we continue to explore the role as a secure data handling broker or intermediary within a Data Trust Framework we will need to establish the core components informing user perceptions of Trust. Working with PETRAS/2 colleagues and wider interested parties we will design and adapt existing features providing indicators based on emerging Trust Framework requirements.

There are clear implications for standards requirements surfacing from the development of Data Trust Frameworks. This could have a significant bearing on the need to identify, programmatically, such items as the license applied to data sources and standards used within metadata and data formats e.g. the Application Profile used to structure the data.

If there is to be a role as the facilitator of data access and aggregation e.g. handling a range of IoT in the home data sources for third party service providers, this will depend on the robustness of significant elements of the system design and associated policy:

- Standards adoption informing quality - Shifts across sectors in the adoption of standard schema and application profiles.
- Application of appropriate licenses - Understanding the role as data provider and data owner
- User confidence with delivery of a secure environment - What will be an appropriate level of security and privacy.

It is clearly understood that building a brand for PEDASI has a long way to go. As an understanding of PEDASI as a service product is developed we will be better placed to establish a service with an informed consideration of the core principles of an informed marketing strategy and product placement. This is becoming increasingly important with a growing demand for value realisation from the secondary and tertiary use of data.

………...