Open data assessment report

Executive Summary
May 2023

Introduction

Terms of reference

PricewaterhouseCoopers LLP (PwC) was commissioned by Ofwat to develop and deliver an assessment of water industry's progress on Open Data in response to H2Open.

Process of work

To assess progress, we developed an assessment framework in collaboration with Ofwat, water companies and water and other industry stakeholders. Our approach to developing the framework involved the following steps:

- We conducted a desktop research of Open Data maturity models, data best practice and progress frameworks
- We organised workshops with relevant stakeholders such as government departments, regulators, companies, and data users to refine the framework and assessment criteria
- We developed a finalised assessment criteria to measure company progress on Open Data
- We created a comprehensive questionnaire for collecting information from water companies to provide evidence of their work in enabling Open Data

Following company responses, we conducted an evaluation of the submissions against the assessment criteria and prepared one-page summaries (hereinafter "one-pagers") for each company highlighting use cases of Open Data and best practices of each company.

Scope of work and limits

The scope of the work consisted of:

- Develop a robust framework for assessing open data maturity in the sector, which can be used for repeated assessment in future
- Engage with water industry and other stakeholders in development of the framework
- Produce a summary report of findings on progress and examples of good practices, and next steps/recommendations

<u>The scope of our work excluded the following:</u>

- Review or account for any progress made by companies on Open Data after end of submission period
- Review whether companies have met their legal or regulatory obligations associated with release of data
- Provide recommendations on the regulatory response for Ofwat based on the findings of the review

We developed an open data framework and assessment approach to meet the assessment objectives set out by Ofwat

Assessment Objectives

- **Showcase the open data progress companies have made** since the H2Open paper was released in October 2021

- Enable progress across the water sector through **highlighting good practice** from both within the water sector and beyond

- Support Ofwat's next steps in encouraging further progress on open data

Guiding principles

- Draw on **existing good practice** from other sectors

- Be capable of being used as a **self-assessment tool**

- **Repeatable** in future to track progress for individual companies and across the sector

- Easily **interpretable** by users, evaluators and the public
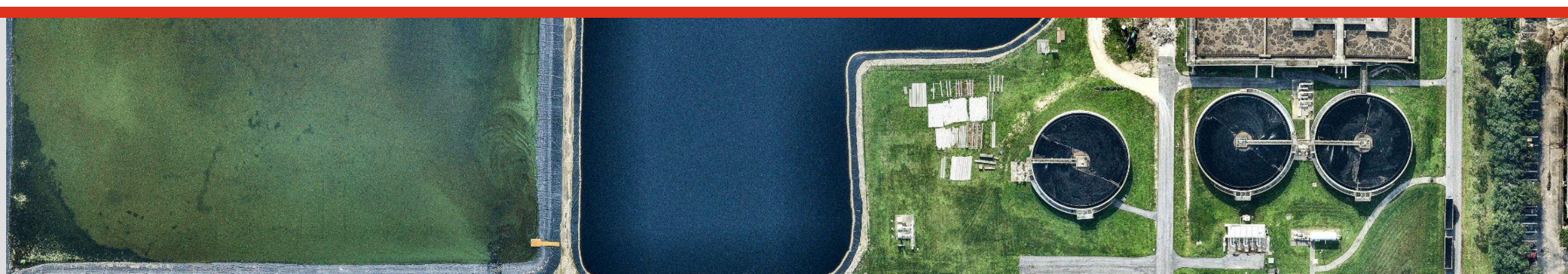
We collaborated with a range of experts whilst developing the assessment framework

The framework was originally developed by drawing on published best practice literature on open data and the knowledge of PwC and Ofwat. The framework was then refined based on feedback collected in two workshops and a series of 1-1 sessions with <u>wider stakeholders, water companies and data users</u>.

Ongoing collaboration between PwC,Ofwat and Open Data experts to create a repeatable self-assessment framework

1. Initial Research

The initial framework structure identifying beneficiaries, outcomes and enablers was developed based on;

- ODI's <u>Open Data Maturity Model</u>'s themes
- Ofgem's <u>Data Best Practice Guidance</u>,
- Catapult Network's <u>Energy Data Taskforce Report</u>
- UKRN's <u>Infrastructure Data Sharing</u>
- Ofwat's <u>H2Open</u>'s Enablers

3. Water Companies Workshop

We engaged with water companies to discuss the types of questions that will be asked to assess progress against the developed open data framework.

The questions were adjusted in line with feedback and a list of potential supporting evidence to be supplied by companies was created to verify company responses.

5. Ofwat's Sign-off and Clarification session

The final Ofwat approved version of the framework was sent to water companies for review and to raise any queries in a clarification session.

Companies were provided a 3 week period to respond to the framework questionnaire.

2. Wider Stakeholder Workshop

We consulted with a range of stakeholders including government departments, independent advisory bodies and regulators to refine the framework structure.

We also identified relevant activities to measure progress against enablers.

4. 1-1 Session Input

1-1 sessions were utilised to gain deeper feedback from regulators and data users to ensure the framework represents the needs of various stakeholders.

Additional details can be found on the following slides

A framework consisting of Activities, Enablers, Outcomes and Beneficiaries was devised, leading to a two-part assessment
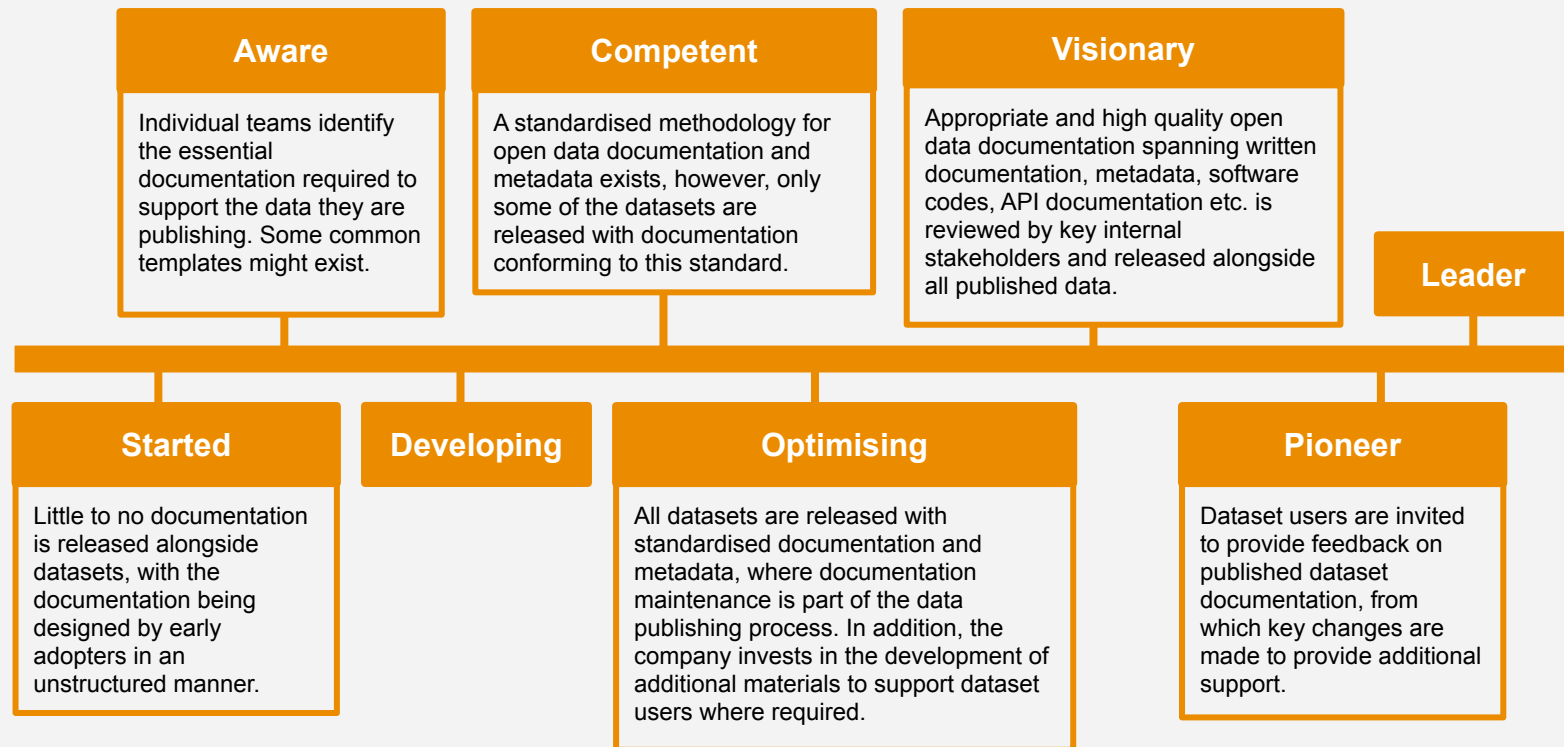
The first section of the assessment asked open ended questions aligned to each of the Outcomes, drawing out case studies of good practice. The second section focused on the Enablers/ Activities, looking to understand the progress that has been made in facilitating Outcome achievement.
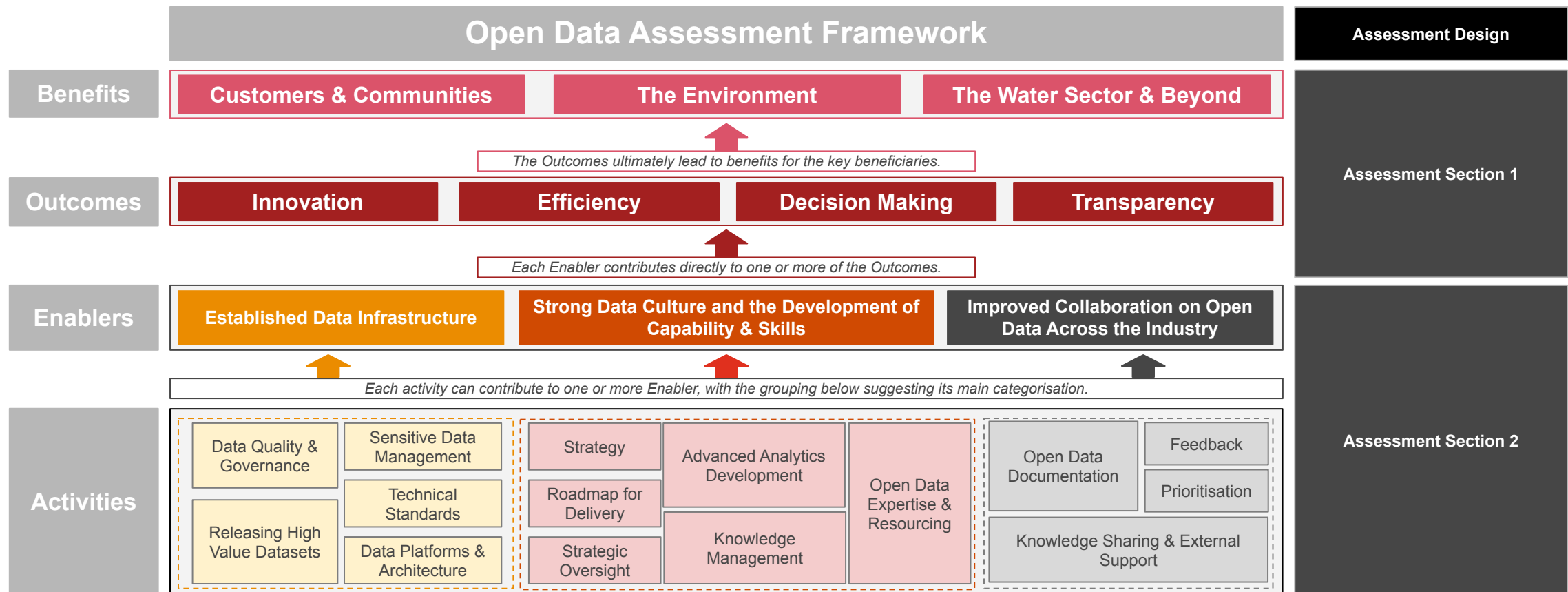
Progress statements were used to map what good looks like for each Activity

To navigate the shared problem of conducting a progress assessment with no baseline to work from, we devised eight progress statements for each activity to define the journey from just starting to becoming a cross industry leader.

The following is an example showing the progress statements for Open Data Documentation:

Aware
Individual teams identify the essential documentation required to support the data they are publishing. Some common templates might exist.

Competent
A standardised methodology for open data documentation and metadata exists, however, only some of the datasets are released with documentation conforming to this standard.

Visionary
Appropriate and high quality open data documentation spanning written documentation, metadata, software codes, API documentation etc. is reviewed by key internal stakeholders and released alongside all published data.

Leader

Started
Little to no documentation is released alongside datasets, with the documentation being designed by early adopters in an unstructured manner.

Developing

Optimising
All datasets are released with standardised documentation and metadata, where documentation maintenance is part of the data publishing process. In addition, the company invests in the development of additional materials to support dataset users where required.

Pioneer
Dataset users are invited to provide feedback on published dataset documentation, from which key changes are made to provide additional support.

Key Design Considerations

- The progress statements were **designed with cross-industry best practice in mind**. Hence, the assessment is a measure against cross-industry best practice rather than the water sector against itself

- The progress statements **provide a benchmark, enabling a repeatable assessment** to measure future progress

Note: Not all activities have a progress statement aligned with all 8 points on the scale, however, if a company response indicates it lies between two non-adjacent progress statements, then the appropriate statement from the full 8 point scale will be used to summarise their performance.

The evidence attached to assessment responses was evaluated and datasets were categorised as either shared or open (1 of 2)

PwC evaluated evidence to make sure it met the required standard, and datasets provided were categorised to be either open or shared based on the Open Data Institute (ODI) definition for open data.

Assessment Approach & Evidence Evaluation

Companies provided written response and attached evidence for each assessment question	PwC evaluated the assessment response and evidence	One-pager progress summaries created based on company response and evidence

- Potential sources of evidence were provided for each assessment question to guide companies on appropriate evidence expected to support their response. (see slide 78 of the main report)

- If evidence against a particular response was deemed insufficient, the case study or the reported dataset has not been mentioned in this report. For instance,
  - where companies have failed to provide any supporting evidence or details about the process or outcome from a given activity in their response.
  - where companies have provided broken links to datasets/supporting evidence published online.

- No additional material outside of the assessment response or supporting evidence was used in the evaluation of companies' progress on open data.

- Industry learnings were devised through aggregating the one-pager progress summaries to understand where the industry as a whole has progressed faster / where it needs to focus its efforts.

Assessment of Progress since H2Open: Companies were informed through the industry workshops and upfront in the framework questionnaire, that the assessment aims to understand the progress companies have made since H2Open. However, the individual questions did not ask about the timeframe for each Open Data related activity/case study that company submits as evidence. As a result, the responses submitted by companies includes applications of Open/Shared data from years preceding the publication of H2Open.

This assessment therefore provides an evaluation of companies progress to date, and establishes a baseline to track progress against in the future. Additionally, where possible, we have highlighted case studies that have either started or made further progress over the past year.

The evidence attached to assessment responses was evaluated and datasets were categorised as either shared or open (2 of 2)

PwC evaluated evidence to make sure it met the required standard, and datasets provided were categorised to be either open or shared based on the Open Data Institute (ODI) definition for open data.

What is open data?

- We have adopted the Open Data Institute's definition for open data, built on the principles developed by the Open Knowledge Foundation
- The ODI defines open data as **data that anyone can access, use and share**
- The Open Knowledge Foundation, whose principles inform the ODI definition, states that open data must be made **available in common, machine - readable formats** i.e. provided in a form readily processable by a computer and where the individual elements of the work can be easily accessed and modified by users
- Open **data must be licensed**. Its licence must permit people to use the data in any way they want, including transforming, combining, and sharing it with others, even commercially

What is shared data?

- Shared data is data made available to specific groups of users
- Access is restricted to specific users via authentication

The Data Spectrum

- The diagram on the right shows the data spectrum developed by the Open Data Institute
- The spectrum is a useful tool in understanding the key differences between open and shared data
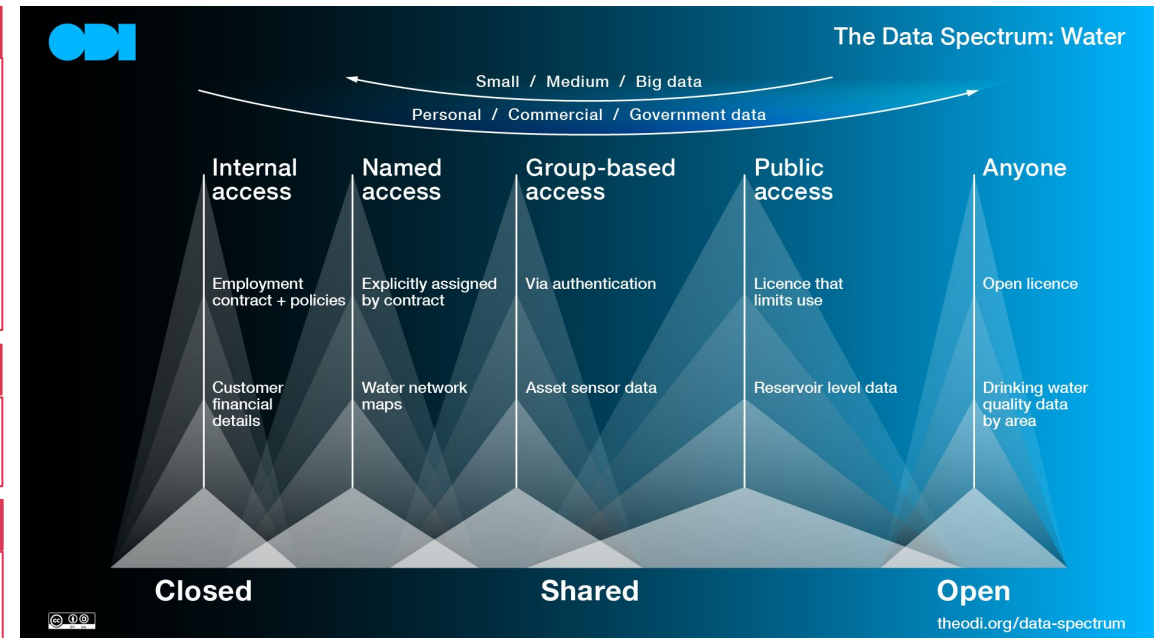
Assessment of Water Company Datasets

- The datasets provided by the water companies on the assessment were categorised into two groups - open or shared.
- Where data was categorised as shared, it was typically due to having restricted access (i.e. not public) or being published in a format that is non-machine readable e.g. pdfs. Our dataset review process first investigated whether the dataset submitted by companies were publicly accessible or not. If a dataset was publicly accessible, a further check was conducted to review if the dataset was machine readable.
- Barring Thames Water's API portal, which provided a data license, an open data license was not discoverable for any of the datasets provided by the water companies as part their response to this framework.
- Therefore, by the definition of ODI, only Thames Water's EDM data published on their API portal will be considered as 'Open'. However, to highlight the number of datasets published by water companies that are **openly accessible** and **published in a common, machine-readable format**, **we categorise any such dataset as open despite a lack of a supporting open data 'license'**.

Our assessment shows that where benefits are being delivered to key beneficiaries, this is typically only through shared data

Water companies have shared data with organisations both within the water sector and beyond to create solutions that provide benefits to the key beneficiaries identified in the assessment framework - Customers & Communities, the Environment, and the Water Sector & Beyond. However, there has been limited evidence of open data releases and hence few benefits achieved from it.

Customers & Communities	The Environment	The Water Sector & Beyond
Customers & Communities, benefit in a number of ways; including:	In response to the current issues facing the water sector, the environment has been a key focus area.	Cross-sectoral collaboration with the water sector on data-sharing initiatives is providing benefits to the whole the economy

Customers & Communities

- **Identification of financially vulnerable customers** and providing targeted services/tariffs **through shared data agreements** between water companies, Department for Work & Pension (DWP) and other utilities companies

- Applications focusing on **sharing maintenance, leakage and water quality information**, e.g. South East Water's In Your Area (IYA) service, are providing increased transparency for customers

- **Near real-time alerts on storm release activity** for example, Southern Water's Beachbouy App*, where alarms and sensors trigger alerts to the public if there is a storm release event

The Environment

- All wastewater companies have **released Event Duration Monitoring** (EDM) data. (see slide 18 for details on publishing requirements). Some companies have taken the next step of making the data more user-friendly. For example, Anglian Water and Thames Water **release the location of their Storm Overflows** alongside their EDM data in an **interactive map** on their website

- **Wessex Water's Intelligent Sewers Competition***, led to the development of a predictive analytics tool in collaboration with StormHarvester that can identify potential sewer blockages earlier. The solution identified >80 blockages over the past year.

The Water Sector & Beyond

- **Streetworks Monitoring and Evaluation tool** developed by Affinity Water in collaboration with Greater London Authority (GLA) has helped in reducing multiple road closures/diversion through sharing of data on planned works programmes with other utilities.

- Initiatives like the **Sewer CCTV AI project***, with multiple companies contributing CCTV footage, help to establish a sector wide repository of CCTV labelled images that companies can use to train their own AI models. These models **replace the manual process of coding defects with an automated solution thereby providing up to 20% of time savings for survey and coding operators**.

Note: Case study examples highlighted with an asterisk (*) were also mentioned in the H2Open publication (link)

# There have only been a small number of open datasets released

Companies have achieved benefits, through varying outcomes, primarily through sharing data with key partners or by merging external third party open datasets with their internal data. A common understanding of the value of open data now exists across the industry but few open datasets have been released
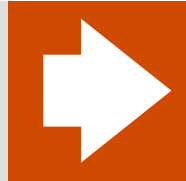
Regular data publications have led to trialling of methods to release data: Companies have used regular data publications to trial methods of releasing data (e.g. Annual Performance Report (APR) and EDM data, see slide 96 of full report for further details), but not all companies have fully met the requirements for Open Data; for instance, by not releasing supporting metadata alongside APR data releases. However, these publications have prompted some companies to trial methods of releasing data that have the potential to facilitate additional data releases in the future. (See evidence on slide 36 of full report)
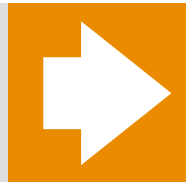
Companies are open data users more than open data providers: Companies are commonly creating benefits for the key beneficiaries by merging data from outside their company with their own data, i.e. they are keen users of open data but have made limited progress in opening of their own datasets. (See evidence on slide 36 of full report)
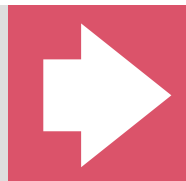
Data is shared rather than making it publicly accessible: When releasing data, companies are generally sharing data with key partners over releasing open datasets with unrestricted public access. Of the datasets referenced by the companies on their assessment responses, 39% (81 of 210) datasets were deemed to be shared due to being restricted to group-based access as per the ODIs open data definition (see slide 8).
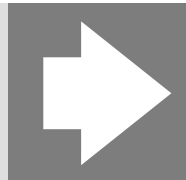
There is a lack of machine-readable open data: Whilst there are some data releases beyond APR and EDM data, these are commonly not done in a machine readable format. For example, through interactive maps with no way of downloading the data are commonplace. Of the datasets referenced by the companies on their assessment responses, 26% (54 of 210) datasets were deemed to be shared due to not being in machine readable format.

Progress in developing open data Enablers has been limited across the water sector, with larger companies further along their open data journey relative to smaller water only companies (1 of 2)

The water sector is at the beginning of its open data journey, with the majority of companies being categorised as Started, Aware or Developing for most Activities.

Enabler Level Learnings

Overall:

- Assessment of the open data Enablers show that the **water sector is still early in its open data development**, with the majority of companies achieving between 'Started' and 'Developing' for most Activities.

- Companies achieving higher progress statements are typically larger and are further along on their digitalisation journeys. Hence, general **Enabler progress is likely driven by broader digitalisation efforts and not solely by open data initiatives.**

Established Data Infrastructure

- **Investment into Data Platforms & Architecture is the most common theme** across the assessment, with many companies currently working towards transitioning towards a cloud-based architecture. This investment is often part of a wider digital transformation with only a few companies making explicit consideration towards how open data may be facilitated on the platform in the future.

- **Releasing High Value Datasets is an area of relative weakness**, where companies have rarely made data open and often fail to provide sufficient metadata or documentation to support releases.

- Some **companies reference that they are engaging with wider collaborative groups** to define/ adopt supporting frameworks and standards. For example, common Platforms, Data Governance Frameworks and common Technical Standards. However, limited evidence was provided by companies to show progress in these areas.

Heatmap showing the Companies' Progress Statements by Activity

		Started	Aware	Developing	Competent	Optimising	Visionary	Pioneer	Leader
Established Data Infrastructure	Data Governance	5	1	3	5	2	0	0	0
	Data Platforms & Architecture	2	0	3	6	4	1	0	0
	Data Quality	1	5	5	2	2	1	0	0
	Releasing High Value Datasets (APR Data)	3	5	5	3	0	0	0	0
	Releasing High Value Datasets (Immediate Challenges. E.g. EDM)	9	1	4	1	1	0	0	0
	Sensitive Data Management	3	10	1	2	0	0	0	0
	Technical Standards	5	2	3	3	3	0	0	0
Improved Collaboration on Open Data Across the Industry	Knowledge Sharing & External Support	2	3	2	8	1	0	0	0
	Open Data Documentation; Feedback	4	8	3	0	0	1	0	0
	Prioratisation	5	7	4	0	0	0	0	0
Strong Data Culture and the Development of Capabilities & Skills	Developing Advanced Analytics	5	5	2	2	2	0	0	0
	Knowledge Management	4	4	5	3	0	0	0	0
	Open Data Expertise & Resourcing	5	4	6	1	0	0	0	0
	Roadmap for Delivery	5	7	4	0	0	0	0	0
	Strategy; Strategic Oversight	3	7	5	1	0	0	0	0

Progress in developing open data Enablers has been limited across the water sector, with larger companies further along their open data journey relative to smaller water only companies (2 of 2)

The water sector is at the beginning of its open data journey, with the majority of companies being categorised as Started, Aware or Developing for most Activities.

| Enabler Level Learnings | Heatmap showing the Companies' Progress Statements by Activity |

Improved Collaboration on Open Data Across the Industry

- Assessment responses indicated that most companies do not have **feedback loops between companies and data users and is an area of weakness** for water companies to improve on.
- **Knowledge Sharing & External Support is an area of relative progress**, with 8 companies scoring competent. This is driven by involvement with collaborative initiatives, however, outputs from these groups were difficult to verify.
- Evidence provided suggests that there are **no collaborative initiatives bringing all 16* water companies together** in relation to open data.

Data Culture and the Development of Capabilities & Skills

- **Strategy, Strategic Oversight & Roadmap for delivery is an area of weakness**, with only 7 companies reporting to have a general data strategy and only 5 of these reporting to having elements specific to open data contained within it.
- **Progress in Developing Advanced Analytics was more common in larger companies** - albeit it focused on the use of other's open data over efforts to open their own data - whereas smaller companies typically opt to leverage external capabilities to support them.

		Started	Aware	Developing	Competent	Optimising	Visionary	Pioneer	Leader
Established Data Infrastructure	Data Governance	5	1	3	5	2	0	0	0
	Data Platforms & Architecture	2	0	3	6	4	1	0	0
	Data Quality	1	5	5	2	2	1	0	0
	Releasing High Value Datasets (APR Data)	3	5	5	3	0	0	0	0
	Releasing High Value Datasets (Immediate Challenges. E.g. EDM)	9	1	4	1	1	0	0	0
	Sensitive Data Management	3	10	1	2	0	0	0	0
	Technical Standards	5	2	3	3	3	0	0	0
Improved Collaboration on Open Data Across the Industry	Knowledge Sharing & External Support	2	3	2	8	1	0	0	0
	Open Data Documentation; Feedback	4	8	3	0	0	1	0	0
	Prioratisation	5	7	4	0	0	0	0	0
Strong Data Culture and the Development of Capabilities & Skills	Developing Advanced Analytics	5	5	2	2	2	0	0	0
	Knowledge Management	4	4	5	3	0	0	0	0
	Open Data Expertise & Resourcing	5	4	6	1	0	0	0	0
	Roadmap for Delivery	5	7	4	0	0	0	0	0
	Strategy; Strategic Oversight	3	7	5	1	0	0	0	0

*Note: Bristol Water and South West Water provided a joint submission in response to the framework questionnaire and have been considered as one for the purpose of this report.

There are lessons to be learned from sectors that are further along on their open data journeys

The transport and energy sectors have made significant progress on their open data journeys and can be looked towards for examples of a greater open data maturity. The development of cross-industry frameworks, centralised open data portals and increased collaboration with data users could help the water sector realise greater benefits.

Examples of good practice	Learnings for the water sector
Development of cross-industry frameworks and data sharing templates	**Adoption of cross-industry frameworks in the water sector**
- The **Rail Technical Strategy (RTS)** highlighted the transport sectors ambitions for improving access and sharing of data across the railway industry. An essential component was the <u>Information Management Framework (IMF)</u>, which included the development of a data standards route map, data-sharing templates and a framework for identifying 'high-value' datasets. **Developing cross-industry data sharing templates and prioritisation frameworks** has enabled the railway sector to move quickly to make stronger use of open data, and has supported the creation of the **Rail Data Marketplace** (RDM); which will combine fragmented sources of railway data to form one digital service.	- The water sector should look to **develop/ adopt similar cross-industry templates and frameworks** to accelerate their path towards a more open data enabled business model. - The companies involved in Stream and CastCo are working towards defining some of these frameworks. Further development of frameworks across the water sector with sufficient external collaboration would help the water sector progress further.
Utilising data portals and APIs to release data	**Developing data portals and APIs to release data**
- Data portals in the energy sector not only provide a centralised source for data users to find multiple datasets, but also act as a feedback mechanism to request additional dataset. For example: the <u>National Grid's Connected Data Portal</u> or the <u>Open Energy Data Portal</u>. - The rail sector is releasing data collected from various monitoring kits through an **Application Programming Interface (API)**. API development has promoted the standardisation and led to the development of third party products and services; for example, travel planner apps.	- The development of **open data portals** for the water sector could accelerate innovation by centralising and signposting the data currently available to potential data users. A data portal could also support the development of an established feedback loop. - Thames Water have developed the first **generic open data API** seen in the water sector. Greater activity in this area, and **sharing of learning**, will see the water sector make considerable progress in open data and support the creation of apps and services in the future.
Collaborating with data users and utilising the feedback loop	**Engage with data users to identify priority datasets and to incorporate feedback**
- The <u>Open Data Communities</u> programme, run by the Open Data Institute, supports collaboration between data users and local authorities to identify and publish open data that is important to local communities. The initiative has been successful in providing useful data to local residents, businesses and organisations, leading to increased transparency and engagement. - <u>OpenActive</u>, a community-led initiative which aims to help people get active using open data, has built an open data ecosystem of over 100 organisations across the sport and physical activity sector, including almost 70 data publishers. It has achieved this through active engagement with individuals working within the sport and physical activity sector.	- Water companies should look to engage with communities and data users to help identify priority data sets. Companies can benefit from having a defined and repeatable approach in place to gather feedback from their data releases. - Wessex Water's Marketplace platform provides a good example of how water companies can both share data and solve common challenges in collaboration with data users.

Next Steps / Recommendations

There are a number of next steps that water companies can take to increase the level of progress on open data across the water sector.

Recommendations to accelerate release of open datasets

As progress in releasing of datasets for public access has been limited across the water sector, companies should look to -

1 Identify priority datasets and set stretching targets to release them such that it accelerates the development of the company's general open data capabilities to support further data releases in the future. Pilot releases of datasets should be conducted and a 'learn by doing' approach should be adopted to iteratively improve on each subsequent data release.

2 Review datasets that have been shared with only specific user groups or organizations and where appropriate assess the steps required to make these openly available.

3 Identify datasets that have been released for public but not in machine readable format and convert these to appropriate formats which can be readily processed by a computer and individual elements can be accessed and modified by users e.g. releasing the data in .csv format.

4 Release datasets with an open license to encourage use and re-use of the information freely and flexibly e.g. <u>License to share and adapt the data</u> provided by Thames Water on its API Portal.

Next Steps / Recommendations

There are a number of next steps that water companies can take to increase the level of progress on open data across the water sector.

Recommendations on development of Enablers

Water companies have started to lay the foundations for open data by investing in data infrastructure, but further progress should focus on development of enablers in a balanced and integrated way -

5 — Companies should develop comprehensive open data strategies, clear roadmaps for delivery with associated delivery KPIs, and set up sufficient accountability and strategic oversight to ensure company board and C-suite buy-in e.g. by tying delivery of strategy to executive performance.

6 — Companies should develop a culture of open data in their organisation e.g. by developing and delivering open data literacy training to create greater understanding of open data across all levels of their organisation.

7 — Industry-wide collaboration is helpful for development of data portals, facilitation of knowledge sharing and development of common standard that enhance interoperability of data. However, such initiatives should be supported with adequate funding, resources and active participation from all companies in order to be effective. Further, priority-based, measurable and time-specific targets should be set to ensure maximum progress is achieved through these collaborative efforts. Such collaboration should also involve engagement with independent experts to draw on learnings from other sectors. Importantly, companies should not rely solely on collaborative groups for ensuring progress.

8 — Companies should actively engage with data users and have a well-defined feedback loop, especially to identify priority datasets and improving quality of data release and associated documentation.

For recommended next steps on each activity in our framework, please see the detailed enabler learnings section of the full report.

Thank you

Annex 1
APR and EDM Data Requirements

Legal requirements and regulatory expectations associated with the publication of APR and EDM data by Water Companies

Requirements on Event Duration Monitoring (EDM) data publication	Expectations on Annual Performance Reporting (APR) Data publication
Currently, under section 141D of the Water Industry Act 1991 the Environment Agency (EA) publishes an annual report on storm overflows (containing EDM data) annually by 1 April. Under section 141C of the Act, companies are also required to publish annual reports on storm overflows annually by 1 April. The Act stipulates that the reports published by companies in general must be "readily accessible to the public." In March 2022, Ofwat asked wastewater companies how they planned to tackle storm overflow and river quality issues. As part of their response, some companies committed to release live sewage discharge notifications or near-real-time data on storm overflows by 2023 (Northumbrian, South West, Thames, and United Utilities), 2024 (Wessex), and 2025 (Dwr Cymru, Yorkshire).	In March 2022 Ofwat set out expectations (link) for companies to make 2021/22 APR data available in a way that can be reused and accessed by all, by publishing their APR tables in an Excel format on their website and publishing metadata such as publication date, version information and the file type, and supporting documentation such as explaining changes in methodology from previous years.