Open Data and Open Source

UBCO Master of Data Science – DATA 541

Today's Class

Open initiatives

Open Data

- Open data in Canada
- Open data in BC
- Open data in Kelowna

Application Program Interface

Open source

Open source software and projects

What Makes a Work “Open”?

Open Work means:

- Any work product (raw data, creative works, software, etc.)
- Made available under a specific license
- That allows access, use and modification of the work by anyone

Open Works share most/all of the following characteristics:

- Free
- Redistributable
- Modifiable
- Formatted for easy use and modification

“Open” Initiatives

There are a wide range of Open Activities and Initiatives:

- Open Data: the Open licensing of data, allowing others to freely use and modify the works
- Open Source: the Open licensing of software and its source code
- Open Access: unrestricted online access to published research
- Open Content: unrestricted access to use and modify creative works
- Open Research: openness in the tools and methodologies for conducting scientific research
- Open Government: openness in access to government documents and proceedings

What is Open Data?

Open Data is the movement to make data freely available to all with no restrictions on use or copyright.

- Availability and Access
  - The data must be available and should be downloadable via the Internet without charge.
- Re-use and Redistribution
  - The data must be provided under terms that permit reuse and redistribution
- Universal Participation
  - Everyone must be able to use, re-use and redistribute. There should be no discrimination against fields of endeavor or persons or groups

What is Open Data?

Governments have been major supporters and providers of open data as data collected by governments is primarily done to benefit its citizens.

- e.g., How many permanent resident visa applications were received in each province?

Corporations and other organizations are both producers and consumers of open data.

Open Data in Canada

Federal, provincial, and local governments have all been involved in the open data movement.

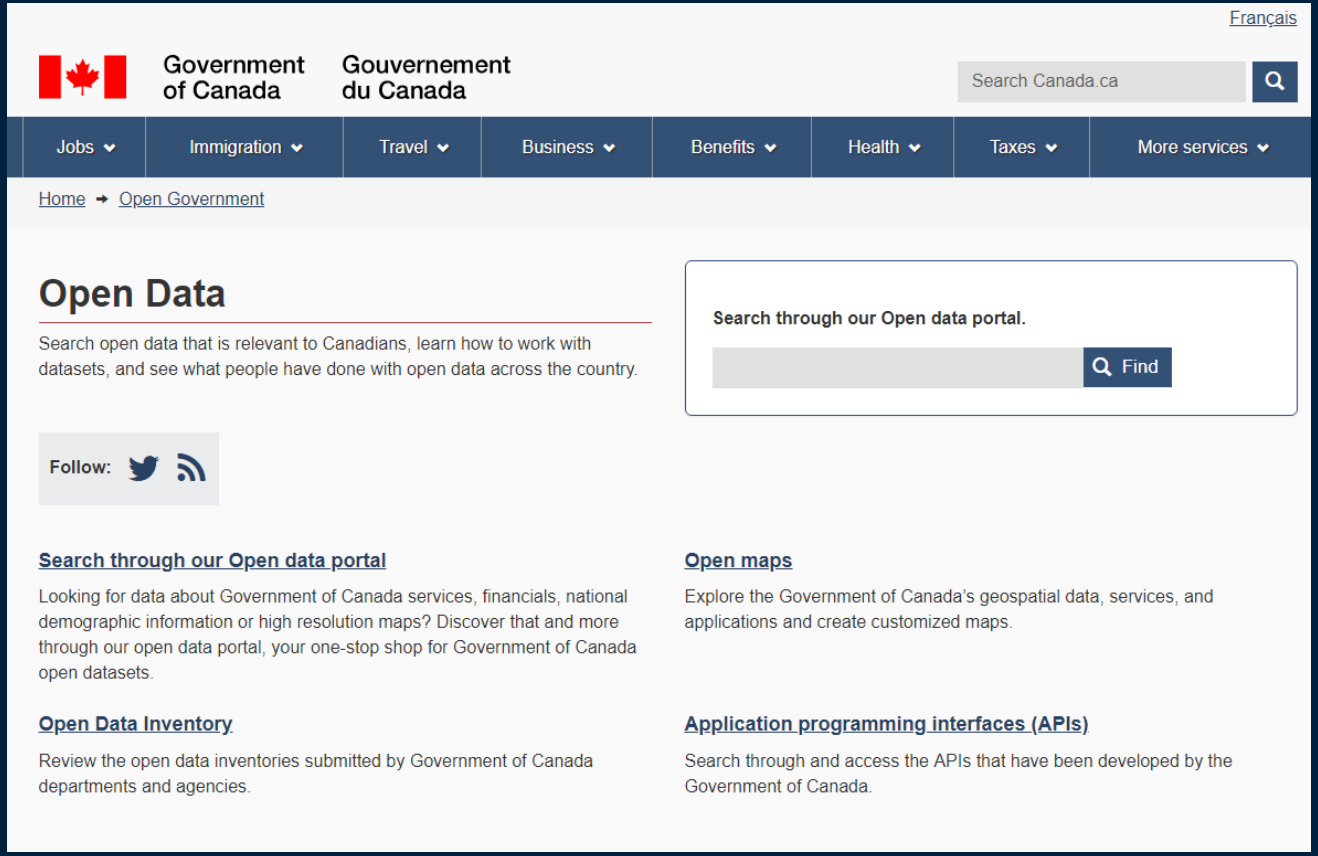
Canadian Federal government: http://open.canada.ca/en

- How to use: http://open.canada.ca/en/working-data
- Statistics Canada: http://www.statcan.gc.ca/eng/rdc/data

British Columbia government: http://www.data.gov.bc.ca/

City of Kelowna:
https://www.kelowna.ca/city-services/city-maps-open-data/open-data-catalogue

Open Data in Canada



Data Format

Structured Data: usually in the form of text but follows a specific format giving it machine readability

- CSV: Comma-Separated Values (CSV)
- JSON: JavaScript Object Notation
- XML: eXtensible Markup Language

CSV

```
id,name,age,gender
1,luis,21,m
2,dave,23,m
```

JSON

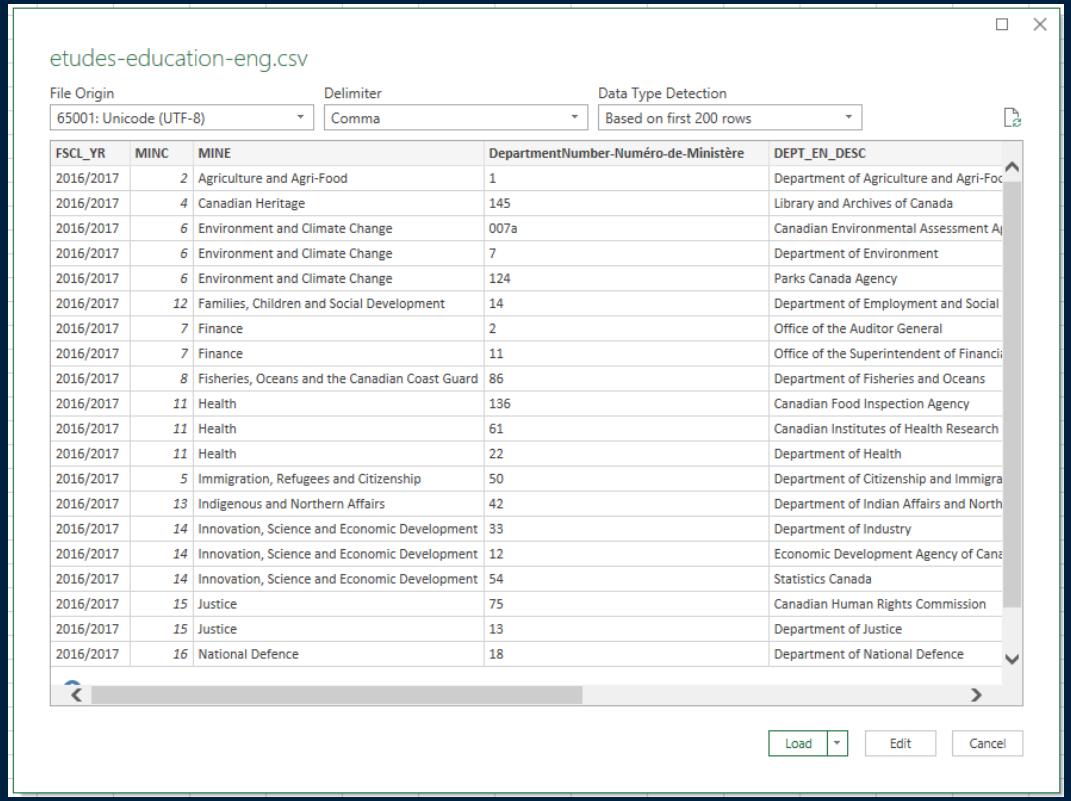
```
{
"person": [
        {
        "id": "1",
        "name": "luis",
        "age": "21",
        "genre": "m"
        }
        ]
}
```

XML

```
<?xml version="1.0"?>
<people>
 <person>
 <id>1</id>
 <name>luis</name>
 <age>21</age>
 <gender>m</gender>
 </person>
</people>
```

Open Data in Canada

How to use: http://open.canada.ca/en/working-data



etudes-education-eng.csv

File Origin: 65001: Unicode (UTF-8) | Delimiter: Comma | Data Type Detection: Based on first 200 rows

FSCL_YR	MINC	MINE	DepartmentNumber-Numéro-de-Ministère	DEPT_EN_DESC
2016/2017	2	Agriculture and Agri-Food	1	Department of Agriculture and Agri-Foc
2016/2017	4	Canadian Heritage	145	Library and Archives of Canada
2016/2017	6	Environment and Climate Change	007a	Canadian Environmental Assessment Ag
2016/2017	6	Environment and Climate Change	7	Department of Environment
2016/2017	6	Environment and Climate Change	124	Parks Canada Agency
2016/2017	12	Families, Children and Social Development	14	Department of Employment and Social
2016/2017	7	Finance	2	Office of the Auditor General
2016/2017	7	Finance	11	Office of the Superintendent of Financi
2016/2017	8	Fisheries, Oceans and the Canadian Coast Guard	86	Department of Fisheries and Oceans
2016/2017	11	Health	136	Canadian Food Inspection Agency
2016/2017	11	Health	61	Canadian Institutes of Health Research
2016/2017	11	Health	22	Department of Health
2016/2017	5	Immigration, Refugees and Citizenship	50	Department of Citizenship and Immigra
2016/2017	13	Indigenous and Northern Affairs	42	Department of Indian Affairs and North
2016/2017	14	Innovation, Science and Economic Development	33	Department of Industry
2016/2017	14	Innovation, Science and Economic Development	12	Economic Development Agency of Cana
2016/2017	14	Innovation, Science and Economic Development	54	Statistics Canada
2016/2017	15	Justice	75	Canadian Human Rights Commission
2016/2017	15	Justice	13	Department of Justice
2016/2017	16	National Defence	18	Department of National Defence

Load | Edit | Cancel

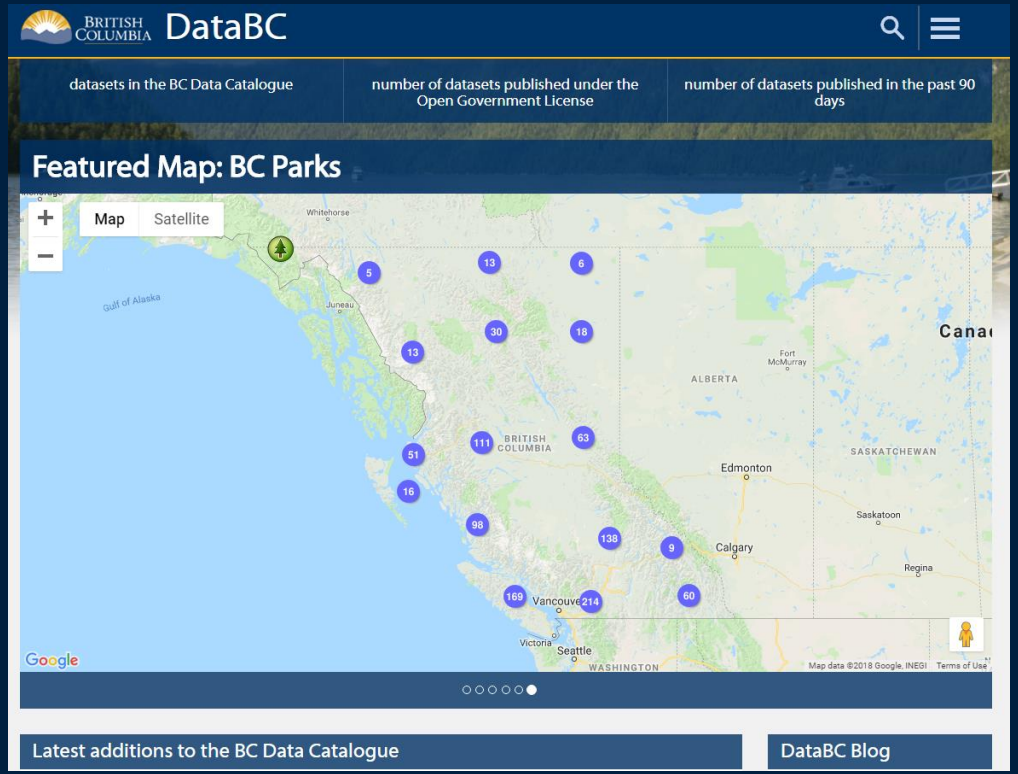
Application Programming Interfaces

In the context of the Government of Canada's Data Portal:

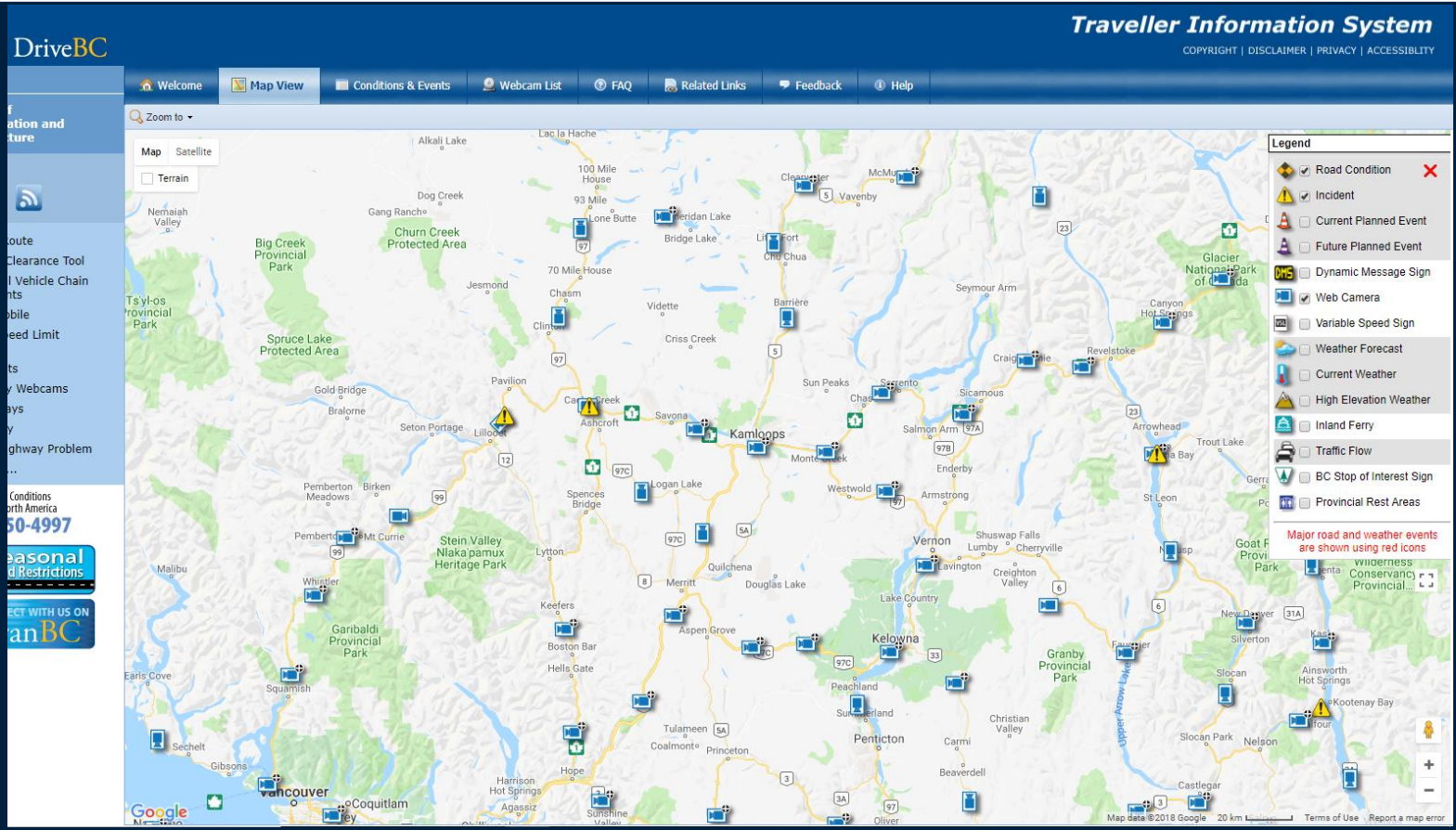
- Internet-connected software interfaces that provide access to open data.
- Commonly these interfaces will use the HTTP protocol.
- APIs provide on-demand access to large, timely or complex data allowing developers to mash data from multiple sources and create new views on information through applications or visualizations.

Open Data in BC

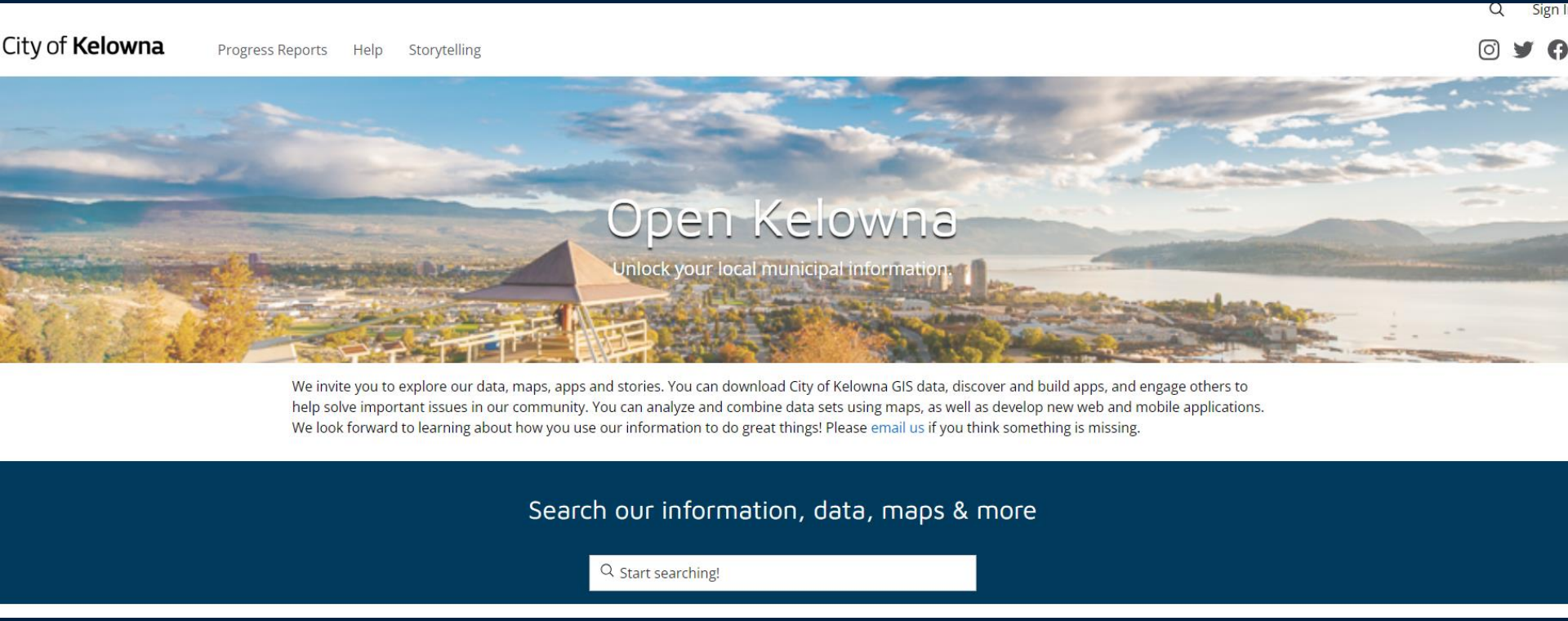
http://www.data.gov.bc.ca/



DriveBC


DriveBC
Traveller Information System
COPYRIGHT | DISCLAIMER | PRIVACY | ACCESSIBLITY
Welcome
Map View
Conditions & Events
Webcam List
FAQ
Related Links
Feedback
Help
Zoom to
Map
Satellite
Terrain
Legend
Road Condition
Incident
Current Planned Event
Future Planned Event
Dynamic Message Sign
Web Camera
Variable Speed Sign
Weather Forecast
Current Weather
High Elevation Weather
Inland Ferry
Traffic Flow
BC Stop of Interest Sign
Provincial Rest Areas
Major road and weather events are shown using red icons
Kamloops
Kelowna
Vernon
Penticton
Vancouver
Coquitlam
Google

Open Data in Kelowna



Open Data in the United States

United States government: https://www.data.gov/

Individual states have their own open data sites as well.

- Example: Washington state: https://data.wa.gov/

United States: Data.gov

Open Data Worldwide

The World Bank: http://data.worldbank.org/

United Nations: http://data.un.org/

Unicef: https://data.unicef.org/

Open Data Aggregators

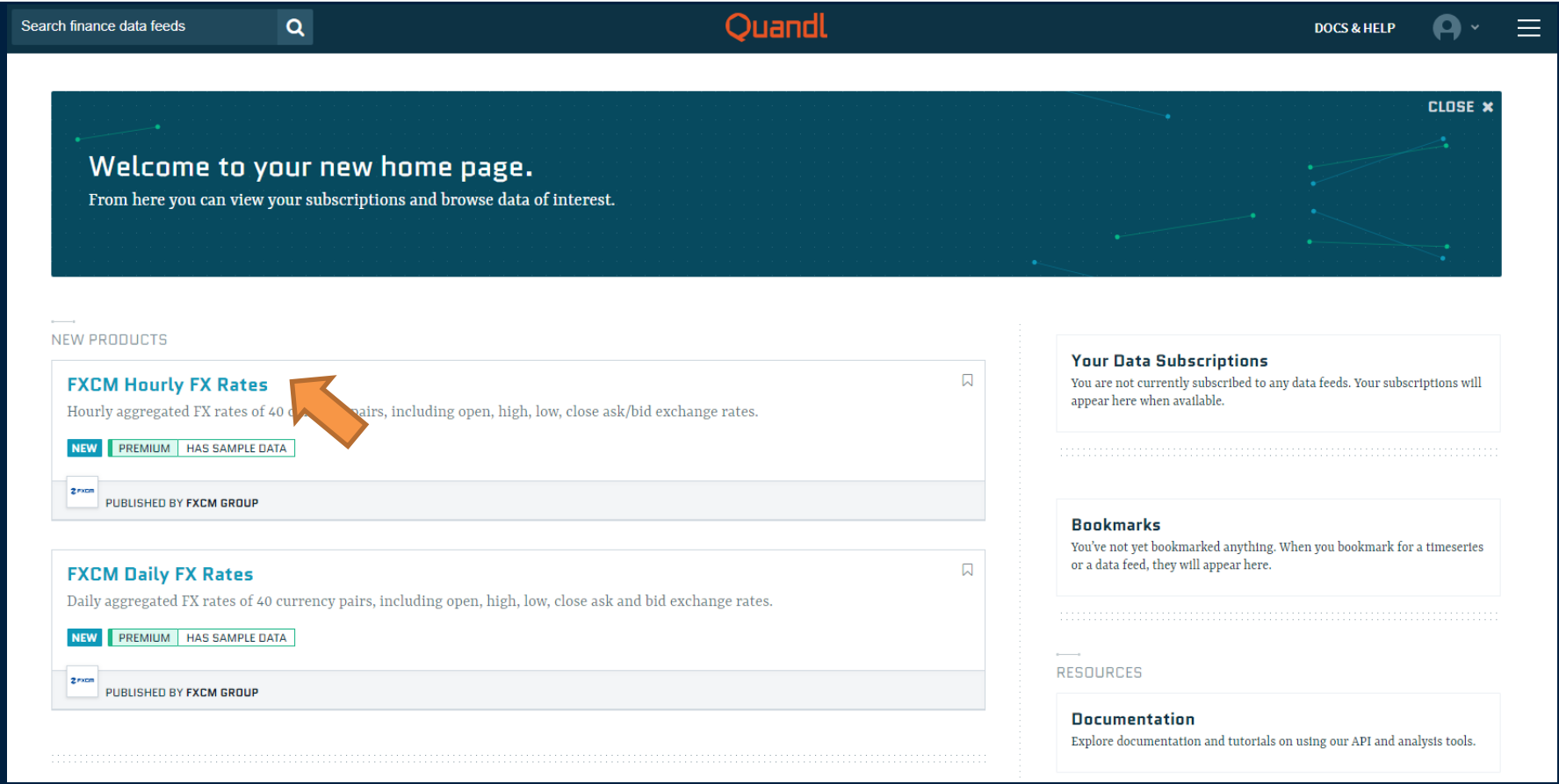
There are many sites that aggregate open data sets (and some data sets for a cost).

A Canadian based site is Quandl (http://www.quandl.com).

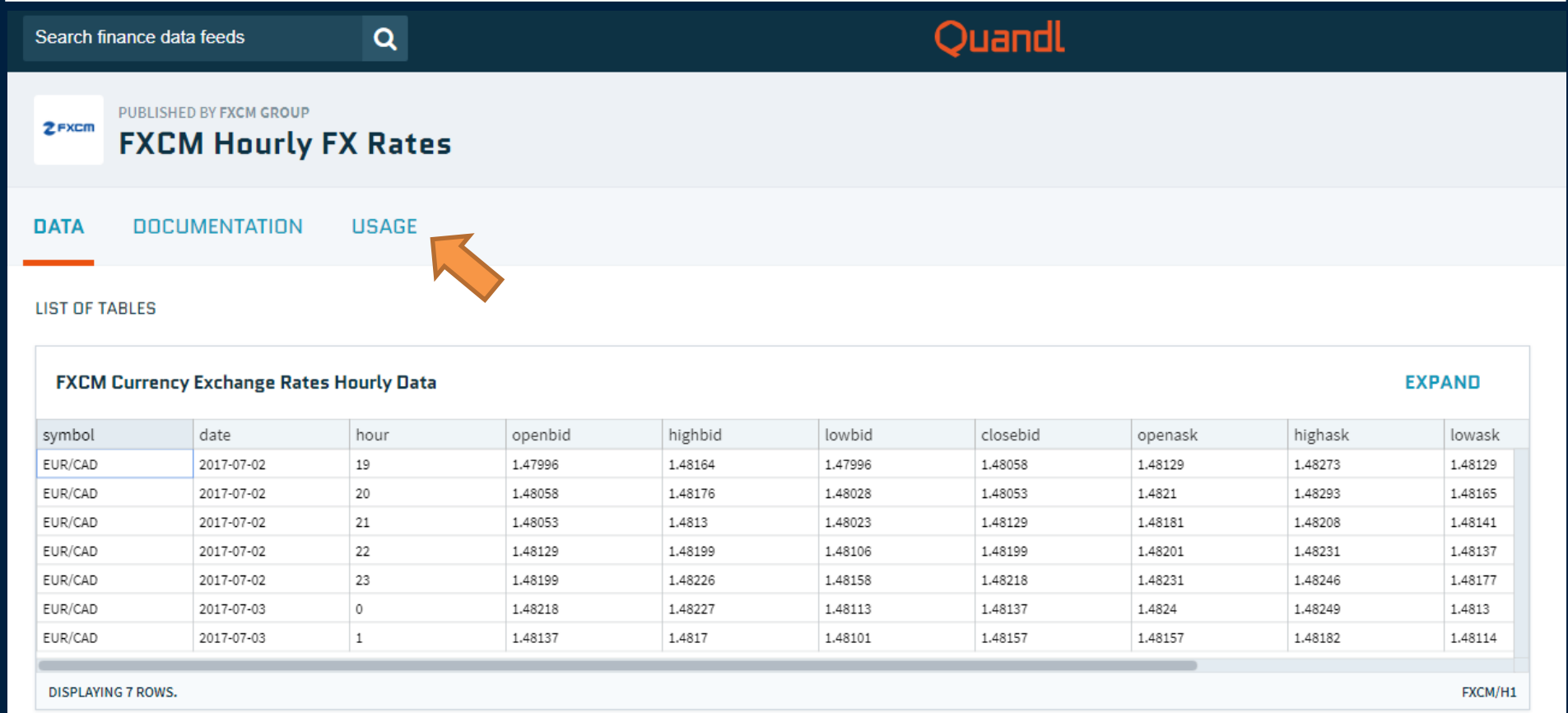
Currently, NASDAQ Data Link (https://data.nasdaq.com/)

Kaggle provides many data sets and competitions and techniques for data analytics and machine learning.
https://www.kaggle.com/datasets

Quandl

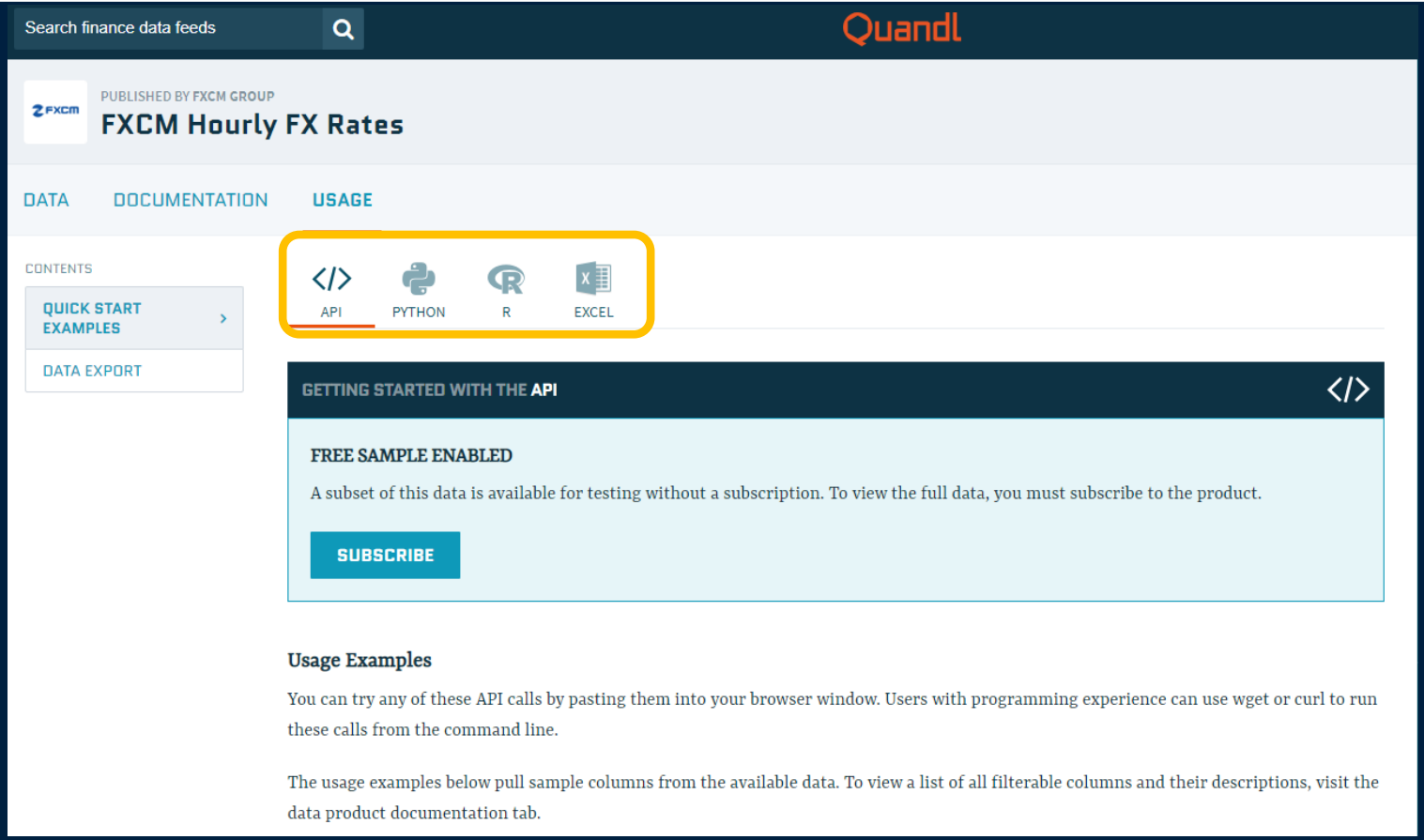

Search finance data feeds
Quandl
DOCS & HELP
CLOSE
Welcome to your new home page.
From here you can view your subscriptions and browse data of interest.
NEW PRODUCTS
FXCM Hourly FX Rates
Hourly aggregated FX rates of 40 c pairs, including open, high, low, close ask/bid exchange rates.
NEW PREMIUM HAS SAMPLE DATA
PUBLISHED BY FXCM GROUP
FXCM Daily FX Rates
Daily aggregated FX rates of 40 currency pairs, including open, high, low, close ask and bid exchange rates.
NEW PREMIUM HAS SAMPLE DATA
PUBLISHED BY FXCM GROUP
Your Data Subscriptions
You are not currently subscribed to any data feeds. Your subscriptions will appear here when available.
Bookmarks
You've not yet bookmarked anything. When you bookmark for a timeseries or a data feed, they will appear here.
RESOURCES
Documentation
Explore documentation and tutorials on using our API and analysis tools.

Quandl



symbol	date	hour	openbid	highbid	lowbid	closebid	openask	highask	lowask
EUR/CAD	2017-07-02	19	1.47996	1.48164	1.47996	1.48058	1.48129	1.48273	1.48129
EUR/CAD	2017-07-02	20	1.48058	1.48176	1.48028	1.48053	1.4821	1.48293	1.48165
EUR/CAD	2017-07-02	21	1.48053	1.4813	1.48023	1.48129	1.48181	1.48208	1.48141
EUR/CAD	2017-07-02	22	1.48129	1.48199	1.48106	1.48199	1.48201	1.48231	1.48137
EUR/CAD	2017-07-02	23	1.48199	1.48226	1.48158	1.48218	1.48231	1.48246	1.48177
EUR/CAD	2017-07-03	0	1.48218	1.48227	1.48113	1.48137	1.4824	1.48249	1.4813
EUR/CAD	2017-07-03	1	1.48137	1.4817	1.48101	1.48157	1.48157	1.48182	1.48114

Quandl


Search finance data feeds
Quandl
PUBLISHED BY FXCM GROUP
FXCM Hourly FX Rates
DATA
DOCUMENTATION
USAGE
CONTENTS
QUICK START EXAMPLES
DATA EXPORT
API
PYTHON
R
EXCEL
GETTING STARTED WITH THE API
FREE SAMPLE ENABLED
A subset of this data is available for testing without a subscription. To view the full data, you must subscribe to the product.
SUBSCRIBE
Usage Examples
You can try any of these API calls by pasting them into your browser window. Users with programming experience can use wget or curl to run these calls from the command line.
The usage examples below pull sample columns from the available data. To view a list of all filterable columns and their descriptions, visit the data product documentation tab.

Quandl

Filter by a single date and symbol

```
https://www.quandl.com/api/v3/datatables/FXCM/H1?date=2017-07-02&symbol=EUR%2FCAD&api_key=qCGcR1znnhhJhuqGXQg3
```

Filter by a single symbol and multiple dates

```
https://www.quandl.com/api/v3/datatables/FXCM/H1?symbol=EUR%2FCAD&date=2017-07-02%2C2017-07-03%2C2017-07-04&api_key=qCGcR1znnhhJhuqGXQg3
```


```
{"datatable":{"data":[["EUR/CAD","2017-07-02",19,1.47996,1.48164,1.47996,1.48058,1.48129,1.48273,1.48129,1.4821,212],["EUR/CAD","2017-07-02",20,1.48058,1.48176,1.48028,1.48053,1.4821,1.48293,1.48165,1.48181,298],["EUR/CAD","2017-07-02",21,1.48053,1.4813,1.48023,1.48129,1.48181,1.48208,1.48141,1.48201,595],["EUR/CAD","2017-07-02",22,1.48129,1.48199,1.48106,1.48199,1.48201,1.48231,1.48137,1.48231,1910],["EUR/CAD","2017-07-02",23,1.48199,1.48226,1.48158,1.48218,1.48231,1.48246,1.48177,1.4824,2307]],"columns":[{"name":"symbol","type":"String"},{"name":"date","type":"Date"},{"name":"hour","type":"Integer"},{"name":"openbid","type":"double"},{"name":"highbid","type":"double"},{"name":"lowbid","type":"double"},{"name":"closebid","type":"double"},{"name":"openask","type":"double"},{"name":"highask","type":"double"},{"name":"lowask","type":"double"},{"name":"closeask","type":"double"},{"name":"totalticks","type":"Integer"}]},"meta":{"next_cursor_id":null}}
```

Open Data from Companies

Many companies either have public data or application programming interfaces (APIs) that allow people to use their data.

- Google: https://www.google.com/publicdata/directory (public data explorer) and https://developers.google.com/maps/ (Google Maps API)
- Facebook: https://developers.facebook.com/ (API)
- reddit: https://www.reddit.com/dev/api (API)
- Twitter: https://dev.twitter.com/rest/public (API)
- Amazon: https://aws.amazon.com/public-data-sets/ (public data sets) and https://developer.amazon.com/ (API for developers)
- Best Buy: https://developer.bestbuy.com/ (API)

Try it: Open Data

Explore the federal, provincial, and City of Kelowna data sets to discover "something interesting". Report to your neighbors and to the class.

From any Canadian government open data site, retrieve a data set and analyze and visualize it using one of our tools: Excel, R, Python, Tableau.

Open Data for Researchers

Increasingly publicly funded researchers are responsible for making their data sets, procedures, and results available to the public (and other researchers).

- Canadian researchers funded by NSERC, SSHRC, CIHR must make their publications freely available within 12 months of publication.
- Researchers in bioinformatics and other fields must make their data sets publically available in a database or repository.

Researchers benefit by having access to public data sets and data sets of other researchers, but there is also a challenge as producing data sets (and perhaps commercializing results) may restrict open access.

Open Data in Computer Science

Computer scientists in various fields create standardized data sets for experimentation and research.

- Machine learning/data mining: UCI ML repository http://archive.ics.uci.edu/ml/

Open Source

Why is it called Open Source

- Open: collaboration is open to all
- Source: source code is freely shared

Open Source Software (OSS)

*"Users have the freedom to run an OSS program for any purpose, have **access to its source code**, may **modify the code**, and may **redistribute copies** of either original or the modified code **without having to pay royalties** to previous developers"*.

Birman, A. and Ritsko, J. "Preface to the Special Issue on Open Systems," *IBM Systems Journal,* Vol. 44, 2, 2005. p. 213.

Example of Open Source Software

Linux

- Free Unix-type Operating System
- Created by Linus Torvalds (1991)
- Free source code and distribution concept
- Many companies released their own version of OS based on LINUX i.e. Red Hat
- More stable and secure, cheaper, and less resources needed

Example of Open Source Software

Mozilla Firefox

- Popular web browser
- Free source code
- Runs on various OS platforms
- Google-Integrated search system
- Compared to IE: faster
  - more secure, customizable

How Open Source Projects Work

There are three major roles in an Open Source Project:

End-Users use the software project

Contributors submit specific change requests to the project code and/or documentation

Committers are the gatekeepers of changes to the project and responsible for planning, reviewing and coordinating changes

How Open Source Projects Work

Open Source projects still retain a copyright holding entity, typically the original author(s) or a Software Foundation.

Most projects that accept contributions require participants to sign a Contributor License Agreement (CLA), which officially transfers the license to use contributions in the project without restriction

How Open Source Projects Work

Open Source Foundations are the copyright holders of many significant Open Source projects.

Foundations can provide **legal and operational infrastructure** and are often set up to **facilitate project donations**.

Projects must answer to the Foundations' Board of Directors and follow the Foundation's by-laws

Most Foundations are technology-specific, like **Linux Foundation**, **Python Foundation**, and **Apache Foundation**

Revision Control

Revision Control System (RCS) is essential for collaborative development and maintenance of all kinds of electronic documents

An RCS provides tools for:

- **tracking** and **reviewing** changes
- **merging** changes made by multiple people.

Common Revision Control Systems include **GIT**, **Subversion**, **Mercurial**, **CVS**, and **Bitkeeper**

Many hosted RCS options are available, some of which are free for Open Source projects (e.g., **GitHub**, **BitBucket**, **GitLab**, and **CodePlex)**

Revision Control

Software development projects require a system for collecting and tracking reports of issues and change requests and assigning them to contributors; this is generally managed through purpose-built issue tracking software

Open Source projects also depend on a number of collaboration tools in order to accept contributions, including:

- Pull requests (a method for reviewing software code changes)
- Mailing lists, forums, and/or group work spaces

Proprietary Software

Proprietary software is any computer software with restrictions on use or private modification, or with restrictions judged to be excessive on copying or publishing of modified or unmodified versions.

Examples:
Microsoft Windows, Adobe Flash Player, iTunes, Adobe Photoshop, Google Earth, Mac OS X, Skype, WinZip.

OSS vs. Proprietary Software

Customizing

Open source software is licensed under a free software license. This enables you to fine-tune your software to your organization's specific needs, giving you a much more tailored solution.

Proprietary software is protected by very strict copyright and licensing agreements which greatly affects what you can do with the software and how it can be used.

OSS vs. Proprietary Software

Technology Support

Open Source Software: Although there isn't a formal support department, open source software has plenty of community-based support options.

Proprietary Software: If you want real tech support beyond the typical FAQ type questions, you'll have to pay a premium either in the form of costly support licenses or pay-per-call fees.

OSS vs. Proprietary Software

Security

Open Source Software**:** Open source offers transparency and reliability of its source code. It has powerful potential security advantages by preventing spyware and promoting encryption.

Proprietary Software: Some proprietary software offers potentially robust security.

Objectives

- Define open data and explain the motivations for making data "open".
- List some of the governments and organizations that provide data in an open fashion.
- Use open data sets when applicable when performing data analysis.
- Define Open Source Software and explain how open source projects works
- Compare and contrast open source data with Proprietary software.


UBC
THE UNIVERSITY OF BRITISH COLUMBIA