OPEN DATA IN HIGHER EDUCATION: AN INTRODUCTORY GUIDE

Universities UK and Open Data Institute

INTRODUCTION

This introductory guide is intended for professionals working in universities who wish to gain a better understanding of the issues associated with using open data within their institutions. It offers advice on how to be strategic in your use of open data, guidance on making the case for open data within your institution, and provides a set of activities for building data skills. The guide is based on training workshops run by the Open Data Institute (ODI) for the 'Creating value from open data' project.

The project was led by Universities UK in partnership with the ODI and Jisc. It set out explore the value of open data in UK higher education and promote a better understanding of how universities can benefit from it.

We ran a series of workshops with professional groups and universities to examine the opportunities and challenges associated with open data in higher education. We learned that open data could potentially be applied to better utilisation of space, student engagement via learner analytics and sharing research equipment. However, a clear focus emerged on the role open data could play in providing services to students.

With this in mind, we ran an Open Data Mashup Day on 17 November 2015 bringing together university managers, student developers and data experts to explore how open data can enhance the student journey. The day culminated in an award by Jisc of support for the development of the winning 'BookMart' app pitched by John Lacey, a graduate trainee at Lancaster University.

I hope that you find this a useful introduction to open data. For more information visit our project google plus site (sign up required) or contact me.

Ian Powling
Digital Programmes Lead, Universities UK
ian.powling@universitiesuk.ac.uk

Open Data Institute, 3rd Floor, 65 Clifton Street, London EC2A 4JE, UK.
http://www.theodi.org

Universities UK, Woburn House, 20 Tavistock Square, London, WC1H 9HQ.

HOW TO BE STRATEGIC ABOUT YOUR USE OF OPEN DATA

Organisations releasing open data often have the same experience when taking their first few steps on their open data journey. Having identified a first dataset to release, the initial focus is typically on the logistics of getting the data published. This covers technical considerations such as formats, but also issues such as licensing and documentation.

This first step is usually a learning exercise for the organisation that involves absorbing more than just the details of how best to publish open data. It often shines a light on existing organisational processes and activities. For example: how is the organisation's data collected and managed? Who owns that data? And what happens when users ask for help or additional data? Answering these questions is an important aspect of developing an organisational strategy for open data.

The Open Data Institute (ODI) has developed two freely available online tools that can help organisations take a strategic approach towards developing their open data practice.

Open Data Certificates

The Open Data Certificates application assesses how well a dataset is being published for reuse by third-parties. By completing a questionnaire, publishers can benchmark their publications against best practices in the following areas:

- **Legal** – what can people do with the data? This covers rights, licences and data protection
- **Practical** – can people rely on the data? This covers timeliness, quality control and guarantees around availability
- **Technical** – how easy is the data to work with? This covers machine readability, suitability of format, whether it uses open standards and URLs as identifiers
- **Social** – are people going to get help working with the data? This covers documentation, feedback to the publisher, community and tooling

The primary output of the process is a badge and a certificate that rates the overall effort being invested into publishing the individual dataset. These examples of a raw certificate and an expert level rating provide some insight into the variety of elements covered by the certification process. The certificate provides a useful report card that summarising the answers to the questionnaire, and the badge can also be embedded into a publisher's website.

The tool also creates an assessment report that provides series of specific actions that a publisher can undertake in order to improve how the dataset is being published. This provides a basic roadmap that can inform further work on this and future datasets.

Open Data Pathway

Open Data Pathway is the second assessment tool developed by the ODI. Rather than focusing on an individual dataset this tool is designed to allow an organisation to assess its overall open data practice across a number of different themes:

- **Data management processes** – e.g. data governance and release processes
- **Knowledge & skills** – open data skills and knowledge management
- **Customer support & engagement** – supporting and interacting with data users
- **Investment and financials** – impacts of open data on procurement and approaches to valuing datasets
- **Strategic oversight** – how open data aligns with and informs organisational strategy

The tool is based on the Open Data Maturity Model which breaks down these themes into 15 different activities and processes. The breadth of the assessment reflects the various ways in which managing, prioritising and supporting open data releases can impact an organisation.

Pathway is also designed as a simple questionnaire which guides users through a maturity assessment, allowing evidence and notes to be added along the way. The report (example) provides a maturity rating from 1-5 for all of the activities and themes and also includes an action plan that suggests steps for improvement.

A maturity assessment provides an excellent baseline against which organisations can assess themselves and also benchmark with peers. The results can be used to create a strategic plan that can guide the development of open data practice in the organisation.

A key first step in that process is building an asset catalogue that identifies key data assets that an organisation is both producing and consuming. The catalogue provides a starting point for making other strategic decisions, including prioritising open data releases, eliminating duplication between datasets, and improving data governance.

Used together, the certificates and pathway tools provide useful guidance for organisations that are at any stage of their open data journey.

How to be strategic about your use of open data slides

THE CASE FOR OPEN DATA

Many organisations are interested in sharing their data but are unsure of how to articulate the associated costs and benefits necessary to make their first steps. Often a focus is placed purely on open data. But it's important to realise that there is a data spectrum consisting of closed, shared and open data.

Some data, including personal and commercially sensitive information should never be published openly. Other data might be shared with a limited group, e.g. with researchers, using specific data sharing agreements. However other datasets might have most impact when published openly for anyone to access, use and share.

As data moves along the spectrum from closed, shared, and to open, more people will have access to it. Greater access provides more opportunity for that data to be used in new and innovative ways. It also becomes possible to combine the data with other sources to yield new insights.

With this in mind, a better framing when making a case for open data is for an organisation to consider:

- what data do we have?
- what data do we already share, internally or with third-parties?
- are there benefits in sharing this data more widely, including publishing it under an open licence?

The Open Data Institute has published a guide that discusses how to make a business case for open data. This guidance highlights a number of key elements to consider, including:

1. **Alignment with organisational goals** – would sharing data help deliver or demonstrate delivery of your specific objectives?
2. **Reviewing current data sharing practices** – how is data being shared already, and is there a case for making it more open?
3. **Identifying internal and external stakeholders** – who owns and manages data internally, and with whom is it being shared already?
4. **Cataloguing data currently being collected?** – what data do you collect and consume as an organisation; and could this be supplemented or replaced with open sources?
5. **Asking how will value be generated?** – would sharing data deliver on transparency objectives; or are there potential cost savings or other economic benefits?

For example, a university may improve delivery of its organisational goals by openly publishing its list of courses, events and equipment. Data on student performance might already be shared with government bodies or researchers, there may be value in sharing this more widely, in aggregated forms.

Addressing Common Challenges

Open data advocates are often faced with a number of challenges and concerns. Anticipating these common questions can help identify additional areas that should be addressed when building a business case.

The issues can be grouped into five broad categories:

- **Concerns over the value of organisation data assets** – this ranges from concerns that organisational isn't interesting or is over poor quality, through to a desire to directly monetise data by selling it
- **The effort to publish and share data** – that publishing the data is too difficult or costly, or that if it were shared then this would only increase demand for new data
- **Legal concerns** – that the rights to share or openly publish data are unclear, or lack of clarity around the merits of different types of licence
- **Fears over misuse of data** – concerns that data might be misused or misinterpreted
- **Concerns with reusing third-party data** – that external datasets, particularly open data, is of lower quality

Several attempts have been made to catalogue the variety of questions along with common answers that may help alleviate concerns.

The Open Data Institute provides a range of freely additional guidance that can also help explore and address these questions, including the following documents:

- How to make a business case for open data
- Engaging with reusers
- What are the impacts of non-open licences

The ODI has also started publishing a range of open data stories and detailed case studies that provide more insight into how a variety of organisations are benefiting from open data.

The case for open data slides

INTRODUCTION TO DATA SKILLS

Getting the most out of your data involves three core skills: cleaning, enriching and visualising your data. This section will give you a brief overview of these skills, and will take you through some practical exercises to get you started.

Cleaning data

Using clean data is an important part of making sure the right conclusions are made. Often errors are unnoticed by data publishers because the data can change over many years. In other cases, errors can be the result of human mistakes in data entry, such as mistyping or incorrect abbreviations.

Be aware that cleaning may often need to take more time cleaning your data than carrying out the subsequent analysis. A careful cleaning process will save substantial time in the long-run.

Following our data cleaning exercise enabled attendees to get started with Open Refine and discover how easy it is for themselves.

Exercise - Cleaning your data

For this exercise you will need a copy of Open Refine (http://bit.ly/1F5VLKu) and an unclean dataset (http://bit.ly/1n8gQg0)

1. Loading your dataset:

To get started, import your dataset into Refine.
This can be done either by uploading the file or pasting the URL of the file.
Once your file has been imported you will need an import options screen. Use this to ensure your file looks correct before naming your project and clicking 'create'.

2. Common Transforms:

Wrong data format? White spaces where there shouldn't be? Numbers that don't appear to be numbers? These are all common dataset errors. Open Refine has the tools to fix these errors. Find the columns that have these types of errors and select **Edit cells** -> **Common transforms** from the column drop down to fix them.

3. Numbers:

Commas can help us to read large numbers, but to a computer this appears as a comma separated list of many numbers. To get rid of commas in Open Refine select **Edit cells** -> **Transform…** from the column drop down and enter the following in the expression box value **replace(',','')**

4. Text:

Spelling errors? Misplaced apostrophes? All common errors in datasets. To find these errors in your dataset try putting a text facet on a column of text select **Facet -> Text Facet** from a column drop down**.** Open Refine can also help fix errors automatically. Try pressing the Cluster button from inside the facet and adjust the options.

Enriching data

One dataset is rarely enough to discover insight. By combining multiple datasets, it is possible to get a more detailed picture. For example, a course reading list is useful in itself as it tells the student what books to read. But combining this with other data, such as library catalogues, could help students make more better decisions on how to access those books. (For example, see: BookMart, winner of the UUK & Jisc Open Data Mashup Challenge, November 2015.)

Open Refine (see above) is a useful tool for combining and enriching datasets using online services and API endpoints.

Exercise - Enrich your data

Use the following exercise to enrich your cleaned dataset:
Enrich data in Refine (http://bit.ly/1OSd2fg)

Visualising data

Visualisations are a useful way of interpreting data, and they help us to unlock insight.

Faceted browsing (or cross filters) allow consumers to filter the data visually. These browsers are commonly used by holiday and flight websites, such as Sky scanner to help filter search results or change input parameters. The exercise below shows you how to create a faceted online browser.

Exercise - Create a Faceted Online Browser

The following exercise uses the clean and enriched dataset to create a faceted online browser.

Get started with visualisation (http://bit.ly/1Rp4vB6) in conjunction with this link (http://bit.ly/1PTfgKg).