k

WPD INNOVATION
Transforming the electricity network

Presumed Open Data (POD)

NIA Major Project Closedown Report

July 2021

WESTERN POWER DISTRIBUTION
Serving the Midlands, South West and Wales

Version Control

Issue	Date
V1.0	July 2021

Publication Control

Name	Role
Sam Rossi Ashton	Author
Stuart Fowler	Reviewer
Yiango Mavrocostanti	Approver

Contact Details

Email

wpdinnovation@westernpower.co.uk

Postal

Innovation Team
Western Power Distribution
Pegasus Business Park
Herald Way
Castle Donington
Derbyshire DE74 2TU

Disclaimer

Western Power Distribution 2021

Contents

1. Executive Summary

There is a large amount of useful data which is published about DNOs through mandatory reports, innovation trials and consumer tools. However, datasets are often published on standalone webpages with limited descriptions. This makes it very difficult for both incumbents and innovators to discover, search and understand datasets.

The scope of this project was to deliver on some of the recommendations made by the Energy Data Task Force, led by Laura Sandys and Energy Systems Catapult, was tasked with investigating how the use of data could be transformed across our energy system, in its Strategy for a Modern Digitalised Energy System Report. Specifically those recommendations were:

> **Recommendation 2 - Maximising the value of data**
> Government and Ofgem should direct the sector to adopt the principle that Energy System Data should be Presumed Open, using their range of existing legislative and regulatory measures as appropriate, supported by requirements that data is 'Discoverable, Searchable, Understandable', with common 'Structures, Interfaces and Standards' and is 'Secure and Resilient'.
>
> **Recommendation 3 - Visibility of data**
> A Data Catalogue should be established to provide visibility through standardised metadata of Energy System Datasets across Government, the regulator and industry. Government and Ofgem should mandate industry participation though regulatory and policy frameworks.

The successful development of a data assessment tool that informs the use of and the extent to which a particular dataset can be shared, and the development of a trial data hub that could host newly available datasets to facilitate a data science challenge has enabled the project to successfully meet its aims and objectives.

2. Project Background

The UK energy system is changing, with an increasing number of distributed energy resources and shifting consumer energy demand profiles. In order to handle the increasing complexity, and support the transition to Net Zero, it will be necessary to utilise digital technologies and data to more effectively manage the whole system.

The Energy Data Taskforce made five key recommendations which, amongst other things, promoted the principle of 'Presumed Open'. This represented a step change for the sector which has the potential to unlock significant value for consumers and stakeholders across the wider energy system. However, opening-up data represents a challenge for distribution network operators due to the magnitude of data and data types collected, when looking to develop and implement their data modernisation strategies and related systems and processes.

The 'Presumed Open Data' NIA project was initiated by WPD in collaboration with the Energy Systems Catapult (ESC) and the Centre for Sustainable Energy (CSE) to demonstrate the value of data which is held by WPD. The goal of the project was to discover what datasets exist, identify the potential demand for those datasets, and develop a process to identify any issues which could limit openness and mitigate these issues where possible. In addition, the project made data available to external stakeholders via a new Open Data Hub and sought to engage data experts through a data science challenge.

3. Scope and Objectives

3.1. Work Package 1 - Data Discovery & Classification

Building on WPD's data inventory work and stakeholder engagement across WPD to identify, review, validate and understand the existing data inventory. The goal was to provide a comprehensive view of the datasets which are available and to review and apply appropriate metadata standards that could be used in the rest of the project. The aim of this was to try to identify datasets which potentially had the highest value across the organisations.

3.2. Work Package 2 - Use Case Development

Utilising the data discovery work from WP1, the goal was to harness expertise from across WPD and ESC to create a number of compelling use cases for the data that WPD holds. This included examples related to WPD's internal challenges as well as providing opportunities for innovators.

The range of use cases will be assessed to identify if the data which exists is sufficient or if there are gaps (data or quality) which needed to be resolved to unlock the potential value. Where limitations are identified, solution options will be proposed. The use cases were to be assessed against a range of criteria (Net Zero impact, WPD's goals, innovation impact, etc.) to rank the importance of WPD data and drive the creation of a data value assessment.

3.3. Work Package 3 - Data Openness Assessment

Building on the work of the Energy Data Taskforce and Energy Data Best Practice Guidance we set upon developing a detailed Data Openness assessment process that couldf be used to identify potential issues with data publication, processes to mitigate issues where they occur and ultimately determine how open a dataset can be and what data licence / terms and conditions should be used.
We would then test the developed process with WPD datasets and review the sensitivity issues related to each dataset (privacy, negative consumer impact, security, and commercials) and develop a methodology to classify the openness of each dataset in accordance with the Open Data Institute's Data Spectrum:

Open: Data is made available for all to use, modify and distribute with no restrictions
Public: Data is made publicly available but with some restrictions on usage
Shared: Data available to a limited group of participants possibly with some restrictions on usage
Closed: Data is only available within a single organisation.

3.4. Work Package 4 – Data Hub Development

Development of recommendations for a public-facing data hub to be hosted online where:

- all data is stored in one central location;

- appropriate means of access (registration/verification of identity) are required for datasets than can be considered Public or Shared;
- data can be easily downloaded upon necessary verification; and
- stakeholders can register to be notified when new data sets are published.

The hub development looked to exploit techniques that are already available regarding the automation of data correction, meta-tagging, and the flagging of data issues when uploading new data.

3.5. Work Package 5 – Data Science Challenge

The aim of work package 5 was to design a well-defined data science challenge relevant to the appropriate stakeholders – A rigorous selection process was followed to select an appropriate data science challenge. Firstly the challenge was chosen from the themes develop through Work Package 2 from engagement with internal and external stakeholders. Secondly, it was designed around some of WPD's largest and most reliable data sets to guarantee there would be sufficient data.

3.6. Work Package 6 – Data Playbook

The purpose of the data playbook was to guide the user through the process of identifying, assessing, publishing or sharing and getting feedback on data within network operator organisations. It will allow compliance with the Energy Data Best Practice guidance introduced as part of the RIIO-2 code changes.

Table 3-1: Status of project objectives

Objective	Status
WP1 - Discover and document as many datasets as possible within a series of teams that exist in WPD	✓
WP2 – Run a series of workshops to establish a series of use cases that can be used to prove the value of data	✓
WP3 – Produce a data openness tool to easily conduct open data triages	✓
WP4 – Development and deployment a data hub	✓
WP5 – Run a successful data science challenge using datasets identified in the project.	✓
WP6 – Draft a data playbook to guide LNOs through the process of identifying, assessing, publishing or sharing and getting feedback on data.	✓

4. Success Criteria

4.1. Work Package 1 - Data Discovery & Classification

Due to the variety of data solutions utilised across any large business there is no single way of discovering datasets. For example, the WPD Innovation Team has project documentation published online, the Network Strategy team has a list of datasets and others provide information on request. Our methodology therefore consisted of a number of different consultative approaches to discover as many datasets as possible and document them appropriately, using a tailored approach to the business areas tools and ways of working. It was expected that the delivery of the data catalogue will go some way to creating an understanding of what data is available throughout the business and to start to align these into a published and maintained data catalogue.

4.2. Work Package 2 - Use Case Development

The Work Package 2 workshops were initiated with an introduction to a number of example data use cases (included below), these were circulated with the workshop participants prior to the meeting to allow them to start thinking about their own use case ideas before the workshop.

The participants were split into smaller subgroups with circa 5 people per breakout room. Within the subgroups, the participants were asked to think about potential use cases. To keep the discussions and ideas flowing barriers were not considered, all ideas were accepted and recorded.

The use cases were recorded as User Stories, this allowed the participants to think about specific goals of a user in the format:

As a **ROLE** I want to **NEED** so that **OUTCOME**

4.3. Work Package 3 - Data Openness Assessment

Within Work Package 3 the team refined the Open Data Triage concept and produced a tool that could be used to systematically identify issues and document the mitigation of any issues. This tool was developed based on the existing Geospatial Commission Data Sharing Assessment Tool but with adjustments to make it more appropriate for WPD's needs and make the guidance more accessible.

4.4. Work Package 4 – Data Hub Development

Work Package 3 took learning from Work Package 2 about what functionality a data hub would require to facilitate the use cases identified and developed, and added feedback from data stakeholders about how they wanted a hub to function to create a hub that could be embedded with WPD web estate to facilitate the data science challenge.

4.5. Work Package 5 – Data Science Challenge

To achieve success that challenge was designed maximum participation in that it was marketed by the three project partners, a LinkedIn ‘community’ was created, and the prizes were consider by the project team as attractive.

4.6. Work Package 6 – Data Playbook

During Work Package 6 the team engaged with the Data and Digitalisation team at WPD to understand what aspects of project and industry learning would be relevant to distil into a document that can be referenced in a ‘lift and shift’ manner to inform data treatment policy.

Table 4-1: Status of project success criteria

Success Criteria	Status
WP1 – Document at least 3 of the main datasets from each of the functional areas identified.	✓
WP2 – Produce a number of compelling use cases from a range of stakeholders across the energy industry that can be used to demonstrate the value of the data sources identified in work package 1	✓
WP3 – Produce an easy to understand and easy to use open data triage tool that can be used to advise on the sensitivities of data and the necessary modifications to the data	✓
WP4 – Deploy a data hub which can host all the data necessary for the data science challenge, which includes a login system and a mechanism for	✓
WP5 – Successfully run a data science challenge which has a comparable number of participants to its peers	✓
WP6 – Publish a data playbook that can be used to inform data treatment policy	✓

5. Details of the Work Carried Out

5.1. Work Package 1 - Data Discovery & Classification

The ESC team applied a consultative approach to the data discovery work by engaging with teams across WPD to learn about their respective areas and document the datasets and systems which they generate or use. Due to the short-time available for work package 1, the plan was to leverage as much existing documentation as possible and seek to elicit more detail from key stakeholders through focused meetings. This approach allowed us to create a first cut of the data catalogue, so that work package 2 can begin and this can be used to develop a set of use cases. To ensure that the ESC team utilised their time effectively they focused on the target areas outlined below:

- Innovation – the team(s) responsible for NIC/NIA projects, external engagement with innovative ideas and delivering projects associated with long term future ideas and market needs
- Network Strategy – the teams responsible for the evolution of the energy network in the medium-term future, including scenario planning for different types of policy changes
- Operations – the teams that manage day-to-day operation of the physical network equipment and assets including network management
- Information Resources (IR) – Digital systems and IT support across the business
- Regulatory Reporting – the teams that provide mandatory information to Ofgem for to compliance with licence obligations and to inform the RIIO price control

With support from WPD's Innovation team, the ESC made contact with a representative from each of the areas above and negotiated an engagement approach to fit in with their day-to-day duties.

Following a review of the Data Improvement Team's outputs, including a face-to-face engagement, the ESC project team aimed to get an understanding of not just which systems were interacting with each other, but which ones had significant data flows between them. This further solidified the understanding that much of the data had provenance in one (or many) of the core management systems (i.e. asset management, network management, and geographical information) even if it was not extracted from them directly. This helped to contextualise the other datasets from other systems in the context of the overall data landscape. These relationships are captured in the data catalogue in the "relates to" category which was very valuable. The first stage of documenting the data sets is to produce metadata. Metadata is a dataset that describes and gives information about another dataset. This can be used by potential data users to understand what they may be able to do with the data technically and legally.

In line with the Energy Data Best Practice Guidance, the metadata is based on the Dublin Core standard with some additions which provide additional value to WPD and the ultimate data users.
In addition to the basic information provided by the metadata, supporting information has been collated. This is designed to help potential users understand the contents of the data more effectively and includes resources such as:

- Data Field Descriptions taken from the documentation that accompanies the data
- Innovation Project Reports produced either during the projects or during the closedown of the projects
- System Overview Documents including where data flows between systems, how to use guides and other system documents
- Interviews with system owners where possible
- Utilising the Data Improvement Team's work as an information source.
- Looking at and directly assessing the data and inferring from the data itself (for example with dates)

Where a dataset has been discovered but metadata is incomplete, this may be because the WPD contact was not able to provide all of the required information or the team have been unable to identify an appropriate contact. Incomplete records will be addressed in the following work packages as more information is uncovered. In addition, further datasets will be added to the catalogue as the required information is uncovered.

5.2. Work Package 2 - Use Case Development

The focus of work package 2 was to develop a greater understanding of the value of WPD data for internal and external stakeholders such that it can be prioritised for open data triage in future work packages. In order to understand the value of data from both internal and external stakeholders a number of use cases were developed which require WPD data. Where a dataset is required more often and will 'unlock' a number of distinct use cases these are considered to be of higher value.

Alongside the identification of data value, ESC has also assessed a range of datasets to understand if they will be able to fulfil the needs of the use case. This could include ensuring that the data is available in the sufficient level of granularity, completeness and accuracy to meet the needs of the user. The following subsections outline the key activities undertaken in this work package.

Internal stakeholder workshops

The Energy Systems Catapult hosted two workshops with WPD stakeholders from a range of departments. These were originally planned to be physical workshops but due to COVID-19 travel restrictions these were held as virtual events via Zoom.

The aim of each 2.5 hour workshop was to introduce an abstracted version of the data catalogue which explained the kinds of data which are held across the organisation and start to develop some high level use case concepts with more detail provided on a subset of the use cases.

External use case workshops

The external stakeholder workshops were organised by the CSE and followed a similar structure to the internal workshops. However, due to the different audience the second half of the workshop additionally requested additional input on the requirements of the Open Data Hub which is being developed in work package 4. These workshops are the subject of a separate CSE report.

Use case analysis

Following the internal and external workshops, 145 individual use cases concepts across all 4 workshops were collated, however, there was some duplication across groups. Rather than consolidating use cases and reducing the richness of the data the use cases were grouped into themes to keep the details of individual use cases. This surfaced a number of common themes across the sessions and has provided a useful way of thinking about the value of data groups.

The themes and use cases were subsequently analysed to understand what data requirements existed. This provides a method of identifying where a dataset was necessary to enable many use cases and therefore where the value of making it more widely available is greater. In addition, this enabled analysis of the relationships between datasets, if a specific dataset is most valuable when made available alongside another this should be considered.

5.3. Work Package 3 - Data Openness Assessment

The recommendations of the Energy Data Taskforce stated that data should be Presumed Open. This has since been endorsed by Ofgem and BEIS, with Ofgem signalling their intention to make this part of the upcoming RIIO2 licence conditions for network companies.

Presumed Open is the principle that data should be as open as possible. Where the raw data cannot be entirely open, the data custodian should provide objective justification for this.
For data to be made ‘as open as possible’, it is necessary to have a formal process which can be used to identify potential issues and mitigate them as necessary, this is referred to as Open Data Triage.

Open Data Triage is a process to systematically identify issues (Privacy, Security, Commercial, Negative Consumer Impact or Legislation and Regulator Barriers) with a dataset which limit their potential openness and then identify what techniques can be used to mitigate these issues.

In order to develop the open data triage process, the project team started by gathering requirements from WPD via a set of workshops, which helped to develop a prototype tool, review and refine the prototype before publishing the tool and associated guidance.

The requirements captured in the workshop were synthesised by the project team and the ESC considered how existing data sharing assessment tools met the needs of WPD. It was identified that the Geospatial Commission Data Sharing Assessment tool met many of the requirements and, as it is made available under the Open Government Licence (OGL), it was a good starting point for prototype development.

The project team made a number of changes to the base tool to simplify the language, remove irrelevant questions and streamline user experience (consolidating the number of tabs). This resulted in the streamlined set of questions.

The prototype tool was presented to stakeholder within WPD in order to help the ESC project team understand how usable the tool is and check if any additional features were required.

5.4. Work Package 4 – Data Hub Development

The POD Data Hub software was developed based on CKAN open source data portal software. This consists of a set of components which work together to run the Open Data Hub. The main CKAN software version used is version 2.9.0 (https://github.com/ckan/ckan/tree/ckan-2.9.0). This has been modified for use within POD. The software is run on a server running nixos (https://nixos.org/). This allows the deployment to be completely scripted through the use of nix files.

The Data Hub consists of the following components

1. **CKAN** The core CKAN application.
2. **Postgresql** A relational database used to store metadata and structured data extracted from suitable format data files (e.g. csv files).
3. **Solr** A search platform used to find data. Version 6.5.1 is used as this is the latest version to work with the standard CKAN sourcecode.
4. **Redis** A key-value database used by CKAN. Used in the management of background jobs.
5. **CKAN datapusher** This is a separate process to the core CKAN process. It converts structured data files to data values which are saved via the CKAN datastore to the Postgres database.
6. **Nginx** Used to provide the external interface to CKAN.

CKAN supports plugins. These are used for a number of purposes:

1. **datastore** A plugin that provides an api used to store data values into the database. Further plugins, such as “recline_view”, then render this data for visualisation by users.
2. **datapusher** This is separate to the CKAN datapusher service. It provides communication between CKAN and CKAN datapusher. When data is uploaded, this plugin passes details about the data to the CKAN datapusher service.
3. **pod** This plugin was written for the POD project. It contains most of the customisations of CKAN for POD. These include changes to text displayed, alterations to the dataset metadata fields re. the “Dubin Core” requirements and changes to the CKAN styling to, for example, use WPD Innovation colours.
4. **Views** Each data file can have a number of “views” configured. When data is uploaded, some views are automatically added (see the ckan.ini configuration “ckan.views.default_views”) and others can be added manually. The ODH install has the recline_view, text_view and image_view views configured.

5.5. Work Package 5 – Data Science Challenge

The data science challenge involved the following main steps:

1. Selection of the data science challenge
2. Collecting and pre-processing the data sets

3. Advertising and recruiting participants
4. Challenge Implementation
5. Collating and summarising the results.

Each of these steps is described in a little more detail below and full details can be found in the data science challenge report.
The challenge was selected based on the following criteria:

1. What are interesting and important use cases for WPD and wider external stakeholders?
    - The latter is relatively important to ensure wider participant engagement.
2. Which problems could be easily formulated as a well-defined challenge?
    - The problem should be simple enough to explain to non-energy experts. It should also have relatively simple scoring criteria.
3. What data would likely be available and of high quality, with minimal missing and anomalous values?
    - This minimises the data pre-processing overhead and means participants can focus on model development.
4. What areas does the project team have domain expertise?
    - This assists the design of a robust and well-defined problem and improves identification of potential pitfalls.

The main theme of the challenge was selected from those collected as part of the stakeholder engagement in work package 2. Four main themes were selected a different challenge was outlined for each. They were then assessed in terms of how they satisfied the above four criteria. The chosen challenge was an optimal control of a storage device within the theme of “Maximising Asset Utilisation”. This had the advantage of being an area of interest to many stakeholders, had relatively reliable and accessible data sources, and was a challenge which could be readily transformed into a well-defined objective, in this case, maximising peak reduction.

The next task was to collect and pre-process the data. There were three main sources of data: demand, photovoltaic generation and weather data. These data sets were selected based on various properties, in particular data quality (there should not be too many missing and anomalous values), location (they should be located close to each other to ensure consistent correlations between variables) and amount of data (enough data is necessary to create robust and reliable models). Given these properties there was minimal pre-processing required. The PV data and weather data were also used to validated each other and ensure that similar variables were consistent with each other (e.g. PV generation and irradiance variables). In addition, the demand data required evening peaks and lower daytime demand to ensure that the problem could be split into charging and discharging period which would simplify scoring criteria.

Finally the data needed to be partitioned into training and test data sets. Hence further analysis was required so that suitable task weeks could be chosen. They should have reliable data and have suitable properties so that different scenarios could be tested.

The next step of the challenge was recruitment and advertising. Multiple sources were used including emailing to relevant researches and companies, social media advertisements (including relevant LinkedIn groups) and popular emailing lists for energy research and data science. The latter include, email lists and newsletters for OpenMod Initative, Power Swarm and Climate Change Ai.Participants were directed to a registration page on WPDs webpage where they could enter details and be given further information about the challenge kick-off event. To encourage participation, two prizes were offered, the first the offer of a publication in the Energies journal and the other a offer to pitch an idea to senior staff of the partners in the project.

To implement the challenge, in addition to the kick-off event, documentation was shared with extensive details about the challenge, the scoring criteria, important dates and rules for the challenge. A LinkedIn group and a dedicated email address were also set up to interact with participants and answer their questions. A practice challenge was used to smooth out the implementation and help teams better understand the rules and the challenge in greater detail and were a useful part of the process in terms of improving quality of submissions in the later tasks. There were five tasks in total (including the practice one) implemented over a seven week period with the first two tasks providing 2 week period of training the models. The final three tasks only had a week between each. This was to ensure that the project could be wrapped up within sufficient time and results disseminated. The day after the deadline of each task a new data set was released and the participants informed via the email list and the LinkedIn page. Submissions needed to be submitted to the dedicated email list by midnight on the day of the deadline using the template provided on the data hub.

The participants also received scoring shortly after each task. A scoring program was developed to take in the submissions using the template and provide the scores for each day of the current task. These were then fed back to the teams as well as errors in submissions to the relevant teams. For the final task teams were also asked to submit answers to a questionnaire which had various questions about the methods and data they used in their models. This was collated together with the final scores to try and understand which methods generally tended to score highest although there were many conflicting factors which make this tricky in practice. The results were fed back to the teams with some overview of the results from the questionnaire.

5.6. Work Package 6 – Data Playbook

Work Package 6 was a short additional work package wherein learning from the project previous work packages coupled with existing industry learning to create a document that could be used to inform LNO data policy. The method for this was that a skeletal document was drafted which was the subject if a comprehensive review session from WPD's Data and Digitalisation Manager. This document was the populated and again reviewed before the final version was submitted for approval.

The final document consists of the following sections:

1. Identify Dataset
2. Capture Metadata
3. Use Case and User Needs Identification

4. Identify / Appoint Internal Data Owner
5. Identify Release Mechanism
6. Open Data Triage
7. Issue Mitigation
8. Documentation
9. Review
10. Sign Off
11. Feedback

6. Performance Compared to Original Aims, Objectives and Success Criteria

6.1. Work Package 1 - Data Discovery & Classification

The ESC team discovered a large number of datasets which were added to the data catalogue. The level of detail in the documentation varies between the following categories:

- Complete metadata with supporting information
- Complete metadata
- Incomplete metadata

During the course of the Work Package over 60 datasets were identified and documented in a data catalogue. During work package 2 the catalogue has been expanded and developed and now includes over 100 separate entries

6.2. Work Package 2 - Use Case Development

Participants were identified and invited by the WPD Innovation team. The knowledge and existing relationships of the Innovation team ensured that there was a good turnout with a mix of individuals from across the organisation and with diverse skills. The attendees for each workshop (excluding facilitators) has been included below. The attendees were further subdivided into breakout groups of 3-5 attendees to enable small group discussions.

Table 6-1: Internal Workshop Attendance

Workshop 1		Workshop 2	
Name	**Department**	**Name**	**Department**
Adam Mealings	Control Systems	Adam Mealings	Control Systems
Dave Tuffery	Network Strategy	Dave Tuffery	Network Strategy
Ian Hatton	Data Improvement	Ian Hatton	Data Improvement
Ian Whatley	Data Improvement	Ian Whatley	Data Improvement
Jenny Woodruff	Innovation	Jenny Woodruff	Innovation
Kester Jones	Network Services	Kester Jones	Network Services
Lewis Williams	Innovation	Lewis Williams	Innovation
Matthew Alderton	Control Room	Matthew Alderton	Control Room
Mitch Golder	Purchasing	Mitch Golder	Purchasing
Paul Dodimead	Procurement Manager	Paul Dodimead	Procurement Manager
Phil Rigden	WPD Telecoms	Phil Rigden	WPD Telecoms
Sam Rossi Ashton	Innovation	Sam Rossi Ashton	Innovation
Simon Catlin	IR/IT Manager	Simon Catlin	IR/IT Manager
Tara Louise Brewin	Legal	Phil Lawson	INM/Data
Yiango Mavrocostanti	Innovation	Sam Rossi Ashton	Innovation
		Steve Quinn	Network Strategy

There were a large number of use cases developed across all workshops and these cannot all be investigated in detail in this report. However, each use case was categorised and the key themes identified. The table below provides an overview of the themes and the number of individual use cases within. Note, in some cases a use case could have been placed into more than one group (e.g. Local Area Energy Planning and Decarbonisation) but we have only counted a use case in the theme which felt most relevant.

Table 6-2: Use Cases by Theme

Use Case Theme	Total Use Cases
Connections and Constraints	31
Cross sector	2
Decarbonisation	35
Education and Research	17
Enabling EV Charging	8
Improving Data Quality and Interoperability	6
Improving Fault Diagnosis	6
Local Area Energy Planning	34
Optimising Network Strategy, Planning and Operation	3
Policy	6
Tariffs and Trading	13
Consumer vulnerability and equality	9

We noted that there were some candidate themes which were prominent but have not been selected for example; decarbonisation of heat and flexibility services. When selecting the themes we tried to select a coherent set of themes with minimal overlap (avoiding double counting of use cases) whilst limiting the number of themes with a small number of use cases (<5).

6.3. Work Package 3 - Data Openness Assessment

The test runs and peer reviews from CSE colleagues identified that it was essential for the tool to be accessible for a novice user. This includes providing more detailed guidance about the various licences and legislation that is referred to in the tool and practical guidance about topics such as identifying security issues.

Also, the decision was made to include an additional field which captured the nature of the mitigation technique used when an issue was identified. This was necessary to ensure that when the 'openness recommendation' which is included on the dataset summary tab provides the right value.

User Guide

A user guide was developed to help prospective users navigate the tool. This was informed by user testing and peer review within the project team. The user guide has been integrated into a tab on the triage spreadsheet for ease of use, but a copy has also been included in the appendix of this report.

Data Sharing Assessment Tool

Following the final review of the tool with the POD project manager and business sponsor the tool is ready for use within WPD.

6.4. Work Package 4 – Data Hub Development

The data hub was successfully deployed for the data science challenge by CSE and used by the number individuals written in the Data Science Challenge section below.

6.5. Work Package 5 – Data Science Challenge

To drive engagement with WPD – Over 360 individuals registered for the event with 55 teams consisting of 142 individuals entering at least one submission for the challenge. Of these 37 finished the full challenge. The challenge also received a wide range of participants, from consultants, energy sector data scientists and university researchers, from at least 72 different institutions and 15 different countries.

In addition to the wide and consistent engagement with the challenge, feedback was extremely positive and at the end of the challenge multiple teams released their code. The opening of WPD data attracted participants who were open to sharing and collaborating on learning more about the approaches and methodologies. This project has not only demonstrated the value in releasing data but how it can also help drive innovation.

There were also other high-profile competitions with which we could benchmark our expected participation. Two of the most well-known ones are the Global Energy Forecasting Competitions (GEFCOM) and the Makridakis Competitions (or M-competitions). These typically are very popular, are run by leaders in the field and have large resources for advertisements (for example the M-competition uses the leading journal in forecasting, the International Journal of Forecasting, to advertise).

The 4th M-competition, M4, ran in 2018 received 50 submissions from 248 registrations. GEFCOM 2012 had over 200 teams over 2 tracks, GEFCOM 2014 gathered 581 participants from 61 countries over four tracks, and GEFCOM 2017 had 177 academic and company teams.

We expected to have a much lower participation in our challenge due to its reduced status and reputation compared to these longer established competitions. The original estimate for participation was 10 teams and perhaps no more than 40 individuals participating in the trial. However, as addressed above our expectations were far exceeded with over

360 individuals registering, and 55 teams of 142 individuals involved in the challenge. Not only this, but participation remained high until the final task with 37 teams finishing. Further 72 different institutions from at least 15 countries were involved.

6.6. Work Package 6 – Data Playbook

A data playbook has been successfully published and WPD plans to use this to develop data treatment policy in the near future.

7. Required Modifications to the Planned Approach during the Course of the Project

The project undertook two modifications to approach:

7.1. CR1 - April 2020

Proposed change: Extended the closure date of WP2 from 5th June 2020 to 30th June 2020 to facilitate the orchestration of virtual workshops instead of face-to-face ones.

Reason for change: COVID restrictions prevented face to face workshops

Effect of NOT making change: The project would not have had enough time to undertake platform testing, invitations, workshops and then write-up before original WP2 closure date.

7.2. CR2 – December 2020

Proposed change: An extension to WP5 (Data Science Challenge) and the addition of a work package to create a document suitable to transition learning into policy

Reason for change: Additional Work Package: The driving motivation behind the project was an Energy Data Task Force recommendation to make data Discoverable, Searchable and Understandable. WPD was keen to implement this as soon as practicable and as such chose to develop documentation for the process outlined below based on the Data Best Practice Guidance, with supporting resources where required, which can be used for policy development.

Effect of NOT making change: Use of project learning and delivery of Energy Data Task Force recommendation would not be optimal.

8. Project Costs

Table 8-1: Project Spend

Activity	Budget	Actual
POD - WPD INTERNAL COSTS	£62,050	£63,922
POD - CONTRACTOR COSTS	£514,250	£460,570
TOTALS	**£576,300**	**£524,492**

Contractors Costs Variance:

Contractor costs came approximately 10% under budget as the post-data science challenge of the POD data hub was not required as WPD when onto develop its own hub using the POD project learning.

9. Lessons Learnt for Future Projects

9.1. Work Package 1 - Data Discovery & Classification

The primary takeaway from this work package is that WPD has a wealth of valuable information across the organisation, the value of which will be easy to maximise if the barriers to entry are lowered. WPD are responsible for an essential part of national infrastructure. Its network connects millions of customers' homes to the wider energy system and it is essential that this is robust and secure. However, some of the decisions that have been taken to make WPD more secure have a negative impact on their ability to collaborate with external teams and adopt remote working. Examples of this are lack of internet access on certain computers which restrict ability to use screenshare software, lack of laptops which make it difficult to work whilst not in the office and lack of an approved remote data sharing solution necessitating the use of USB sticks to transfer large amounts of data.

Internally WPD faces similar challenges to the wider industry on the acquisition and usage of data internally, outside of the operational management of assets and the network. This has an advantage in that the areas that work on specific tasks and use specific datasets are very well defined in their responsibilities, but it is clear that there is value to a company wide data strategy to enable the functional areas to utilise core data more effectively without having to know specifically who to go to for something. This makes it difficult to find data in WPD and even more difficult for 3rd parties to get access to data, even if triaged effectively.

Although made up of 4 different geographical areas, the WPD core systems that are used for assets and physical infrastructure are consistent across the 4 networks. Although not all of the data could be accessed from the core systems during this work package, we have found many derived data sources that utilise the data from these core systems and there is alignment on the way in which data is captured from the assets themselves as they are ingested into the core systems.

When it comes to the innovation and network strategy datasets, they are typically well documented as they are published in a publicly facing environment. It is worth noting that when innovation projects are closed down it is uncommon for data owners to be formally appointed. This is not a problem which is restricted to WPD, but it does restrict the future value which is extracted from the data. This issue can be easily rectified through a formal process to nominate ongoing data owners at project close down and where possible, the publication of data in a discoverable, searchable and understandable way.

In general, with the correct tools and processes in place, WPD will be well placed to start providing data to the wider industry. Given the alignment of systems and the provenance of the asset data being in aligned systems it will be possible to start running valuable regular extracts that will be of value to third parties. In addition to this, given that innovation and network strategy already publish much of their work, it will only take a small amount of coordination to align the datasets into common formats and common metadata in order to make them more discoverable, searchable and understandable.

9.2. Work Package 2 - Use Case Development

Within Work Package 2 the POD team has engaged with a large number of stakeholders across the business and industry. This was a challenging task during the UK COVID 19 lockdown period but with the support of the WPD innovation team and with access to the right digital collaboration tools we were able to keep up momentum and deliver valuable insight in a timely manner. In the sections below we focus on some of the most noteworthy areas of learning and insight.

Digital Workshops

At the start of the COVID 19 lockdown many organisations had to spend time adjusting to the 'new normal' of increased remote working and the variety of IT and logistical challenges that presented. However, the POD team was quickly able to pivot towards a more digital way of working and reshape the planned workshops to replicate the dynamic of a physical workshop.

We found that both WPD and external stakeholders were keen to participate in our workshops and were able to offer some very valuable insight. In fact, given the distributed nature of WPD staff and external stakeholders we believe that providing virtual events may have increased the diversity and number of participants.

Stakeholder Engagement

One of the primary goals within this work package was to build a base of interested and engaged stakeholders. We know that staff across WPD are busy and it is therefore challenging to find time to engage with activities outside of their day to day duties so it was important for the project team to show stakeholders that we value their time and input so that they will be keen to engage further later on in the project. The initial feedback received seems to suggest that stakeholders enjoyed the workshop format and felt that it was a useful activity.

When working with WPD stakeholders it quickly became clear that there are a significant number of innovative members of staff outside of the core innovation team who have good ideas and are keen to use data more effectively in their roles. From Network Strategy through to Customer Contact there were an array of strong use cases which have been gathered.

We note that there is significant potential to use the workshop approach to find and develop innovation ideas within WPD which can benefit customers, external stakeholders and WPD alike.

Data Value

The workshops have shown that WPD data has significant value to stakeholders outside of the normal groups. From sharing data more effectively within WPD to making more information available to external parties there is significant value to be extracted.

9.3. Work Package 3 - Data Openness Assessment

Within Work Package 3, the POD team has engaged with stakeholders across the business to review and test the developed tool with a number of different datasets. In the sections below we focus on some of the most noteworthy areas of learning and insight.

Business Support

In order to truly embed Presumed Open, it will be critical to combine 'top down' leadership with 'grass routes' support across the business. It will be essential that individuals across the organisation are empowered to triage their own datasets and make informed openness recommendations which are supported by business leaders.

It was evident from the engagement received in work package 2 and during the requirements gathering activity at the start of work package 3 that there is a good level of support for Presumed Open Data across the business. WPD stakeholders understand that the data they hold has value to others across the company and broader industry.

However, despite the interest we found that many individuals struggled to provide support to the project due to time and resource constraints. This is likely due to unfortunate timing with other projects and commitments, but it highlights that there is a need to transition Presumed Open Data to be BAU with a share commitment across the organisation.

The appointment of the DSO Digitalisation and Data Manager is a very positive step forward for WPD and provides a good point of contact for Digital and Data topics. To ensure the organisation as a whole continues to push forward with Presumed Open Data it will be critical to educate and empower other managers and directors, so they are able to champion Presumed Open Data in their own teams and areas.

User Guide

The initial drafts of the Data Sharing Assessment Tool were released with little user documentation but triage was supported by the project team who were available to provide guidance and support to help individuals across WPD to provide the right inputs and deliver a robust triage of the dataset. However, without the additional guidance from the project team the tool, some individuals found struggled or took more time to complete the assessment.

To aid in the use of the tool the guidance sections were simplified to make them more accessible, the extent of the guidance was expanded, and an additional user guide was included. This provided more support for users who were not given a full introduction to the tool.

Validation and Iteration

Within this work package the project team have tested the Data Sharing Assessment Tool with real datasets to understand if it meets the need of the end users. This has been critical to identify additional needs, find where the tool is not working as planned or as the user expected and check that the tool delivers valid recommendations.

Whilst we have been able to verify with a number of datasets these have primarily been from innovation with the exception of the Embedded Capacity Register. This has provided some useful feedback but has meant that we have not tested with more sensitive internal data which may have thrown up more issues.

To mitigate this issue we would recommend that we continue testing during the next work package as the relevant people become available.

User Confidence

The concept of Presumed Open is relatively new for the energy sector and does not always come naturally for an organisation that operates critical national infrastructure and is naturally risk averse. In our experience, individuals who have assisted with the triage of datasets have all of the knowledge and skills needed to identify and sensitively mitigate issues however the risk averse nature of the energy sector creates doubt and worry that they have 'missed something'.

For Data Sharing Assessment Tool and the Presumed Open Data project to gain traction and truly become Business As Usual it will be essential to ramp up engagement with stakeholders across the organisations. Providing feedback on assessments completed and additional training where required. It is advised that in the first instance the DSO Digitalisation and Data team provide support to the business to review assessments and give users the confidence to provide their honest assessment rather than a restrained view.

9.4. Work Package 4 – Data Hub Development

The key lessons learned from hub development were that:

- It is possible to develop and implement a hub that can cater for a future where large operational datasets are made more sharable.
- An off-the-shelf data management system, such as CKAN, is being suited to this.
- Consideration needs to be given to IT changes that need to take place in order to enable the uploading of large datasets to a hosted site.

9.5. Work Package 5 – Data Science Challenge

The learnings from the data science challenge have suggested some key points for development of future challenges:

- Allowing participants to use external data would make a much more diverse solution and perhaps identify useful data sets for the challenge. This must be carefully considered, since it allows performance to be driven by data accessibility and could lead to uneven playing field for teams with lower resources or accessibility. One option is to split the challenge into two tracks, one with and one without restrictions.
- Time for participants to work on a task should be considered. Participants often take part in the challenge in their own time and hence longer gaps between submissions can be advantageous but also require longer commitments. One compromise could be to reduce the number of tasks but make the test set much larger to allow more robust assessment.

- It is beneficial to choose a realistic problem which can be easily framed as a well-defined competition problem. However, this is not always trivial. More realistic problems are easier to understand but are not always easy to score, and well-defined competitions are not always realistic which can cause confusion. A balance must be sought when developing future challenges.
- Make sure the data is as clean as possible. Although this challenge used high quality data even the small number of erroneous values provided some difficulty and added preprocessing time for some participants.
- Consider skill scores rather than absolute metrics to reduce biases and variations in the tasks.
- A practice challenge is essential. It helps to work out bugs in the submission process and helps clarify the challenge requirements to the participants.
- Regular engagement with the teams is well-received. The LinkedIn Forum and emailing list helped facilitate information sharing and for teams to discuss with each other and clarify minor points from the challenge. It also ensured that information was fairly shared with all participants.

9.6. Work Package 6 – Data Playbook

As this was a collation of existing learning, there was no specific learning from this work package.

10. The Outcomes of the Project

The outcomes from this project include:

- Learning Reports generated at the end of each of the six work packages.
- Our Data Process Team is now using the Data Playbook developed during Work Package 6 to formalise the policy and process required to fully adopt the Data Sharing Assessment Tool developed during Work Package 3. This will enable third parties to request previously unshared WPD datasets
- The Energy Networks Association (ENA), with the help of WPD, used the Data Playbook developed by the project in Work Package 6 to create the ENA Data Triage Playbook, this in-turn was used to develop the Energy Data Request Tool which enables a single point for data requests and standardising networks' approach to servicing these requests.
- Additionally, due to the success of the Work Package 5 Data Science Challenge, we are looking to setup additional challenges to engage the data science community with energy problems.

11. Data Access Details

The POD project was largely concerned with the study and treatment of existing datasets rather than creating new ones. There was however analysis done on the use cases generated in Work Package 2 to group them into themes. This analysis is available upon request from: wpdinnovation@westernpower.co.uk

12. Foreground IPR

The IP generated throughout this project includes:

- The WPD POD Data Catalogue
- The Use Case Assessment Analysis
- The Data Sharing Assessment Tool
- The POD Open Data Hub
- The Data Playbook

13. Planned Implementation

WPD is scaling up its capability in digitalisation immediately as a result of this project. Following on from the project there are the following intentions.

13.1. Data science challenges

Following the success of the first data science challenge and to incentivise a regular data release schedule, the intention is to run a new data science challenge every quarter with new data published and documented. Over time this will allow the teams involve to upskill and learn by doing practical projects, lead the rest of the energy sector in terms of data publication, and allow WPD to provide new data use cases to guide the digitalisation strategy.

13.2. Digitalisation strategy

WPD intends to continue and embed the learnings from this project by continuing to grow the team that are working on digitalisation. By working with external parties that want to access and use data, publishing data on a regular cycle and responding to user feedback, WPD will develop a much closer relationship with interested parties in the sector and take a leading role in enabling innovation in energy. The data release playbook will be used as a basis for future iterations and the processes will be improved an updated over time as more data is published and further data science challenges are conducted.

14. Contact

Further details on this project can be made available from the following points of contact:

Innovation Team
Western Power Distribution,
Pegasus Business Park,
Herald Way,
Castle Donington,
Derbyshire
DE74 2TU
Email: wpdinnovation@westernpower.co.uk

Glossary

Abbreviation	Term
BEIS	The Department for Business Energy & Industrial Strategy
CSE	Centre for Sustainable Energy
DNO	Distribution Network Operator
DSO	Distribution System Operator
ESC	Energy Systems Catapult
GEFCOM	Global Energy Forecasting Competitions
IT	Information Technology
IR	Information Resources
LNO	Licensed Network Operator
NIA	Network Innovation Allowance
NIC	Network Innovation Competition
RIIO	Ofgem Network Price Control
WPD	Western Power Distribution

Western Power Distribution (East Midlands) plc, No2366923
Western Power Distribution (West Midlands) plc, No3600574
Western Power Distribution (South West) plc, No2366894
Western Power Distribution (South Wales) plc, No2366985

Registered in England and Wales
Registered Office: Avonbank, Feeder Road, Bristol BS2 0TB

wpdinnovation@westernpower.co.uk
www.westernpower.co.uk/innovation

WESTERN POWER DISTRIBUTION
Serving the Midlands, South West and Wales