Open Data Management in Agriculture and Nutrition

This e-learning course is the result of a collaboration between ***GODAN Action*** *partners, including* ***Wageningen Environmental Research*** *(WUR),* ***AgroKnow****,* ***AidData****,* ***the Food and Agriculture Organization of the United Nations*** *(FAO),* ***the Global Forum on Agricultural Research*** *(GFAR), and* ***the Institute of Development Studies*** *(IDS),* ***the Land Portal****,* ***the Open Data Institute*** *(ODI) and* ***the Technical Centre for Agriculture and Rural Cooperation*** *(CTA).*

GODAN ACTION
Global Open Data for Agriculture & Nutrition

UKaid
from the British people

GODAN Action is a three-year project UK's Department for International Development to enable data users, producers and intermediaries to engage effectively with open data and maximise its potential for impact in the agriculture and nutrition sectors. In particular we work to strengthen capacity, to promote common standards and best practice and to improve how we measure impact. [www.godan.info]

UNIT 2: USING OPEN DATA

LESSON 2.1: DISCOVERING OPEN DATA

Aims and learning outcomes

The lesson aims to provide a foundation in how to discover and access data that is available on the web. From data downloads and data service providers to publishers of linked open data, this lesson will cover the complete toolkit to obtain the right data from the web, faster.

After studying this lesson, you should be able to:

- *list different types of services that provide access open data*
- *list different methods by which data can be accessed from these services.*
- *explain the difference between these types of services*
- *use the different types of services to access open data*
- *discover downloadable data*
- *discover hidden data*
- *identify whether a data source is open data for these types of services*
- *describe the advantages and disadvantages of these types of services.*

Contents

List of figures

List of tables

1. Introduction

As the web has evolved, so has its continued use for sharing ever more complex resources and data, but it challenges existing paradigms. The World Wide Web was central to the information data and is seen as an information space where documents and other web resources are accessed (World Wide Web).

Seeing the web in this way has led to the development of a web of documents, or webpages as they are more commonly known, designed for humans to access and read. About 4.75 billion of them [http://www.worldwidewebsize.com]. Many of the documents are linked together, and those connections add value (Hyperlink). In a blog, newspaper article or academic paper, we can use links to build on previous discussions or point to factual sources. They help us to explore the web of documents.

2. The Generations of the web[1]

There were many fewer documents in the early days of the web but people still needed to be able to discover things. The first efforts at manually maintaining an index[2] were performed by Sir Tim Berners-Lee, the ODI's President and Co-Founder[3]. People could go to the list and then jump to pages that looked interesting.

We then created portals, such as DMOZ[4] and the early Yahoo![5] These were curated lists of websites and pages organised by particular topics. As the web scaled up, portals were no longer viable, and people moved to metadata search engines, such as Altavista[6] and Lycos[7], which used metadata that had been manually set in the webpage and provided information about the document.

Search was more scalable because pages were discovered automatically, but results were unreliable and easily manipulated. The next generation of web discovery came with PageRank-style[8] search, such as Google, which used many more cues for search, including an understanding of content, usage and

[1] https://theodi.org/blog/we-need-to-learn-how-to-search-the-web-of-data
[2] https://www.w3.org/History/19921103-hypertext/hypertext/DataSources/WWW/Servers.html
[3] https://theodi.org/team/timbl
[4] https://www.dmoz.org
[5] https://en.wikipedia.org/wiki/Yahoo!_Directory
[6] https://en.wikipedia.org/wiki/AltaVista
[7] https://en.wikipedia.org/wiki/Lycos
[8] https://en.wikipedia.org/wiki/PageRank

linking. This third generation learnt how to look within the web of documents to discover how relevant each document would be for users.

It was vital that we learnt how to build these different types of search. The web of documents could not scale until search[9] became the primary means of discovery. All of these methods still exist, and meet different needs, one of them being to fulfil the requirements of putting open data on the web.

3. The Generations of open data on the web

The early open data publishing techniques mirror the same generations as those of the web of documents. We first created portals, such as data.gov.uk, opendata.go.tz and data.sncf.com. These are curated catalogues of datasets organised by particular topics or organisational structures.

As the amount of open data begins to grow, data aggregators begin to appear that provide services to ease the discovery of data related to particular topics or regions. Examples include enigma.io[10], European data portal[11] and transport API[12]. Such services rely on the availability of metadata from other portals, websites and servicers to provide information about and access to the data.

The third generation of search engines for data is still very focused on the web being an information space, thus we have to rely on the same search engines to find data as well as information. Development of search engines for data is still in its infancy and this is greatly related to the methods being used to publish data on (or in) the web.

4. Data is just another resource on the web

As the web evolved, it started to become a place to share multimedia resources. The inclusion of images, audio and video unlocked new potential to deliver services such as streaming services. Audio was the pioneer here, where early sites like last.fm used metadata about music tracks to build customised recommendations for people. This technology precedes and forms the basis of many of the recommendation systems in use today. Last.fm provided recommendations but didn't allow people to listen to the music itself. This functionality didn't emerge until 3 years later in 2005 with the launch of Pandora, a personalised radio station application. Another 3 years on saw the

[9] http://onlinelibrary.wiley.com/doi/10.1002/1097-4571(2000)9999:9999%3C::AID-ASI1607%3E3.0.CO%3B2-F/full
[10] https://www.enigma.com
[11] https://www.europeandataportal.eu
[12] https://www.transportapi.com

launch of Spotify, which was the first streaming service which used web technologies to deliver a dedicated audio platform outside of the web browser.

Wind forward and the web and internet are now the delivery platform for huge amounts of different resources. Sticking inside the web browser, search engines like Google have added multimedia-specific search capacity to find images and videos, while specific portals like YouTube now provide web-based access through their website as well as via applications and connected TVs.

The same is not quite true of data. Search engines still do not have specific searches for data, perhaps they never will. But finding data on the web can be a challenge, one that starts with the definition of data itself.

What is data?

An image is a visual representation of something (a picture). An audio file makes a sound when played. A video combines multiple images with audio to make a moving picture. Data is… difficult to define and thus provide search for.

Data is the lowest level of abstraction from which information and knowledge are derived. These are abstract terms and thus data could be an image, or spreadsheet, or audio file. Additionally if the web is an information space then data is something which is of a lower level?

Traditionally, data is thought of as a spreadsheet or set of numbers that can be analysed in some way. On the web, such data is often shared via data portals simply as a file that can be downloaded. Some portals provide YouTube-like functionality where the data can be explored without downloading, however the data itself is still a static resource, uploaded ready to be downloaded by someone else.

While data remains a static resource, second- and third-generation web services are perfectly suited to harvest metadata about these static resources and provide entries in search results. Static resources can be linked to directly and thus algorithms like PageRank remain relevant.

This approach suits the web of documents approach, where the metadata is still the key way in which the data can be found. This means that existing search engines can be used to find data if you can ask the correct query.

However not all data is static.

When you visit a shopping website, travel website or weather website. The content is likely to be different each time depending on the raw data. On a shopping website, certain products may vary in price and stock level. On a

travel website, options will vary in availability and price depending on the search criteria while the weather changes constantly.

The data that is powering these sites is vast and hidden in the web. Machine-readable data is creating a new web of data which experts claim has the potential to unlock a data age. Perhaps the web then transforms from an information space into a data and information space.

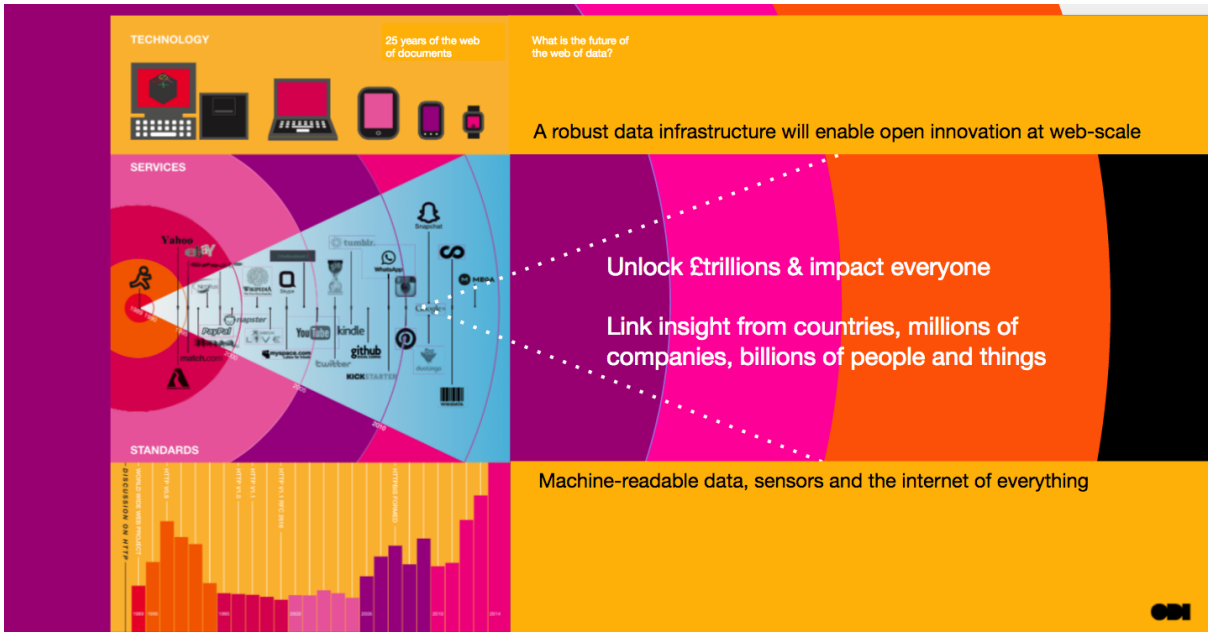

Figure 1 The expanding scope of the web

Applications that use such data are already prevalent, from travel planners, weather and shopping apps to games. Such applications exchange data to help them function, however this data is often hidden, making it difficult for others to access and use.

The rest of this lesson explores the two approaches of data on the web of documents and how to search it before exploring the web of data and how to begin to unlock its potential.

5. Government portals

The history of open data is closely connected with laws that govern access to public information. Such laws ensure that the public has a right to access information from those providing public services. The open data movement attempts a complete reversal of this logic. Rather than having to request data, that data should already be 'open by default'; to close a dataset should require a good reason rather than the opposite. Governments and public service providers should work in the open proactively.

Going a step further, governments in many countries signed up to the Open Government Partnership (OGP). OGP was launched in 2011 to provide an international platform for domestic reformers committed to making their governments more open, accountable, and responsive to citizens. A key part of OGP is the commitments on being 'open by default' and open data. This lead to the launch of many government open data portals that were built to hold government data and make it easily accessible for the public.

Over the years, open data activities have evolved to the extent where governments are now scored on how well their activity is going and how sustainable it is.

The Open Data Barometer (ODB) by the World Wide Web Foundation scores governments in three aspects: readiness, implementation and impact. The implementation score is measured by looking for key open datasets to be present, accessible and up to data. In the 2016 barometer the implementation score looks for the availability of 15 types of data.

1. Map data
2. Land ownership data
3. Detailed census data
4. Detailed government budget
5. Detailed government spend
6. Companies register
7. Legislation
8. Public transport timetables
9. International trade data
10. Health sector performance
11. Primary and/or secondary education performance data
12. Crime statistics
13. National environment statistics
14. National election results
15. Public contracts

Datasets such as election results, government spend and education performance are records of historical significance. Such datasets are very static in nature and can be made easily available in spreadsheet to download via a portal. At the same time it is these records that are linked more with access to information laws and have less economic potential for wide reuse.

Conversely datasets such as map, companies and trade data are much more dynamic and as such are more suited to a different level of service from a simple file download. This is especially true as map data takes the form of many complex formats and trade data can exceed sensible file sizes for access via download.

Agriculture is an area which cross cuts so many of these areas and as such many different datasets exist in portals. For example mapping data is critical to inform agriculture, beyond land ownership aspects such as water catchment and runoff, land use as well as protected/restricted areas can all help inform land use.

Many governments now run a data.gov.XX portal (or variants of in the local language, e.g. datos.gob.mx) that provide a central catalogue for government data. Many portals have dedicated agriculture sections, often closely tied with the relevant government department, that offer large amounts of links and access to downloadable data.

As mentioned previously, finding data in these portals relies heavily on metadata search. As such the title and description of each dataset is critical to enable discovery. However as the portal will often be organised by department or activity, finding a dataset requires either specialist domain knowledge and/or knowledge of how the local government organises and describes its data in the portal. A good starting point is to browse the portal for yourself to gain clues on how it is organised and discover how data is described before targeting your search using this new knowledge.

6. Obtaining data from 'on the web'

As mentioned previously much of the available open data out there is only available 'on the web'; either via a download button or contained within the web pages themselves. This section looks at the techniques to start discovering and unlocking this data ready to be used.

6.1. Finding downloadable data files

Many suppliers are helpful and give you human-readable download links in order to obtain their data. Most of the datasets on government data portals work in this way, explore the links below to discover more:

- UK National Statistics - Latest cereal stock statistics
- Tanzania Agricultural Census[13]

Many search engines, including Google, give you the ability to use advanced searches and prefixes in order to dig out data from sources such as the ones above. Advanced searches make use of filters and prefixes in order to limit the type of search and results. A list of prefixes and linked examples can be seen in Table 1.

[13] http://opendata.go.tz/dataset/tanzania-utafiti-wa-sampluli-sensa-ya-kilimo-2007-2008

Table 1 Prefixes for advanced search

Prefix	Description	Example search
filetype:	Search for specific file types only	filetype: xls cereal stocks
site:	Search with a specific domain or site only	site: opendata.go.tz agriculture
related:	Search for content related to a known page	related: https://www.gov.uk/government/statistics/cereal-stocks
link:	List only pages that link to the one given	link: https://www.gov.uk/government/statistics/cereal-stocks

Each of these can help refine your search for data. While the top two help narrow your research, the bottom two broaden it again once you have found a relevant resource. While 'related' can help find other relevant content, those who link to your dataset might have used the data and help provide context over the existing usage. Given that the majority of openly licensed dataset require attribution, the 'link:' specific search should return at least a number of results if the dataset has been used.

6.2. Data aggregators

One of the main challenges facing data 'on the web' is the lack of ability to search within the data itself. Existing search engines only enable metadata search, regardless of the format of the file (the same is true for audio and video, but google does allow you to search using an existing image).

Many data portals act as aggregators and allow some exploration of data, either from a single data provider or a set. In relation to agriculture the World Bank aggregates key statistical data from many countries and allows the exploration and download of this data.

One example from the World Bank is the Agriculture and Rural Development indicators[14]. Aggregating the data together allows the exploration of indicators such as agriculture land area vs rural population. The example linked to above looks at the comparison between a number of countries in east and west Africa. The DataBank service from the World Bank provides easy access to explore and then download the data ready for use.

[14] https://data.worldbank.org/topic/agriculture-and-rural-development?locations=KE-TZ-RW-GH-NG-ML-BF

Such aggregators can be a crucial source of open data when country portals are either out of date or simply not available.

Enigma.io, winner of Techcrunch Disrupt in 2013[15], brings together data from a multitude of open data sources and enables fine-grained search within the data itself. This is in effect a reverse search on data; rather than searching the metadata to find the data, enigma.io searches the data and shows you which datasets your search term is in. For example searching for 'Monsanto'[16] returns all sorts of interesting datasets, from Federal Campaign Contributions in 2016 to weather reports from 1949.

6.3. Scrapers

Sometimes data will not be available for download in a usable format. Sometimes the data will be available only from within the webpage as a table or list. In other cases the data may be available in a document format (such as PDF) rather than a data format. In both cases the use of data scrapers can help extract this visible data.

6.4. Web data scrapers

Web scrapers allow the automatic extraction of structured data from a web page. Tools like grepsr[17] allow the automatic extraction of data from structured websites in seconds, including the ability to handle pagination and infinite scrolling results. Currently such tools usually involve a per-record cost for extraction with a limited number of free credits per month.

6.5. PDF data scrapers

Another place where useful data is often embedded is within PDF reports produced by statistics agencies. These will often contain long appendices of tabular data which can be extracted using tools like PDFTables.com. Try it for yourself with some agricultural statistics[18] from Tanzania. The data starts on page 124 of this report and it is advisable to reduce the PDF to the exact pages that contain the data required before uploading to a PDF extractor like PDFTables.com.

7. Obtaining data from 'in the web'

[15] http://www.businessinsider.com/techcrunch-disrupt-winner-enigma-2013-5?IR=T
[16] https://public.enigma.com/search/Monsanto
[17] https://www.grepsr.com
[18] http://harvestchoice.org/sites/default/files/downloads/publications/Tanzania_2007-8_Vol_5g.pdf

The evolution of the web has led to the requirement to separate back-end infrastructure and data from the presentation layer such as websites and mobile applications that all use this same data. Shopping, weather and travel applications all offer various options for users to interact with essentially the same data.

These applications are all using dedicated data services to access and query the data. Many of these data services are documented for anyone to use while many remain hidden for various commercial or budgetary reasons. This section looks at the different techniques that can be tried to access data that is 'in the web' which helps service such applications.

7.1. Filetype extensions

Some websites have been built to offer a way to extract data by adding a file extension to the URL of the web page being viewed. For such websites, usually maintained by organisations who also publish downloadable open data, adding the correct extension will trigger a download of that page in a data format, as opposed to a document format.

A good example of this is the UK Government website (www.gov.uk), which provides any page in a data format simply by adding the relevant extension like '.json', for example www.gov.uk/browse/business.json. To view the data in a more human-readable form, copy it into jsonlint.io.

The UK Trade Tariff[19] also has the same functionality and contains details on the international trade codes that can be linked to the trade data available from Revenue and Customs[20].

Unfortunately not many websites make it clear to humans that alternative formats (such as JSON) are available. A good indicator is to find modern-looking websites where pages clearly contain data, such as records about individual companies[21], where such extensions can be tried. Table 2.1.2 lists common data formats available for data 'in the web'.

Table 2 Common formats for data 'in the web'

Extension	Description
.csv	Comma Separated Values. Tabular data format like excel but stripped back to just contain data in a simple structure.

[19] https://www.trade-tariff.service.gov.uk/trade-tariff/sections
[20] https://www.uktradeinfo.com/Statistics/BuildYourOwnTables/Pages/Home.aspx
[21] https://opencorporates.com/companies/ch/471356

.json	JavaScript Object Notation. A hierarchical data format native to the JavaScript language which is used widely on the web as it forms part of the HTML5 specification.
.xml	eXtensible Markup Language. A markup specification that has a wide range of uses. Has been criticised for its complexity and verbosity in comparison to JSON.
.rdf	Although RDF should not be a data format (not covered here). RDF defines a formal data structure which can be applied in xml, json and csv formats. Use of the extension implies that the structure is used and most commonly the data itself is in XML format.
.rss	Another specific XML structure that is often used for data feeds that regularly update such as news and weather.

7.2. Application programming interfaces (APIs)

APIs are one of the best ways to access data. APIs are a service best described as a 'promise' by one system to constantly and consistently provide a service to another that allows to the two to interact. For this reason APIs have many advantages over any other form of data access, as listed below.

1. *Service agreements.* As an API is a service, this guarantees access to data and can often be accompanied with service level agreements for those who wish to use them.

2. *Live access.* APIs provide a mechanism whereby data can be included live within an application. The most common example of a data API is live transport times. On the back of a single API, many hundreds of applications can be created.

3. *Designed for data.* Perhaps the biggest advantage of an API is how they are designed for data and machines rather than for humans. This means that data availability is no longer constrained by the paradigms of how humans use the web, however this does create challenges when searching data that might be within an API.

The major disadvantage of APIs is that data is not as easily accessible to download and use straight away. Some third-party applications, such as enigma.io already use APIs to access data from other services to allow easy access, while others like OpenCorporates allow downloads by file extension as part of their API.

Examples of services that have APIs include: OpenCorporates[22], OpenSteetMap[23], Twitter[24], Flickr[25], and LinkedIn[26]. These APIs provide direct access to the raw data as well as broad queries to allow faceted search.

Many of the open data platforms provide APIs to access the data including Socrata and OpenDataSoft. Such platforms are used by a number of governments and departments, Socrata is mainly in the US while OpenDataSoft throughout Europe. CKAN, an open source alternative, also has an API although this API only gives access to the metadata records in many instances.

Table 3 contains some examples of each platform and some of the available agricultural datasets, some of which were mentioned earlier.

Table 3 Examples of open data platforms and agricultural datasets available there

CKAN	Web page: https://data.gov.uk/dataset/cereal_stocks_england_and_wales API: https://data.gov.uk/api/3/action/package_show?id=cereal_stocks_england_and_wales
Socrata	Web Page: https://data.code4sa.org/dataset/List-of-Registered-Dams-2014/iety-gmha API: https://data.code4sa.org/resource/cig6-sz38.csv
OpenDataSoft	*No agricultural examples found.*

7.3. Using APIs

Many web APIs take the form of REST APIs. REpresentational State Transfer (REST) is an API designed specifically for the web. It has a specific set of guidelines and rules that control if something is a RESTful API.

Broadly speaking a REST API requires the use of resource identifiers which are then interacted with to upload/download the required resource. In the case of the Socrata example from the previous section the API location is the web-based identifier of the resource (https://data.code4sa.org/resource/cig6-sz38). Clicking this resource will redirect you to the web page, not because this is what click this link does, but because that is what was asked for when the

[22] https://api.opencorporates.com
[23] http://wiki.openstreetmap.org/wiki/API
[24] https://developer.twitter.com/en/docs
[25] https://www.flickr.com/services/api/
[26] https://developer.linkedin.com/docs/rest-api

linked was clicked in a web browser. The web browser goes to a GET request for a web page (text/html) representation of this resource as shown below.

GET /resource/cig6-sz38 HTTP/1.1
HOST: data.code4sa.org
ACCEPT: text/html

A REST API specifies that a machine should be able to change the request in order to ask for different representations of the same resource. This is a bit like adding file extensions, except where the requested resource does not change any part of its location on the web (as adding '.csv' effectively changes the URL). The example below shows two example requests for a JSON and CSV version of the same resource using a REST API.

JSON	CSV
GET /resource/cig6-sz38 HTTP/1.1 HOST: data.code4sa.org **ACCEPT: text/csv**	GET /resource/cig6-sz38 HTTP/1.1 HOST: data.code4sa.org **ACCEPT: application/json**

REST APIs are simply extensions of the webs existing HyperText Transfer Protocol except used for data. Thus it is possible to change the type of request from a GET to a PUT and then send structured data to the server to replace the existing data with new data (using authentication obviously). The City of Chicago use the POST method to send updated crime statistics[27] to their data portal and have been doing so daily since 2001.

APIs not only allow access for users to data, they also form a key part of the provider's data infrastructure allowing data to be managed and kept up to date.

7.4. Hidden APIs

Not all websites that dynamically load data make their API known public, even if one exists. However it is possible to discover them. Doing so requires a fair amount of technical knowledge; however a good Google search often turns up communities of people who may have already built something for the particular service you want to extract data from.

As many APIs are based upon the REST API design, in many cases it is fairly straightforward for someone familiar with REST APIs to quickly find if a service

[27] https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2/data

has one and how it works. This can be done by trialling out some REST requests with browser extensions like Postman for Google Chrome[28].

The ODI's experimental Hidden Data Extractor[29] tool has been built to automatically look for REST APIs that exchange JSON data when a web page is loaded.

2.1.7 Checking Your Rights to Use Data

There are many ways to obtain data from the web, be it clearly visible via a download button or available through a public or hidden API. Regardless of the method of acquisition of data it is critical to check your rights both to use that method and to use the subsequent data.

Just like the data itself, some rights statements will be only human-readable, some only machine-readable and some a combination of both. Commonly, however, service providers will have a human-readable version of their terms of use and/or data licence that will cover both the terms of use of the service and rights to use data once acquired.

Many government data portals will have the data licence listed as a piece of metadata against the record being viewed. For example, in data.gov.uk all licences are listed directly under the title of the dataset as a clickable link. The CKAN platform (which data.gov.uk is a version of) is particularly good at exposing rights statements. This helps users ensure the data they are viewing is open data.

Services like Flickr also have licences against each photo. Each flickr user is able to specify licenses for their own photos. Flickr even provides a search that allows others to find photos with specific licenses.

If licences are also machine-readable (as is the case with CKAN and Flickr) then search engines can use this as a piece of metadata, meaning that search results can be instantly filtered to contain only openly licensed content (try out google advanced search[30]).

If using a REST API, the rights statement might be returned as part of a Link header[31]. This separates the rights statement from the content, allowing the response to still be the pure data, e.g. CSV file.

If none of these options exist, then it might be necessary to read the terms and conditions of the providers to ensure that your method of access and rights to

---

[28] https://chrome.google.com/webstore/detail/postman/fhbjgbiflinjbdggehcddcbncdddomop?hl=en

[29] http://odinprac.theodi.org/hidden-data-extractor/

[30] https://www.google.co.uk/advanced_search

[31] https://theodi.org/guides/publishers-guide-to-the-open-data-rights-statement-vocabulary#linkingtorightsstatementsfromwebapis

use the data are permitted. Just because something is accessible on the web does not give everyone the right to use it.

Summary

This lesson has introduced many of the methods and shortcomings of discovering open data on the web. It is still early in the evolution of a 'data age' following the 'information age' and services that specialise in providing fast access to data are evolving.

At the same time the number of services providing data is also growing, mirroring the early days of the web. There are still lessons to be learnt, however methods to access data are beginning to stabilise with the emergence of common APIs such as REST.

Data formats have also evolved and thus so too have methods to discover and access data. Search engines are becoming much more intelligent and can be customised to perform highly targeted queries. At the same time tools to help extract and work with data have evolved such that it is very easy to start working with data regardless of the format.

The evolution of mobile applications that demand instant access to data has also increased the number of APIs available, even if some of them remain hidden.

It is clear that we live in the age of data, however we need to be careful over our rights to use such data. Having clear open data licenses is critical to the future of our data infrastructure.