Exercise: Discovering Open Data

The idea of this exercise is to help you gain a better understanding and be able to critically appraise the different types of open data publication practice on the web. While publishing data may just be about meeting targets, publishing usable data that can be consumed by others in a reliable fashion is entirely different. The aim of this exercise is to take a critical view on the ability to effectively consume open data and identify a number of key areas for improvement.

Throughout this exercise you should think exclusively as a consumer of open data. When analysing each dataset think about the ease of using the data, availability and opportunities to answer complex questions and perform data analysis. Questions may include:

- Do I understand what the dataset is?
- Am I able to access the data?
- Am I permitted to use the data?
- Do I understand the data itself?
- What questions can I answer using the data? Is the data too granular, too generic?
- Is the dataset supported long term?
- Is the data consistent?
- Is the data clean?
- How much effort would be required to make the data usable?
- Can I get support on the data and find what else it has been used for?

Once you have analysed the dataset as a consumer, think about how you would change the publication practices of the datasets in order to improve them. Do the best practice guidelines help here? Would you recommend this style of data publication to others?

Exercise Datasets

Dataset 1: Spend over £25,000 in the Foreign and Commonwealth Office
Location: http://data.gov.uk/dataset/financial-transactions-data-fco

Dataset 2: *Your own*
Location: *Please explain how you discovered the dataset when presenting your findings.*

Dataset 3: Land Registry Monthly Property Transaction Data
Location: http://data.gov.uk/dataset/monthly-land-registry-property-transaction-data

Exercise 2 - Data of the Web

This exercise requires you to use a tool that allows you to browse a “web of data”. As in exercise 1, there are a number of datasets for you to analyse, critique and compare.

For the purposes of browsing the data we shall be using the graphite tool available at the following URL:

http://graphite.ecs.soton.ac.uk/browser/

For the purposes of this exercise you should use the **URI** listed with graphite and the **URL** to browse to the location of the data in a normal web browser.

Dataset 4: University of Southampton: Building 32
URI and URL: http://id.southampton.ac.uk/building/32

Before answering the questions set out in the first exercise think a little about the reasons for publishing data “on the web” in this format. Firstly is the data **reference** or **transactional** data? Does the type of data make it more suited to this type of publication?

Be sure to browse the data both via graphite and as normal in a web browser. What are the differences between the types of representation? Is it clear what the relationship between the two is?

Finally, think about how you would improve this dataset and if you would recommend this style of publication to others?

Once done take a look at these two additional examples and compare and contrast the three examples. Which example is the best one in your opinion?

Dataset 5: BBC Music Data
URL: http://www.bbc.co.uk/music/artists/d24fb461-dee8-41fc-bb15-2f13bb2644a6
URI : http://www.bbc.co.uk/music/artists/d24fb461-dee8-41fc-bb15-2f13bb2644a6.rdf

Extension Exercise - Hybrids

Exercise 1 looked at how data is being published “on the web”, while exercise 2 looked at how we can create a “web of data”. There are a number of examples where an attempt has been made to fulfil both aims.

The aim of this exercise is to critically analyse a hybrid approach to discover if it solves the problems the other methods introduce, or compounds them.

Dataset 6: Land Registry Monthly Property Transaction Data (Exercise 1)
Location: http://data.gov.uk/dataset/monthly-land-registry-property-transaction-data

In order to view this data in the form needed for exercise 2, we require a URI, something that is not easy to discover from this data. The following URI shows all the applications in England and Wales divided by authority in August 2013. There is an equivalent CSV you can compare with this.

URI: http://landregistry.data.gov.uk/data/trans/applications-by-region?countPeriod=2013-08

Dataset 7: Elevation benchmarks in Chicago
URI & URL: https://data.cityofchicago.org/resource/zgvr-7yfd

Once you have analysed all the datasets, which one do you feel was the best approach for publishing open data and why?