

Open Access Week Seminar Series

University of Adelaide - October 2011

Open data – access to and re-use of data

What it is, where it is, how you can find it and use it, why data should be open and some interesting case studies and examples.

Initial presentation on Thursday 27 October, 3:00pm – 4:00pm, Ira Raymond Room

Prezi can be viewed at <http://prezi.com/pkbfhjlibag4/open-data/>

To move through the Prezi use the forward arrow on screen or use the right arrow key on your keyboard.

<INITIAL PREZI SCREEN – OPEN DATA>

What is open data?

This week the University Libraries are conducting a seminar series about Open Access. Open Access Week is now in its fifth year, and the open access to the results of scholarly research is growing. But what about open data?

Making data available for reuse for different purposes, including reuse in research, is not a new idea. However, although open data is a recognizable concept it is not yet formally acknowledged in declarations or statements in the same manner as Open Access. Therefore no single definition for open data exists. For today's purposes, I will be referring to open data as data that are freely and immediately available for reuse. Open data can include data that are a product or outcome of research as well as data that are collected or generated by public and private entities or individuals. The scope of open data is very broad.

Open data offers a range of opportunities and possibilities but also raises some important questions. This talk will provide an introductory overview of open data; I will not be covering open data standards, data interoperability, data mining or open APIs. But today I will present some examples of open data and data reuse, consider some of the questions that arise with open data, look at some reasons why data should be open, and consider how open data can impact you and the University of Adelaide. The presentation will run for approximately thirty minutes and there will be time at the end for questions and comments.

<NEXT PREZI SCREEN - DATA.GOV.AU>

Examples of open data

One readily identifiable example of open data is <http://data.gov.au>. This website provides a service for finding, accessing and reusing public datasets from the Australia Government as well as state and territory governments. The website uses tagging and categories to aid discovery of datasets. Each dataset has a data licence that sets out the terms of its reuse. Relatively speaking, this Australian initiative is in its infancy compared with international equivalents. It currently contains just over 720 datasets compared with over 7,500 datasets

from <http://data.gov.uk>. Interestingly one of the most downloaded datasets from the Australian website is the National Public Toilet Map! Other top downloaded datasets include Commonwealth Electoral Boundaries and NSW Crime data.

The website states that its main purpose is “to encourage public access to and reuse of government data” and users of the datasets are encouraged to share the tools and applications they create. A list of apps can be found on the website.

<NEXT PREZI SCREEN – SUBURBAN TRENDS>

One such example is www.suburbantrends.com.au. The Suburban Trends product is a mashup based on a range of publicly available datasets. It allows users to search on suburbs by name or postcode, and then to compare different suburbs by viewing graphical indicators on different aspects of each geographical area. It includes indicators of education levels and perceived safety levels.

In the screenshot you can see in the main part of the screen a map of Adelaide with the city and its immediate surrounding suburbs highlighted. These suburbs are listed in the table on the right of the screen and the indicator that we are viewing in this case is ‘perceived safety using transport’. The perceived level of safety is represented for each suburb, where a longer bar indicates higher levels of perceived safety. As you can see Unley as the highest level, and up the top Adelaide’s level is lower than Unley’s yet higher than North Adelaide shown just below it. If you look at the ‘About’ link from the bottom of the page it does provide information on the data sources and explains that the data for this particular indicator is based on a question from the 2004 International Crime Victimization Survey from the Australian Institute of Criminology, which asked respondents to rate how safe they felt waiting for or using public transport in their area after dark.

This is just one example of reusing data in applications or mashups, but there are several ways to reuse data. Datasets can of course be used in research. The Australian Bureau of Statistics springs to mind as a source of data that can be reused in the course of study and research. All statistics on the ABS website are freely accessible and have been since 2005. I am not going to look at that today because it is such a well known source of open data, but we will look at some other collections of research data that can be used in further study or research.

<NEXT PREZI SCREEN – ATLAS OF LIVING AUSTRALIA>

One such example is the Atlas of Living Australia. The Atlas of Living Australia is a national database of Australian flora and fauna, bringing together biodiversity information and data through support from several partner organisations including CSIRO, museums and the Australian Government.

In the screenshot you can see the results of browsing the website by Local Government Areas, and in this case we’ve drilled down to the Adelaide Hills region. There are several categories listed and on the screen that relate to this region, and I have collapsed them all except the frogs window. There are eight relevant species listed, and six with images that appear there. Clicking on any of those images or links will bring up further data including an

overview, images, accepted and common names, scientific classification and more. This is just one view of the Atlas of Living Australia but there are several other ways of exploring the data.

<NEXT PREZI SCREEN – AUSSTAGE>

In a completely different discipline is AusStage: Gateway to the Australian Performing Arts. AusStage is another collaborative initiative that provides data on live performance in Australia. It includes data on events, venues, contributors, organisations and resources. The University of Adelaide Libraries has contributed to the information in this resource.

In the screenshot you can see an entry for A Midsummer Night's Dream performed at The Playhouse in June and July 1992. It lists actors and other contributors involved in the production. At the bottom of the screen you can just see the beginning of the list of reviews related to this production.

<NEXT PREZI SCREEN – PHIDU HOME>

Another data sharing initiative even closer to home is the Public Health Information Development Unit, which is located here at the University of Adelaide. The Unit provides data atlases and graphical representations of Australian health and wellbeing in a socioeconomic context. In addition to the data being available for download, PHIDU offers interactive graphic software that allows the data to be displayed on a map. Another feature available through the website is the 'Find Your Area' data module, that allows you to type in a postcode to see for which corresponding Statistical Local Areas or Local Government Areas there are data available.

<NEXT PREZI SCREEN – PHIDU DATA>

The screenshot shows an example of the map functionality from PHIDU, in this case looking at a representation of data on dwellings in South Australia with no motor vehicle in 2006, sorted by statistical local area. The darker areas on the map represent statistical local areas with a higher percentage of dwellings with no motor vehicles. The legend in the top right hand corner of the screen explains the percentage range associated with each colour. The data is also available for download and there are several interactive features on the website that can be explored. There are some videos on the PHIDU website that I encourage you to view if you are interested in a more detailed explanation.

There are countless examples of databanks and mashups that provide access as well as tools and services that provide new representations of the raw data. It would be impossible to cover all the categories of open data and example of data reuse in every discipline. Let's look at just one more example of data reuse.

<NEXT PREZI SCREEN – SAHULTIME>

This product is called Sahul Time and is the result of project from Monash University. It provides a reconstructed representation of the ancient Earth based on sourced data. Sahul Time provides an interactive map that allows the user to view the earth across a time span

up to 100 thousand years ago. It uses data from GeoScience Australia's oceanic surveys and the GEBCO Digital Atlas from the British Oceanographic Data Centre.

In the screenshot I have dragged the red bar at the top of the screen to 60 thousand years ago and the image on the map depicts a satellite-style reconstruction of the coastline at that time based on the oceanic data. There are some other features to be explored in SahulTime also.

Questions for open data

The examples we have just looked at show the very tip of the iceberg of the possibilities for reusing data. Reusing data presents different challenges in comparison to reusing other outputs such as publications. The copyright associated with reusing publications is quite clear and the practice of citing publications is well established; rights and responsibilities with publications are generally understood. But how does this translate to reuse of data?

<NEXT PREZI SCREEN – CREATIVE COMMONS>

Copyright does not protect facts or information but it can protect databases, compilations, tables and forms. Therefore to aid reuse of data, making the terms and conditions of its reuse clearly documented and available allows others to readily reuse the data within the terms and conditions specified, without concern that they are breaching copyright. Datasets are often made available under a non-exclusive licence, such as those licences on data.gov.au, which are in fact Creative Commons licences. The ABS introduced Creative Commons licensing in 2008. On the screen is one of the Creative Commons licences under which one of the data.gov.au datasets is made available.

Licences may require that the author of the dataset is attributed. Acknowledging the source of the data may seem fairly logical, but how this is done is not as straight forward or as established-a-practice as citing a publication in a bibliography. In research communities the practice of citing datasets is growing but data citation standards vary across disciplines. Developing a culture of data citation is important and can only be of benefit to the open data movement as receiving acknowledgement and attribution for one's datasets provides further incentives to share one's data.

These questions of copyright, licences and data citation should not be barriers to open data but they are important considerations for the open data movement and when managed effectively they should promote the uptake of open data.

Why should data be open?

Perhaps the easiest way to see the benefits of open data is to look at some open data success stories. Let's take a look at a short video from TED Talks, given in 2010 by Tim Berners-Lee, the inventor of the World Wide Web. This talk is entitled 'The year open data went world-wide'.

<NEXT PREZI SCREEN – TED TALKS VIDEO>

(http://www.ted.com/talks/tim_berniers_lee_the_year_open_data_went_worldwide.html)

<NEXT PREZI SCREEN – ANDS COST BENEFITS REPORT>

I love this video because it captures so many benefits of open data for communities and individuals. It is just a short video but it provides a powerful snapshot of the value of open data. There are also more formal investigations into the benefits of open data. This year the Australian National Data Service commissioned the Centre for Strategic Economic Studies at Victoria University of Technology to report on the costs and benefits of open public sector data. The report, by Professor John Houghton, was published last month and models costs and benefits of open data for a number of government agencies. While the benefit to cost ratio varied by agency, in every case it demonstrated that the benefits considerably outweigh the costs of open data. The report is available through the Australian National Data Service website if you would like to look at it.

The report includes lessons for the research sector, although it acknowledges that research data are different from public sector data. Making research data open and available for reuse allows other people to take the data and represent it in new ways, draw new conclusions, and generate new outcomes. Open data can save time and money in research.

The value of open data is certainly being recognized by funding bodies including the National Health and Medical Research Council (NHMRC). Their Australian Code for the Responsible Conduct of Research, which has been adopted by the University of Adelaide, has an entire section on management of research data and primary materials and includes the statement that “research data should be made available for use by other researchers unless this is prevented by ethical, privacy or confidentiality matters.”

<NEXT PREZI SCREEN – DRYAD>

One example of open data within research is Dryad, an international repository of data related to peer-reviewed articles in the basic and applied biosciences. Dryad is associated with a number of journal partners. On the screen is an example of a dataset associated with an article published in the Journal of Heredity. This dataset is also made available under a Creative Commons licence. The uptake of research data repositories like Dryad varies considerably across disciplines.

<NEXT PREZI SCREEN – RESEARCH DATA AUSTRALIA>

On a national level, open data is being advocated by the Australian National Data Service. ANDS is building the Australian Research Data Commons, which is accessible through Research Data Australia as shown on the screen here. Research Data Australia is not exclusively populated with open data, as it includes descriptions of data collections that are not necessarily open although contact details may be available for requesting permission to access the dataset. The University of Adelaide is contributing descriptions of research data collections to Research Data Australia, with three ANDS funded projects at the moment.

What can you do?

There are countless stories of data reuse; I hope some of those that I have shared today have inspired you, but how does open data apply to you?

Firstly on reusing others' data, consider the opportunities for accessing and using open data and if and when you do reuse data, consider the impact and benefit and attributing the data authors through appropriate data citation. If you are a producer of data, consider how, when and where your data can be made open, and how it might be licensed and available for reuse, and the guidance you can provide to others on correctly citing your datasets. Depending on your area of research there may be different tools or resources that can support you in this.

Resources and further reading

Data.gov.au <http://data.gov.au/>

Data.gov.uk <http://data.gov.uk>

Suburban Trends <http://www.suburbantrends.com.au/>

Atlas of Living Australia <http://www.ala.org.au/>

AusStage <http://www.ausstage.edu.au/>

Public Health Information Development Unit (PHIDU) <http://www.publichealth.gov.au/>

SahulTime <http://sahultime.monash.edu.au/>

The year open data went worldwide

http://www.ted.com/talks/tim_berners_lee_the_year_open_data_went_worldwide.html

Costs and Benefits of Data Provision: Report to the Australian National Data Service by Professor John Houghton, ANDS, 2011 <http://ands.org.au/resource/cost-benefit.html>

The Australian Code for the Responsible Conduct of Research

<http://www.nhmrc.gov.au/guidelines/publications/r39>

Dryad <http://datadryad.org/>

The Australian National Data Services <http://www.ands.org.au/>

Research Data Australia <http://services.ands.org.au/home/orca/rda/>

Engage: Getting on with Government 2.0

<http://www.finance.gov.au/publications/gov20taskforcereport/index.html>

Australian Copyright Council INFORMATION SHEET G060v10 'Databases, Compilations, Tables & Forms' from <http://www.copyright.org.au/>

ANDS guide: Copyright and data <http://www.ands.org.au/guides/copyright-and-data-awareness.html>

ANDS guide: Creative Commons and data <http://www.ands.org.au/guides/cc-and-data.html>

ANDS guide: Data citation <http://www.ands.org.au/guides/data-citation-awareness.html>