OPEN FOR FEEDBACK

WORK IN PROGRESS

What data publishers need: research and recommendations

Synthesis of user research

Open Data Institute

Contents

About

This report has been researched and produced by the Open Data Institute. It was published in January 2018 and **last updated in May 2018**. Its lead authors were Myriam Wiesenfeld, Phil Lang and Olivier Thereaux, with contributions from Caley Dewhurst, Tom Hunter, Anna Scott, Dave Tarrant, Emily Vacher, Cai Williamson and Jeni Tennison.

If you would like to send us feedback or comment on this document, please get in touch by filling this online form.[1] Or if you want to share feedback by email or would like to get in touch, contact the publishing tools project lead Olivier Thereaux at ot@theodi.org.

WORK IN PROGRESS

This is work in progress. It is likely to be updated as we continue our work. Keep an eye out for updates!

OPEN FOR FEEDBACK

How can it be improved? We welcome suggestions from the community in the comments.

[1] See: https://goo.gl/forms/91PmG0pGjAoEUTnj1

Key

User needs

Link between need & issues

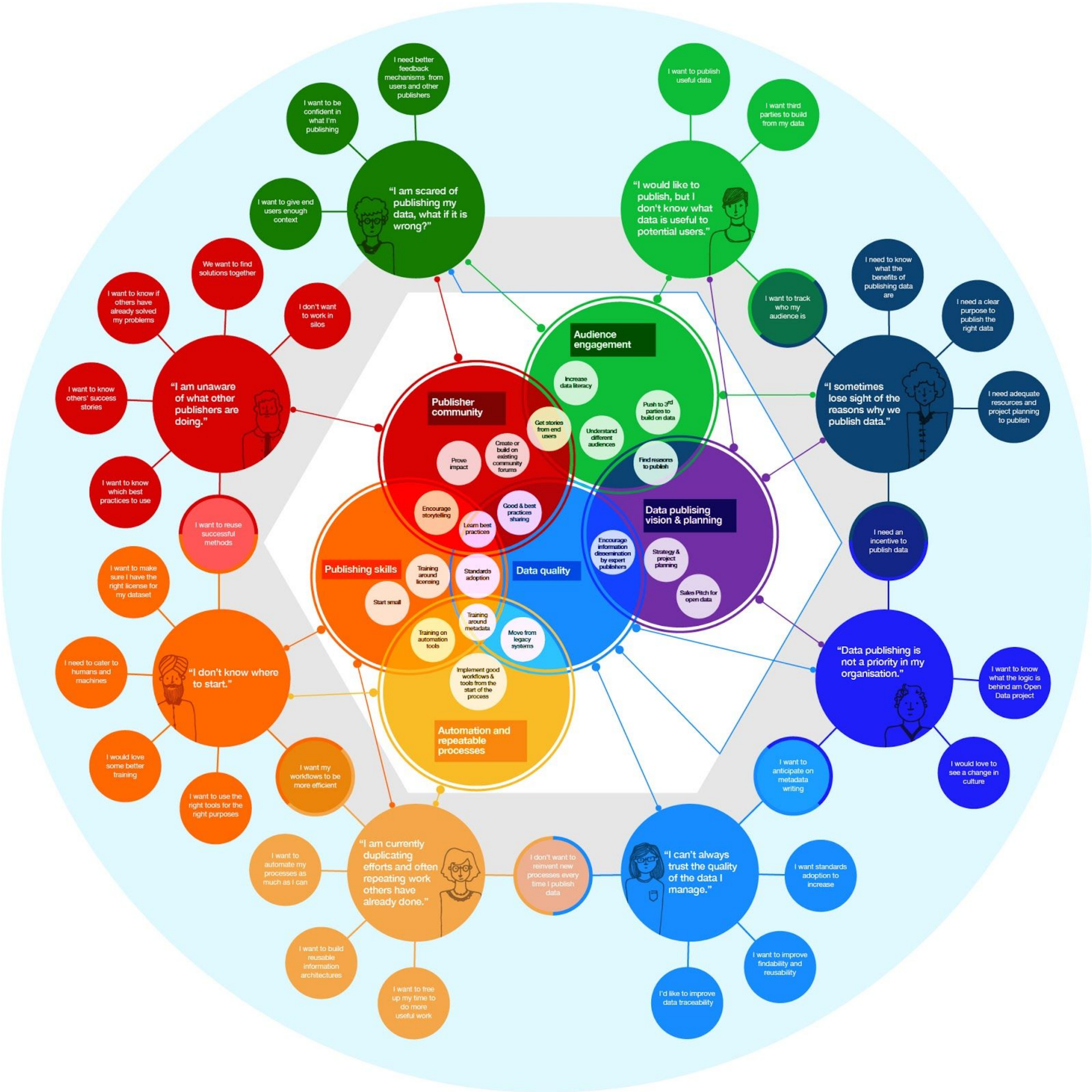
Overarching issues

Recommendations

Organisations and people publishing open data use a variety of activities and tools as they create, gather, clean up, process, describe, vet and publish data, so others can access, use and share it.

When talking to publishers about how we can improve data quality, and reduce publishing costs and time, the ODI found a number of common issues around their unmet needs.

This diagram gives a high-level view of those needs and issues, and some of the proposed solutions we discovered through this research. You can download it here.



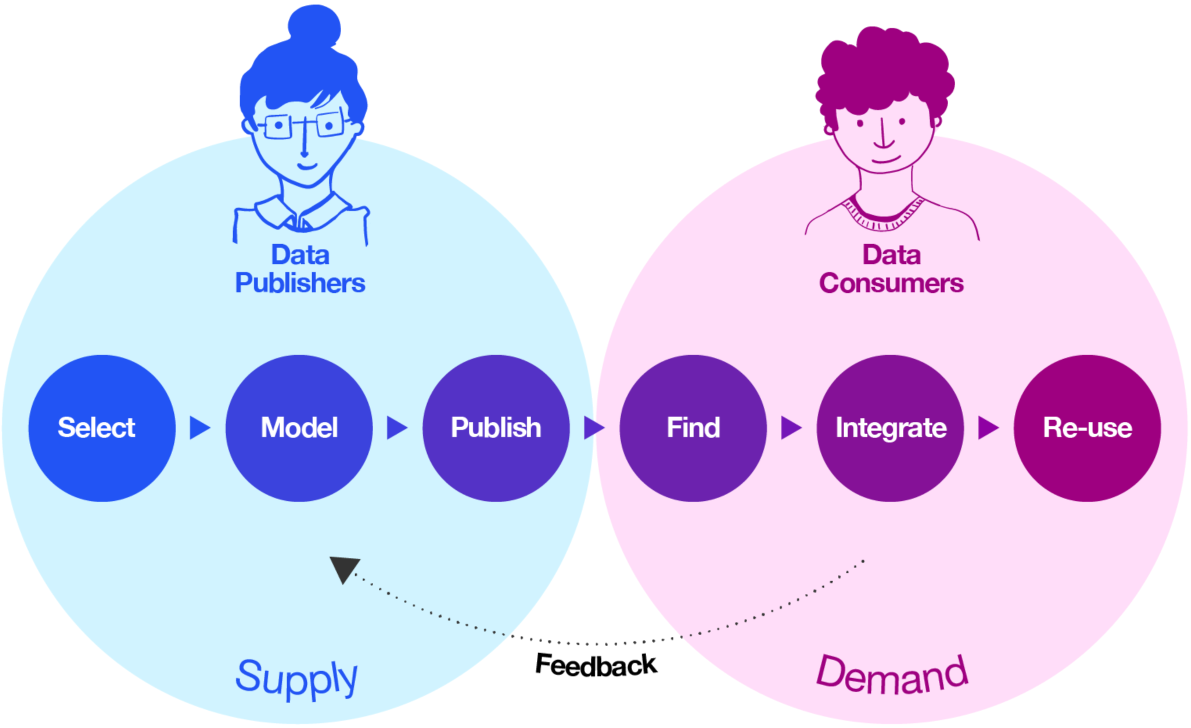
"I am scared of publishing my data, what if it is wrong?"
I want to give end users enough context
I want to be confident in what I'm publishing
I need better feedback mechanisms from users and other publishers
"I would like to publish, but I don't know what data is useful to potential users."
I want to publish useful data
I want third parties to build from my data
I want to track who my audience is
"I sometimes lose sight of the reasons why we publish data."
I need to know what the benefits of publishing data are
I need a clear purpose to publish the right data
I need adequate resources and project planning to publish
I need an incentive to publish data
"Data publishing is not a priority in my organisation."
I want to know what the logic is behind am Open Data project
I would love to see a change in culture
I want to anticipate on metadata writing
"I can't always trust the quality of the data I manage."
I want standards adoption to increase
I want to improve findability and reusability
I'd like to improve data traceability
I don't want to reinvent new processes every time I publish data
"I am currently duplicating efforts and often repeating work others have already done."
I want to automate my processes as much as I can
I want to build reusable information architectures
I want to free up my time to do more useful work
I want my workflows to be more efficient
"I don't know where to start."
I want to make sure I have the right license for my dataset
I need to cater to humans and machines
I would love some better training
I want to use the right tools for the right purposes
I want to reuse successful methods
"I am unaware of what other publishers are doing."
We want to find solutions together
I want to know if others have already solved my problems
I don't want to work in silos
I want to know others' success stories
I want to know which best practices to use
Publisher community
Prove impact
Create or build on existing community forums
Get stories from end users
Encourage storytelling
Learn best practices
Good & best practices sharing
Audience engagement
Increase data literacy
Push to 3rd parties to build on data
Understand different audiences
Find reasons to publish
Data publising vision & planning
Encourage information dissemination by expert publishers
Strategy & project planning
Sales Pitch for open data
Publishing skills
Training around licensing
Start small
Standards adoption
Data quality
Training around metadata
Training on automation tools
Move from legacy systems
Implement good workflows & tools from the start of the process
Automation and repeatable processes

Introduction

Between July and October 2017, the ODI conducted user-focused research to help understand and explain how open data publishing practices could be improved.

How do people and organisations create, gather, clean, process, describe, vet and publish open data so others can access, use and share it?

As one of the first activities in the ODI's 2017 Research and Development programme, we set out to better understand the life of those on the supply-side of the open data lifecycle.



This focus was motivated by a belief that this aspect of the data ecosystem – open data publishing – is often taken for granted, but that its success is equally key to creating a robust data infrastructure.

Plenty of time and effort is spent advocating that open data be published with appropriate licences, rich descriptions and metadata, using efficient formats and standards. What if those outcomes could be radically increased by improving the experience of those people publishing this data, creating this metadata, and choosing these formats and standards?

We set out to test three hypotheses – that quality, speed and cost-effectiveness of open data publishing could be improved through better tools and automation.

Research question

How can better publishing tools improve the quality, speed and cost-effectiveness of open data publishing?

Research methodology:

The research in this discovery phase aimed to help us understand:

- who are the main actors in the open data publishing landscape?
- what pain points do open data publishers face when publishing data?
- what are the potential solutions to these pain points?

The research included two phases: an audit of existing open data publishing tools, and user research to understand the needs of publishers.

The audit looked at more than 30 different tools: who they are for, what type of data they are used for, what business models they operate on, how mature they are.

The schema of categories is available here; the publishing tools register is available here.

The user research focused on understanding data publishers via workshops and interviews, including with the companies that built publishing tools. The ODI talked to 60 people, a majority of whom were from the UK public sector, with some representation from the private sector and outside of the United Kingdom. We collated this research with one of our partners in Northern Ireland (Lintol), who the ODI commissioned to create personas that represent individuals actively involved within the open data supply chain.

When talking to data users and publishers, the ODI found that their needs varied but all revolved around the same themes: how to improve data quality, reduce its publishing cost and the time spent on it.

Our research mostly focused on what support could be given to open data publishers to meet both their own needs and those of data users.

Actors in the open data publishing landscape

The research identified three key groups of actors in the data-publishing landscape: publishers, with varied expertise; the people and organisations developing tools to help publishers; and data users – consumers of the published data, whether people or machines.

Open data publishers

‘Data publishers’ can range from full-time expert publishers to hobbyists. Throughout the research, the ODI found three broad clusters of behaviour within this group.

i) Enthusiastic novice publishers

This group is new to publishing but keen to open their data. Their intentions are not always matched by the skills and knowledge they have.

ii) Expert publishers

These have been publishing for several years and remain enthusiastic and motivated. They champion the concept of open data.

iii) Disillusioned experienced publishers

This group has been publishing for years but their initial enthusiasm has waned and their motivation for publishing has diminished.

Open data consumers

While not the main focus of our research, data consumers are the ones ultimately benefiting from better, more effective data publishing. They are a broad group, at different stages of maturity, from novices to seasoned data publishers. Their needs are varied and they have different approaches to finding and reading data.

i) Novice data users

Novice data users will often have a project they believe data could help with, but need help understanding the data they are accessing.

ii) Expert value creators

These are highly skilled and know how to extract value and build products and services with open data.

iii) End users

End users, in the context of this report, are the people and organisations benefiting from the use of data. They are worth distinguishing from the data users above as they may never see or use the data directly, or even be aware of its existence.

iv) Machines

Machines are a key data consumer that should not be overlooked. There is a need to increase machine readability of data to make it more useful - increasing interoperability.

Tool developers

Tool developers in many ways determine what publishers can do. They fall into two categories.

i) Commercial tool developers

Developers of data publishing tools want to help their users in publishing and working with data by creating tools that empower the user in these areas.

ii) Open-source tool developers

Open source tools lower the barriers associated with successfully publishing data - democratising the ability to share quality data for reuse. Developers of such tools commonly draw focus on the community that surrounds their creations, as such communities foster growth and adoption of their tools and practices.

Mapping the needs of data publishers

Looking at eight data publishing archetypes, and the issues they face in their role.

The research behind this paper began with identifying existing open data publishing tools[2]. It quickly became apparent that a holistic view of data publisher needs was required. Existing tools satisfy many publisher needs, such as data cleaning, automatic validation and visualisation. This section aims to focus on publisher needs that we have found to be unmet, and areas where there are opportunities to make interventions and build new solutions.

The research identified eight key user-needs for publishers that are currently unmet (or only partially met). All of these are, to varying degrees, relevant to the three publisher profiles identified above. When it is particularly relevant to one type of publisher, this is described below.

1. “I don’t know where to start”

This user-need is particularly relevant to publishers who are new to open data publishing. Because of the very rich ecosystem of existing tools and the varying levels of technical proficiency required to use them, non-experts struggle to know what to do when starting to publish. Knowing which format to use, how to create the right information architecture, what a successful open data project looks like, etc, can be difficult.

Novice publishers need support to navigate the difficulties of publishing open data: understanding how to get the best value out of the data they control, who to cater it to, or how to make it machine readable and human-friendly at the same time.

Interviewees were often excited about the idea of publishing as they appreciated its value. The first hurdle was finding the right tools to meet their needs, and then understanding and using them correctly.

The need for good, clear workflows is not limited to novice open data publishers. For those who have been publishing open data for longer, the challenge often revolves around finding the correct workflow that will be compatible with the tools that they use, and vice versa. The process of publishing can be clunky and frustrating as it often involves jumping between tools that do not easily integrate with each other. Legacy systems also play an important role in the frustration of more experienced publishers who wish they could improve the way publishing is done by implementing new and better work structures.

[2] See goo.gl/8aFqTc

2. “I sometimes lose sight of the reasons why we publish data”

A common issue shared by interviewees is that the “why” we publish open data question can often be forgotten: publishing becomes a repetitive task that has to be done, rather than one with a clear goal.

Publishers feel that there is a directive coming from above that states that data must be published openly; this directive comes without reason, guidance, planning or real resources.

As open data publishing is often an additional, temporary assignment on top of existing responsibilities, it can be frustrating not to get insight or reason for the task.

This is combined with insufficient feedback on how much datasets are used by the audience (downloads, usage, reuse, API creation based on datasets…), which is discouraging for data publishers who sometimes feel that they push open data into a vacuum.

3. “I would like to publish, but I don't know what data is useful to potential users”

Prioritising which datasets to release and maintain is a difficult challenge when resources are limited. As different audiences do not have the same needs, levels of expertise or maturity, data published incorrectly not reaching the right end-users reduces the impact and benefit of the published data.

Our research also shows that data is typically made more human-readable to help data publishers and data consumers. This often comes at the expense of machine-readability, resulting in poor interoperability and making the data typically harder to find.

These factors lead to a recurring issue of incentivising open data publication: “if no one will find this data, then why should my team waste time, effort and money publishing it?”.

About half of the data publishers interviewed told us they spend a lot of time trying to understand how their publication approach can have the greatest impact. Understanding who their audiences are and how to help them overcome their challenges is a significant consideration in the publication process.

4. “I am scared of publishing my data, what if it is wrong?”

A persisting issue was the fear of publishing due to potential errors in datasets. A lot of publishers seem to think that every publication must be perfect. Striving for perfection can lead to a fear that is paralysing, especially in larger projects where no data gets published unless a hard deadline is set by senior leaders.

It is important to note that the community of open data users are often forgiving of imperfect data. They prefer to see imperfect data being published, rather than no data being published.

5. “I am currently duplicating effort and often repeating work others have already done”

The lack of standard methods to publish similar datasets was a common frustration in the workshops and interviews. For example, local authorities or city councils publishing the same types of data didn’t appear to use the same approaches, tools or workflows.

Similar Freedom of Information (FOI) requests, for example, are often answered multiple times by separate workers, leading to the frustration of government publishers who feel they are duplicating effort.

A lack of interoperability often restricts the ability for publishers to collaborate as they use different tools and processes, and reduces the chance that the data will be beneficial, as data consumers will not want to duplicate their work dealing with similar, but un-interoperable, data.

6. “I am unaware of what other publishers are doing”

From data publishers to tool developers, we found that all actors in the data publishing process put an emphasis on how collaboration facilitates a successful journey through publishing.

Publishers vary widely in their levels of expertise. Some are great advocates for the open data publishing cause, disseminate their knowledge and take it upon themselves to help others. Some don’t know that there is a wider community of publishers facing similar issues and challenges.

Premium publishing platforms often put a lot of time and effort into creating discussion forums and communities where the platform users can ask questions to their peers for answers and solutions. However, a lot of publishers do not have access to these premium platforms or don’t know that communities of this sort exist.

7. “I can’t always trust the quality of the data I manage”

As there is a lack of standards adoption, little use of automation tools and sub-optimal metadata entry, the quality of published datasets is not always high. Data publishers often receive their data, in an unrefined format, from various departments in their organisation. It is left to the individual(s) responsible for publishing to clean and refine this data. Limited resources and a lack of tools to guarantee data quality across multiple formats and use cases makes it nearly impossible to maintain a high standard of quality across all of the datasets they receive.

Publishers said they would like to have ways to mark their datasets as being high-quality, aligned to certain standards and flagged as usable, up to date, etc. Such mechanisms could foster trust in the quality of published data, but there are not many tools to do that easily.

8. “Data publishing is not a priority in my organisation”

When discussing the challenges in publishing open data, many publishers mention culture change before technical considerations. They often bemoan their small budget and the (at times) indifferent attitude towards open data within their organisation.

Some publishers feel they are not being taken seriously by the people in charge. They have a need to plan data publishing as ‘real’ projects in order to get the right resources, understand challenges, set goals, get to know their audience, and work out what type of data to publish and how.

Understanding the issues and exploring solutions

The eight user-needs identified fit into larger themes, and share a number of root causes, some of them technical, many cultural, others organisational.

A. Insufficient publishing skills

Which user-needs does this relate to?

- **1.** "I want to start publishing data, but don't know where to start"
- **5.** "I am currently duplicating effort and often repeating work others have already done"

Data publishers have great intentions and want to publish, but often lack the proper skills and technical expertise. The workflows they employ are often complex, employing multiple tools that don't integrate well and require varying technical skills.

Data publishers can be very technical but are equally likely to be uncomfortable with tools requiring them to understand complex technology or code. For some, their day job is to publish data; for others, it is an additional side-task. Publishers don't necessarily always know where to start and how to approach data publishing projects.

Helping publishers to learn and employ the correct skills and tools from the beginning would help create logical workflows, better approaches and adoption of appropriate standards. This would allow publishers to learn and therefore decrease the time it takes to publish subsequent datasets.

Key solutions

- **Start small.** Tools providers should encourage publishers to begin with smaller datasets, making it easier for the novice publisher to go through the learning curve.
- **Training around licensing.** A better understanding of licensing options and best practices increases the potential for use and reuse of the data.
- **Make standards more accessible.** Adopting standards is key to creating impact with open data, making it easier for users of the data to adopt it. However, standards are not well-known, and are hard to find. Increasing awareness of existing standards could be a 'quick win'.

- **Training around metadata.** Metadata is key to promoting open data, but is often tacked on at the end of the publishing process rather than thought through at the start.

B. Inadequate vision and planning for data publishing

Which user-needs does this relate to?

- **2.** “I sometimes lose sight of the reasons why we publish data”
- **8.** “Data publishing is not a priority in my organisation and is therefore under resourced.”
- **3.** “I would like to publish but I don’t know what data is useful to potential users”

An organisation's vision for publishing data should aide in team planning, get the correct resources, decide on which tools to use and which form the data should take. Publishing teams can decide on a strong information architecture, anticipate metadata needs and which licensing to use ahead of time. Having a robust plan will help publishers understand the bigger picture around their project and therefore incentivise data-publishing.

A plan will also help quantify the end use of the datasets, which in turn will give publishers feedback on whether or not their approach is legitimate. It can also allow to publish less data at a time, in order to test it with a known audience and therefore iterate on the first datasets, and improve their quality and value to data users.

Key solutions

- **Strategy & project planning.** A project that is planned correctly ensures the use of the right tools at the right time. A strong vision for open data publishing plan can help preempt all kinds of pitfalls from licensing, to metadata creation or highlighting data provenance. Clear open data project plans also become repeatable and more trackable.

- **Sales pitch for open data.** Publishing open data as a business strategy, like any idea, needs to be sold internally. Developing a short elevator pitch to use throughout your organization can do wonders to get the message across.

- **Find reasons to publish.** Publishers don’t always have strategic reasons around who they’re publishing for and why, which can be a deterrent to publishing in general and to qualitative publishing in particular. Having a strategy around a publishing project which is aligned with your organisational mission and strategy will give publishers the drive to publish.

C. Low trust in the quality of published data

Which user-needs does this relate to?

- **7.** “I really need to be able to trust the quality of the data I see”
- **4.** “I am scared of publishing my data, what if it is wrong?

One problem publishers and users face is trusting the quality of the data they publish and use. There is a need for data standards to be adopted more and for data and metadata to have more consistent structure.

Because many publishers do not plan how to publish their data, they often miss selecting the right type of licence, anticipating metadata creation or highlighting data provenance. Moreover, there is no obvious way of discerning good data from bad data: few datasets come with quality or standards compliance badges. This is an obstacle to data being trusted, findable, interoperable and machine readable.

Lack of standards adoption is a real problem; this happens partly because standards are seen as too complex and difficult to comply with. Sometimes the problem is simply that the standards are not well known and are hard to find. There is a need for standards to be better exposed along with what they should be used for and how.

Giving publishers incentives to publish data according to certain standards in order to get them started and help them understand how to use those standards that exist, could prove useful.

Data standardisation could improve quality and interoperability, in turn increasing machine-reading potential, better automated visualisation programs and general findability.

Adhering to standards, organising publishing projects thoroughly, anticipating known challenges (like metadata or licensing) would help reduce fear of publishing. If a publisher follows guidelines, it is harder to make mistakes, easier to pinpoint errors in datasets and ultimately easier to publish and get constructive feedback from users.

Key solutions

- **Standards adoption.** Standards are seen as complex and hard to comply with. Changing this perception could go a long way to ensure a more consistent adoption across (and within) organisations. Through standardisation, quality and interoperability would improve automatically, increasing machine-reading potential and findability.

- **Encourage information dissemination by expert publishers.** Some leaders in open data publishing have gained fantastic insights from their experience. Sharing that knowledge more widely would be hugely beneficial.

- **Move from legacy systems.** While this switch can be time consuming and daunting at first, the benefits of moving beyond out of date systems are undeniable.

D. Lack of automation and repeatable processes

Which user-needs does this relate to?

- **5.** "I am currently duplicating effort and often repeating work others have already done"
- **1.** "I want to start publishing data, but don't know where to start"

Publishers do not want to spend their time repeating the same tasks over and over. There is a need for systems automation: tools to clean or validate data, to publish it visually, to create consistent metadata, etc.

By reducing the number of manual processes, there would (theoretically) be fewer mistakes made and the published data would be more homogenous.

Automation also allows publishers to save time once they have published their first datasets correctly. There will of course always be some new challenges to overcome and new automation systems to put in place. Automating the easy components is therefore crucial to focus on new and more complex obstacles.

Automation is also repeatable and shareable: if data publishers find great solutions to some known problems that other organisations have, that technique can be shared and remove pain points for a lot of other publishers.

Key solutions

- **Implement good workflows and tools from the start of the process.** The current process of publishing often involves jumping from tool to tool, which don't always integrate with one another. A good workflow will help address this issue.

- **Training on automation tools.** Tools such as OpenRefine, fusion tables or GoodTables offer automation of data quality checks. Learning these tools is a relatively quick process, which could save time and mistakes compared to a manual data quality assurance process.

E. The publisher community is not accessible to novices

Which user-needs does this relate to?

- **6.** "I am unaware of what other publishers are doing, and often miss out on best practices"
- **4.** "I am scared of publishing my data, what if it is wrong?"

Many publishers seem to work on their own without knowing that many of their peers may share their issues, worries and ideas.

Forums and communities exist where publishers share good methodologies, best practices, tips and ideas, as well as guidance on how to publish better data. Online communities such as these can be beneficial in bringing publishers out of their silos. According to the publishers interviewed in this research, the communities are hard to find and do not reach out enough to novice publishers.

Key solutions

- **Create or build on existing community forums.** Interactions on these forums encourage peer-learning between communities of open data publishers and users.

- **Encourage storytelling.** Stories excite the community about the potential of open data and provide a way for publishers to sell open data within their organisations, and therefore push for more resource for their team.

- **Prove impact.** Statistics on the economic and social value, and job creation from open data help maintain the momentum and drive behind publishing.

- **Learning and sharing best practice.** Good habits can spread more quickly if publishers and tool-developers are learning from one another and sharing both their successes and mistakes.

F. Not enough engagement with data consumers

Which user-needs does this relate to?

- **2.** "I sometimes lose sight of the reasons why we publish data"
- **3.** "I would like to publish, but I don't know what data is useful to potential users"
- **4.** "I am scared of publishing my data, what if it is wrong?

When the processes to clean, validate, license and publish data have been improved, the next step is ensuring that the consumers publishers have targeted as their audience are engaged and empowered to use the published data. This can be a challenge as it means improving data literacy of consumers, providing inspirational stories and streamlining access to data.

The steps required to meet these challenges are broad. The production of content that focuses on conveying stories of successes and learnings offer paths that users can follow, aiding people on a journey of learning and exploration with data. There is also a technical element where delivery of data can be improved by building mechanisms that facilitate such delivery.

By taking a broad and open approach, engaging with potential data users can not only increase the interaction with published data, but also act as a vehicle to improve data literacy and increase the dispersal of inspiring data stories.

Key solutions

- **Increase data literacy.** Strengthening the ability to read, create and communicate data will unleash the potential of open data and ensure reuse comes from a wider pool of innovators.

- **Understand different audiences.** Data publishing is not ‘one size fits all’, and different audiences will have different needs in terms of formats and readability.

- **Push to third parties to build on data (e.g. APIs, apps).** These provide both the impact and the stories that will make open data publishing sustainable in the long-run.

- **Stories from end users.** Understanding how end-users are using open data to solve problems and build applications strengthens the reasons to publish and makes it easier to sell open data internally.

This is a living document

The needs and practice of open data publishing will evolve over time, as the ecosystem matures.

Created as part of the ODI's innovation programme, this document aimed to help our team prioritise our efforts, both for the development of new or existing tools and the support we could offer to others, making sure we also strengthen the network of organisations creating a stronger and better data infrastructure.

The insights generated from our user research were key to awarding support to the Lintol[3] and Frictionless Data[4] projects, and supporting them as they navigated their development roadmap. The research made it evident that extra effort should be put into integrations, workflows, and generally making the tools used by open data publishers easier to use, by novices and veterans alike.

We also applied the learnings documented in this report to drive the development of one of the ODI's own toolbox. One of them, called Octopub, had been developed earlier as an exciting experiment in publishing data to the Github platform. We used Octopub as a platform to refine our understanding of the issues explored in this report, and vice-versa, exploring how to create a truly useful and usable tool for teams and individuals to work through the crucially important steps of preparing data for publication.

As we publish a new version of this report, we know that this work was only the beginning. The true impact of the development initiatives spurred by this user research will only become apparent in the future, and we intend to document the changing user needs in future iterations of this document - offering, hopefully, a fuller and up to date view of evolving landscape and practice of publishing data.

This will therefore be, for now, a living document and we plan on documenting changes and revisions in the appendix below.

If you would like to send us feedback or comments on this document, please get in touch by completing this online form.

[3] A project undertaken by a team of partners in Northern Ireland. See https://lintol.io/

[4] A project by Open Knowledge International. See https://okfn.org/ and https://frictionlessdata.io/

Appendix 1
Versions and changelog

Version 1.0 — 21 November 2017

- Initial revision

Version 1.1 — 22 May 2018

- Editorial updates to improve flow and legibility
- Shortened and updated section on the actors of data publishing. Made language used to describe actors more consistent throughout
- Re-organised sections on understanding the issues and proposed solutions into a single section
- Made the list of issues (now in section “Understanding the issues and exploring solutions) read more like a list of issues than a mix of issues and solutions
- Updated the illustration of user needs and issues to match new wording in the report
- Updated final section based on recent developments. Renamed last section