Open Data

Lecture 3

TU Wien, 193.067 Free and Open Technologies (WS 2019/2020)
Christoph Derndorfer and Lukas F. Lang

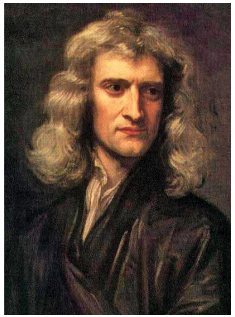
Lecture outline

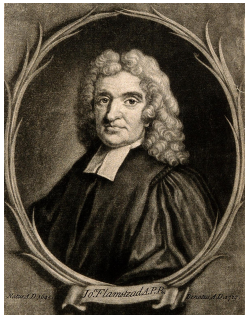
1. ~~October 8, 2019: FLOSS (Free/Libre and Open Source Software)~~
2. ~~October 15, 2019: Open Hardware~~
3. **October 22, 2019: Open Data**
4. October 29, 2019: Open Content/Open Educational Resources
5. November 5, 2019: Open Science/Research
6. November 12, 2019: Open Access
7. November 19, 2019: Open Spaces/Open Practices at Metalab Vienna
   - Location: Metalab, **Rathausstraße 6, 1010 Vienna**
8. November 26, 2019: Guest Lecture: Stefanie Wuschitz (Mz* Baltazar's Lab)

Motivating examples

Isaac Newton vs. John Flamsteed

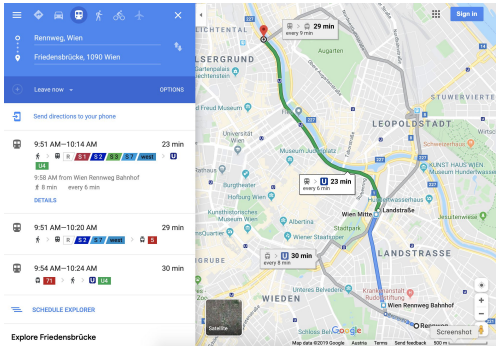
- Feisty dispute over astronomical data around 1700 [1]

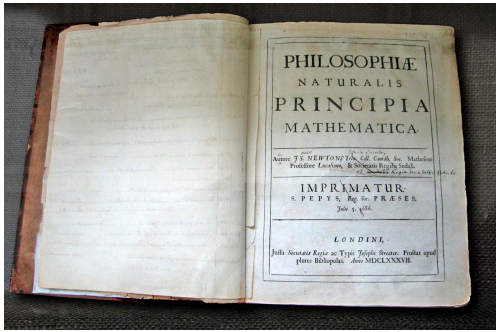




Wiener Linien vs. Google Maps

- After years, the City of Vienna eventually agreed to publish public transport data under Creative Commons license and in de-facto standard format [5].





"Philosophiae Naturalis Principia Mathematica" at Wren Library, Trinity College Cambridge, UK

[1] Freistetter, F. (2018), "Isaac Newton, The Asshole Who Reinvented the Universe". Amherst, NY: Prometheus.
[2] Andrew Dunn, CC BY-SA 2.0
[3] John Flamsteed, Godfrey Kneller, CC BY 4.0
[4] Isaac, Godfrey Kneller, public domain
[5] Der Standard (2017/08/21), "Wiener Linien: Daten ab sofort in Google Maps integriert", accessed 2019/10/18

Open data sources

Government portals:

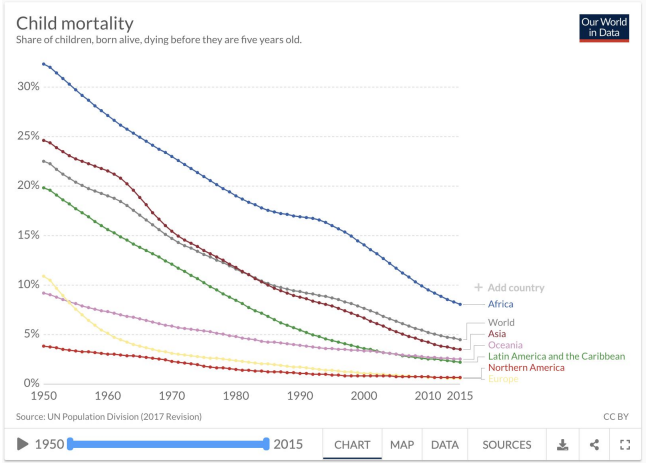
- Europeandataportal.eu: more than 900,000 data sets
- Data.gov (USA): more than 250,000 data sets
- Data.gov.uk: more than 52,000 data sets
- Data.gv.at: more than 26,000 data sets
- World Bank Open Data: more than 20,000 data sets
- Digitales.wien.gv.at: 490 data sets

Scientific repositories:

- Zenodo.org: more than 1,373,504 data sets (code, images, raw data, etc.)
- Clinicaltrials.gov (USA): more than 300,000 studies (40,000 with results)

Application examples

- Public transportation apps:
  - Wegfinder, Öffi, etc. use open transportation data for routing applications
- Real estate:
  - Zoomsquare (ceased in 2018) used open data by the City of Vienna to provide additional information for flat listings (e.g. schools, doctors, carsharing stations, etc.)
  - Bodenpreise.at uses basemap.at data to display land prices on a map
- Tourism:
  - Bergfex.at uses basemap.at data for skiing and hiking information
- Science:
  - Ourworldindata.org aggregates and visualizes open data from publications, WHO, World Bank, etc.

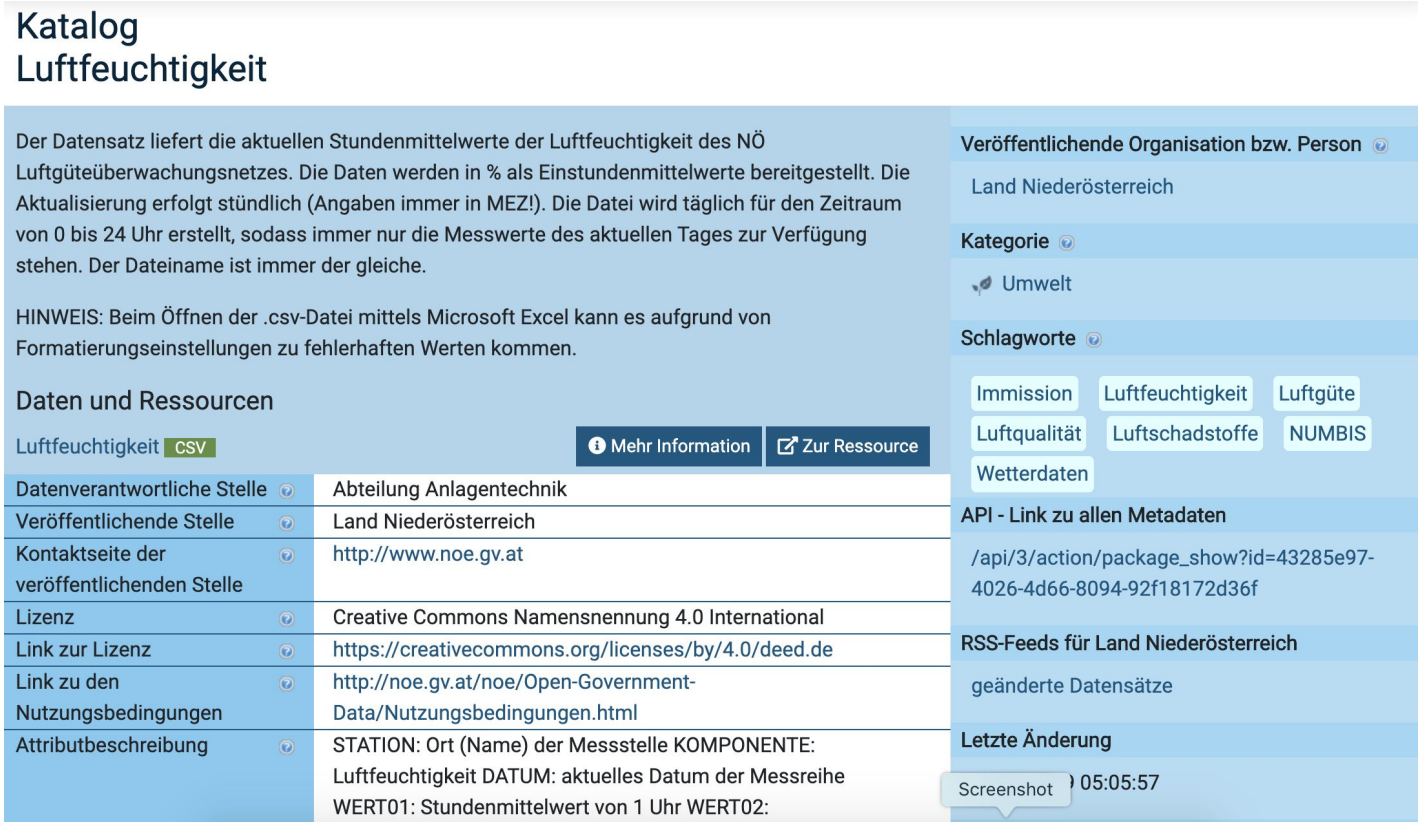


[1] Our World in Data, The Our World in Data-Grapher, CC BY 4.0

Data examples

Categories of data.gv.at:

Data examples



Katalog
Luftfeuchtigkeit

Der Datensatz liefert die aktuellen Stundenmittelwerte der Luftfeuchtigkeit des NÖ Luftgüteüberwachungsnetzes. Die Daten werden in % als Einstundenmittelwerte bereitgestellt. Die Aktualisierung erfolgt stündlich (Angaben immer in MEZ!). Die Datei wird täglich für den Zeitraum von 0 bis 24 Uhr erstellt, sodass immer nur die Messwerte des aktuellen Tages zur Verfügung stehen. Der Dateiname ist immer der gleiche.

HINWEIS: Beim Öffnen der .csv-Datei mittels Microsoft Excel kann es aufgrund von Formatierungseinstellungen zu fehlerhaften Werten kommen.

Daten und Ressourcen

Luftfeuchtigkeit CSV

Mehr Information | Zur Ressource

Datenverantwortliche Stelle	Abteilung Anlagentechnik
Veröffentlichende Stelle	Land Niederösterreich
Kontaktseite der veröffentlichenden Stelle	http://www.noe.gv.at
Lizenz	Creative Commons Namensnennung 4.0 International
Link zur Lizenz	https://creativecommons.org/licenses/by/4.0/deed.de
Link zu den Nutzungsbedingungen	http://noe.gv.at/noe/Open-Government-Data/Nutzungsbedingungen.html
Attributbeschreibung	STATION: Ort (Name) der Messstelle KOMPONENTE: Luftfeuchtigkeit DATUM: aktuelles Datum der Messreihe WERT01: Stundenmittelwert von 1 Uhr WERT02:

Veröffentlichende Organisation bzw. Person

Land Niederösterreich

Kategorie

Umwelt

Schlagworte

Immission | Luftfeuchtigkeit | Luftgüte | Luftqualität | Luftschadstoffe | NUMBIS | Wetterdaten

API - Link zu allen Metadaten

/api/3/action/package_show?id=43285e97-4026-4d66-8094-92f18172d36f

RSS-Feeds für Land Niederösterreich

geänderte Datensätze

Letzte Änderung

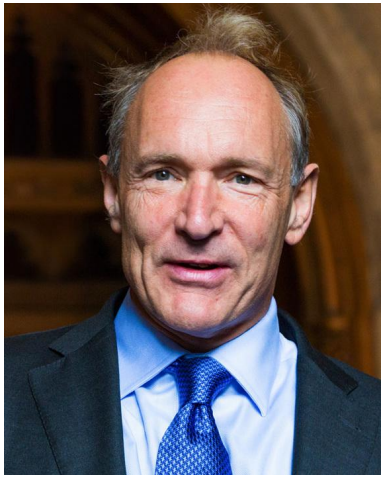
05:05:57

Screenshot

Why do we need open data?

“Greater openness, accountability and transparency in government will give people greater choice and make it easier for individuals to get more directly involved in issues that matter to them.” [1]

Sir Tim Berners-Lee



[1] BBC News (2010/04/01), “Ordnance Survey offers free data access”, accessed 2019/10/17
[2] Image by KOKUYO, CC BY-SA 4.0

Open data and community

Steven Levy's famous book "Hackers" [1] lists ethics and principles:

- Access to computers should be unlimited and total
- All information should be free
- ...

The Chaos Computer Club recently added [2]:

> **"Öffentliche Daten nützen,**
> **private Daten schützen."**
>
> ["use common/public data, protect private data"]

"open data" on media.ccc.de

[1] Levy, S. (1984), "Hackers: Heroes of the Computer Revolution". New York, NY: Penguin Books.
[2] Chaos Computer Club, "Hackerethik", accessed 2019/10/15

Goals

Open data movement demands that **all non-personal data** “produced” in politics, public administration, and science must to be made available and accessible at no cost in a timely manner, in open and machine-readable formats, and without posing any restrictions on its use.

Examples:

- Environmental data, education data
- Infrastructure data, geospatial data, public transportation data
- Scientific studies, publications
- Statistical data, census, crime statistics, law, etc.

Why? → funded by public money and should be public goods

Important clarification

EU General Data Protection Regulation (GDPR) [1]:

- GDPR concerns exclusively **personal data**
- GDPR and open data are **not mutually exclusive**
- Exceptions:
  - Legitimate reasons
  - Anonymized data

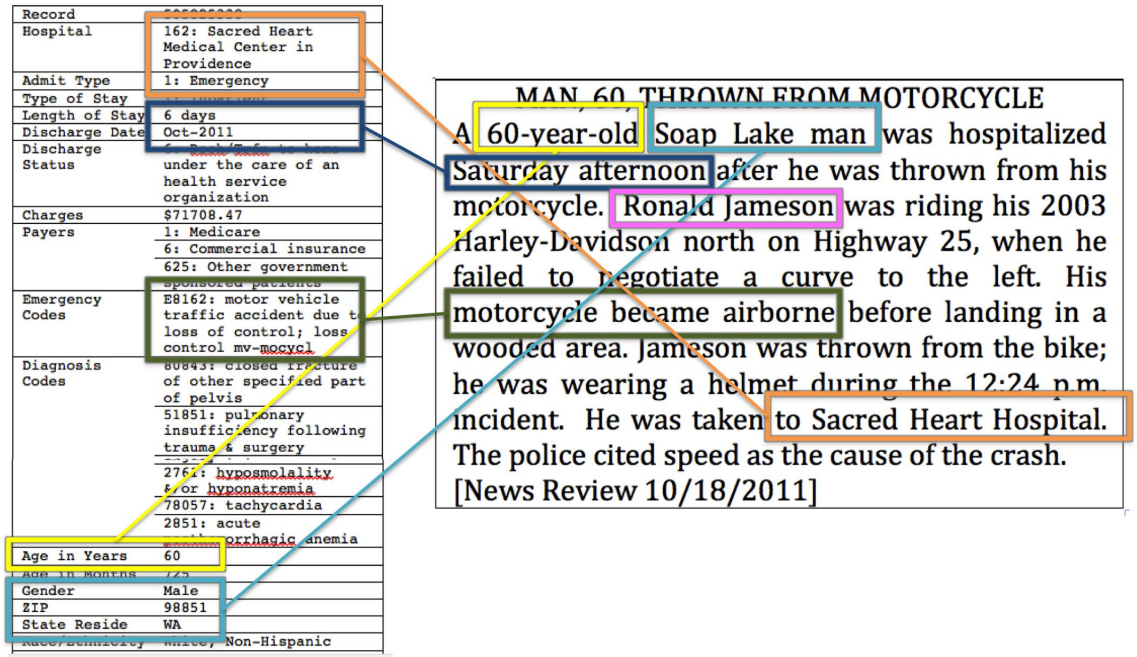
However: anonymizing data is very hard and sometimes impossible!

- UK NHS medical records worth billions [2]
- Google failed to comply with privacy requirements for NHS data set of 1.6 million patients [3]

[1] European Data Portal, “Protecting data and opening data”, accessed 2019/10/20
[2] The Guardian (2019/06/10), “NHS data is worth billions”, accessed 2019/10/20
[3] ICO (2017/07/03), "Royal Free - Google DeepMind trial failed to comply with data protection law", accessed 2019/10/21

Risk of de-anonymizing via (open) data

- Combination of only three publicly available data sets (anonymized health records, news, voters lists) leads to de-anonymization of medical records [1]
- Can be made safer with small improvements on data level (e.g. obfuscate ZIP, remove age in months, etc.)
- Process improvements (e.g. only sharing more granular data with pre-approved organizations)



[1] Sweeney, L. (2015/09/29), "Only You, Your Doctor, and Many Others May Know", accessed 2019/10/20

Potential of open data

- Improving and transforming government/administration:
  - Tackling corruption
  - Increasing transparency
- Empowering citizens:
  - Making informed decisions – e.g. abgeordnetenwatch.de, votewatch.eu
  - Citizen participation
- Creating new business opportunities:
  - E.g. use of public transport and map data → data.gv.at/anwendungen/ lists 485 apps
- Solving public problems:
  - Data-driven assessment and decisions
  - E.g. environmental issues

[1] Open Data Impact, accessed 2019/10/17
[2] Davies, T. (2010), "Open data, democracy and public sector reform.", accessed 2019/10/18
[3] Janssen, M. et al. (2012), Benefits, Adoption Barriers and Myths of Open Data and Open Government. Information Systems Management 29, 4.

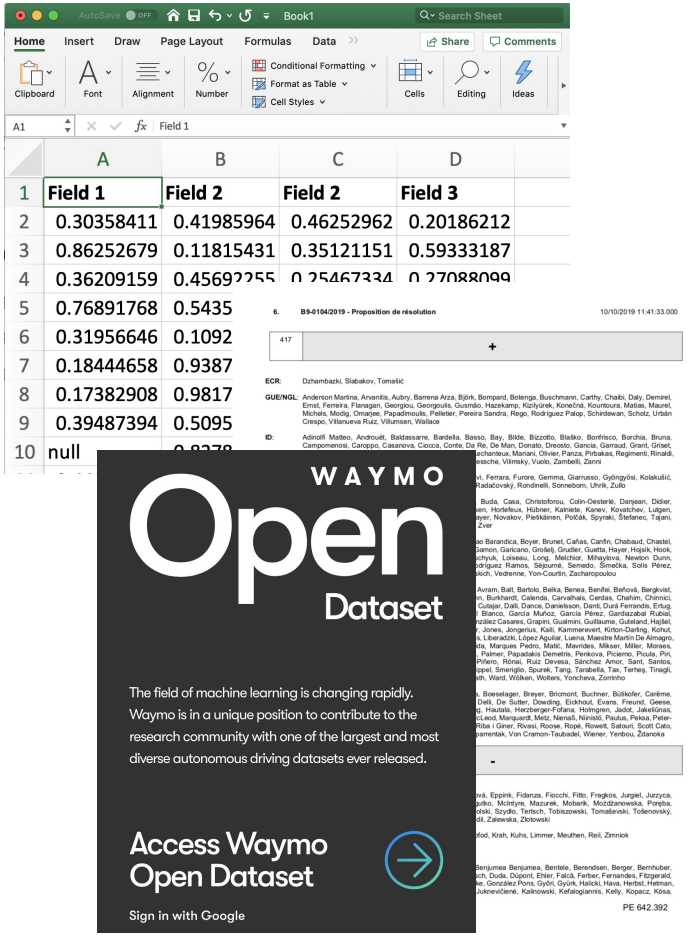
What makes data non-open?

Examples:

- MS Excel sheet on my publicly accessible web server
- European parliament publishes results of votes as PDF
- Open Waymo Dataset accessible via Google Sign-in

What are some problems?

- What are the data about?
- What do the fields mean?
- No-one knows about it
- Not an open format* → need to buy MS Excel
- Who owns the data? How can I use them?
- Does copyright apply to data?
- ...



* Technically, Office Open XML is open and PDF is royalty-free but patented by Adobe

[1] Image © European Union, 2019 – Source: European Parliament

What are open data?

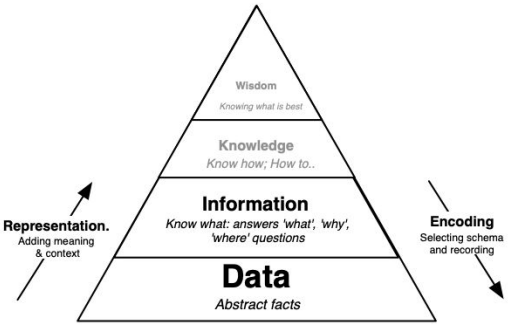
One of many definitions by The Open Definition:

> “Open data and content can be **freely accessed**, **used**, **modified**, and **shared** by **anyone** for **any purpose.**” [1]

→ Synonymous with “free” or “libre” in the Free Software Definition sense

Includes data (“raw material”) in the classical sense
but also **non-textual material**:

- Maps, genomes, chemical compounds,
- Mathematical and scientific formulae, medical data, etc.



[1] The Open Definition, accessed 2019/10/17
[2] Rowley, J. E. (2007), The wisdom hierarchy: representations of the DIKW hierarchy. Journal of Information Science 33, 2.

What are open data? – criteria

- Access:
  - Provided as a whole and only at reasonable one-time reproduction cost
- Machine readability
- Open format:
  - At least one free/libre/open-source software tool available for processing
- Open licenses or public domain:
  - Must allow free **use**, **redistribution** (even commercially), **modification**, **separation**, **compilation**, **non-discrimination**, and **propagation** for **any purpose** and at **no charge**
- Acceptable conditions:
  - Attribution, integrity, share-alike, notice, source, no technical restrictions, non-aggression

[1] The Open Definition, accessed 2019/10/17

Brief historical context

1940s: Robert King Merton: Research (data) should be shared freely for the common good

1950s: International Council of Scientific Unions: open access to scientific data

1960s: US Freedom of Information Act (disclosure of agency records, procedures, and exemptions)

1990s: Human Genome Project: genomic sequence information should be freely available

2000s: People (such as Tim Berners-Lee) started advocating for open data and open government

- 2003: European Public Sector Information Directive
- 2005: UK Freedom of Information Act (“right of access to information held by public authorities”)
- 2006: Guardian launches “Free Our Data” campaign
- 2007: Tim O’Reilly, Lawrence Lessig and others defined 8+7 principles of open government data
- 2009: data.gov launched with 47 data sets (more than 250k now)
- 2009: Obama issues memo on Transparency and Open Government
- 2010/11: data.gov.uk launched; fragdenstaat.de launched by the Open Knowledge Foundation (OKFN)

[1] Davies, T. (2010), “Open data, democracy and public sector reform.”, accessed 2019/10/18
[2] Chignard, S. (2013), “A brief history of Open Data, Paris Innovation Review”, accessed 2019/10/18

Arguments for and against open data (selection)

Pros:

- Public money → public goods
- Help build public trust
- Some data belong to everybody (human genome, environmental data)
- Opportunity to give feedback to data and learn from the public (e.g. change wrong information, add to data)
- Can encourage and support innovation, stimulate economic growth

Cons:

- Cost of data provisioning
- Data quality
- Harder to manage than “closed” data: open data are more affected by external factors (often unpredictable!)
- Potential for unintentional privacy violations
- Potential misinterpretation and misuse
- Potentially “waste” of resources to make invaluable data public

[1] Janssen, M. et al. (2012), Benefits, Adoption Barriers and Myths of Open Data and Open Government. Information Systems Management 29, 4.
[2] Zuiderwijk, A., Janssen, M. (2014), The negative effects of open government data - Investigating the dark side of open data. Proceedings of the 15th Annual International Conference on Digital Government Research.

Economic impact

- European Data Portal study (2015) [1]:
  - Expected 100,000 open data jobs by 2020
  - Public sector in EU to save €1.7 billion in 2020
  - 2020 version of study is currently underway
- Misuse of charities in Canada [2]:
  - Analyses of data about charities showed significant misuse
  - $3.2 billion and 100,000 tax filers involved
- Danish address data [3]:
  - In 2005 Denmark started to release address data free of charge
  - Economic impact between 2005-2009 estimated at €62 million (versus cost of €2 million)

→ See odimpact.org for more case studies

[1] European Data Portal, “Creating Value through Open Data”, accessed 2019/10/20
[2] Eaves, D. (2010), "Case Study: How Open data saved Canada $3.2 Billion", accessed 2019/10/20
[3] GovLab (2016), "Denmark's Open Address Data Set", accessed 2019/10/20

Main sources of open data

- Government:
  - Open government agencies' data
- Commercial/not-for-profit organizations:
  - Media: e.g. Offshore Leaks and Panama Papers published as open data
  - R&D: e.g. Kaggle (owned by Google/Alphabet) provides open data sets for competitions
- Science:
  - See upcoming Open Science/Open Research/Open Access lectures

Licenses

Difficulty: some jurisdictions distinguish **content** vs. **data(base)**

Copyright may apply in a different way to data(bases):

- No copyright in the USA
- EU introduced directive for databases

Example 1: database of photographs:

- May want to license each image separately

Example 2: geospatial database:

- App/platform generates map (image) from data → map is content

[1] Open Definition, "Conformant Licenses", accessed 2019/10/17

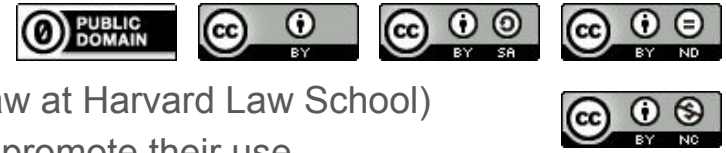
Licenses: main types

- Copyleft (“share-alike”: freely use, study, share, modify) licenses:
  - Creative Commons licenses
  - Open Data Commons licenses
- Open government licenses:
  - UK Open Government License
  - Data License Germany
- Bespoke licenses:
  - E.g. commercial organizations

[1] Open Definition, “Conformant Licenses”, accessed 2019/10/17

Licenses

Creative Commons



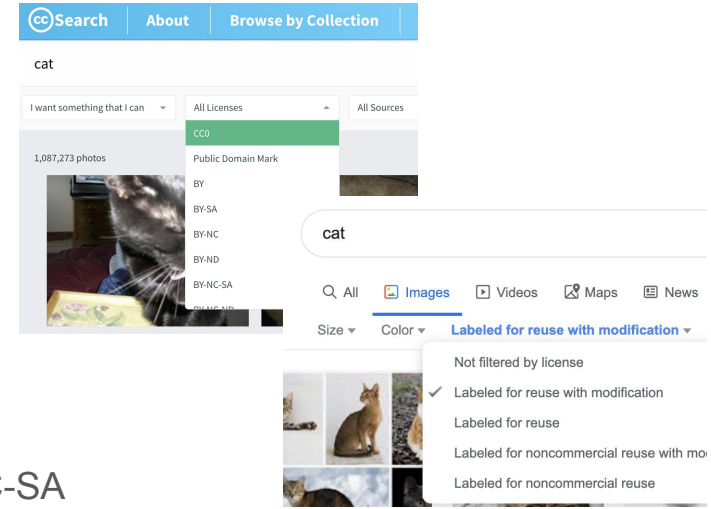
- Co-founded 2001 by Lawrence Lessig (professor of Law at Harvard Law School)
- Goal: create works in public domain to build upon and promote their use
- Introduced a set of licenses:
  - Gives creators a simple, standardized way to grant copyright permissions to their creative work
  - Aimed to be internationally valid
- Basic types (+ combinations) [1]
  - CC0: “No Rights Reserved”
  - CC BY: attribution (distribute, remix, tweak, build upon, even commercially)
  - CC BY-SA: share-alike (new creations under identical terms) → “copyleft”
  - CC BY-ND: no derivatives
  - CC BY-NC: non-commercial use only
- Three-layer design: legal, human readable, machine readable

[1] Creative Commons licenses, “About The Licenses”, accessed 2019/10/08
[2] Wikipedia, “Comparison of free and open-source software licenses”, accessed 2019/10/03

Licenses

Examples of usages (see [1] for a list):

- 2018: 1.4 billion works under CC licenses
  - 415 million pictures on Flickr under CC license
- PLOS (Public Library of Science)
  - publishes ~50.000 articles/year under CC BY
- Wikipedia and Wikimedia Commons widely uses CC BY-SA
- Arduino hard/software under CC BY-SA
- Wired.com photography releases photos under CC BY-NC
- OmegaTau podcast publishes episodes under CC BY-NC
- MIT Open CourseWare materials released under CC BY-NC-SA
- TED Talks videos licensed under CC BY-NC-ND



[1] Wikipedia, "List of major Creative Commons licensed works", accessed 2019/10/07

Licenses

Open Data Commons

- Launched by Open Knowledge Foundation (OKFN) in 2007
- Based on Open Source Definition and Creative Commons licenses
- Addresses the need to differentiate between **“database”** and **“content”**:
  - Example: database of photographs
- Set of licenses:
  - Public Domain Dedication and License (PDDL)
  - Attribution License (ODC-BY) → Attribution
  - Open Database License (ODC-ODbL) → Attribution & ShareAlike
    - Database Contents License (DbCL)
- E.g. used by OpenStreetMap, Open Corporates, OpenSpending, FarmSubsidy, Open Food Facts.

OPEN KNOWLEDGE FOUNDATION

You are free to share, create, adapt; as long as you: attribute, share-alike, keep open.

[1] Open Data Commons, “Licenses FAQ”, accessed 2019/10/17
[2] Summary icons, by Open Knowledge Foundation, CC BY 3.0
[3] OKF logo, by Open Knowledge Foundation, CC BY 4.0

Licenses

Example statement from [1]:

“The Open Food Facts database is available under the Open Database License.
Individual contents of the database are available under the Database Contents License.
Products images are available under the Creative Commons Attribution ShareAlike license.”

[1] Open Food Facts, “Terms of use, contribution and re-use”, accessed 2019/10/17

Recommended licenses

Licenses that are **conformant** with The Open Definition (freely accessed, used, modified, and shared by anyone for any purpose) **and** are:

- Reusable (not specific to organization or jurisdiction)
- Compatible
- Widely used and considered best practice

Conformant but not recommended:

- [UK/Germany/USA/etc.] Open Government licenses → non-reusable

Examples of non-conformant licenses:

- Creative Commons licenses:
  - No-derivatives: doesn't allow modification of data
  - Non-commercial: restricts fields of endeavor

[1] Open Definition, "Conformant Licenses", accessed 2019/10/17

Practical issues: consumer's perspective

Accessibility

- Available where?
- How to access (dump, API)?
- What format (CSV, XML, RDF, PDF)?
- Who can access it (gender, education, etc.)?

Ownership and licensing

- Who is publisher?
- What license?
- Privacy considerations (personal data)?

Support

- How are the data documented?
- Support procedures (error reporting, contact information)?

Form

- Raw?
- Have the data been processed? How?
- How to transform the data before use?
- Compatible with my use case?

Quality

- How current are the data?
- Regularly updated?
- How accurate are the data?
- Are all fields documented and understandable?
- Missing fields (e.g. NaN, null)?
- Are the data biased (only "perfect"/selected data published)?

[1] Open Data Initiative, "The Open Data Consumer Checklist", accessed 2019/10/17; [2] Zuiderwijk, A. et al. (2012), Socio-technical Impediments of Open Data. Electronic Journal of e-Government, 10, 2.; [3] Davies, T. (2010), "Open data, democracy and public sector reform.", accessed 2019/10/18; [4] Zuiderwijk, A., Janssen, M. (2014), The negative effects of open government data - Investigating the dark side of open data. Proceedings of the 15th Annual International Conference on Digital Government Research.

Practical issues: publisher's perspective

Three key principles:

- Keep it simple
- Engage early and often (e.g. with citizens, businesses, developers)
- Address common fears and misunderstandings

Four basic steps:

- Choose data set
- Legal openness: apply an open license
- Technical openness: make data available
  - "Free", in bulk, machine-readable (e.g. PDF is bad)
- Make data discoverable

[1] Open Data Handbook, "How to Open up Data", accessed 2019/10/17

Practical issues: barriers to offering open data

- General risk-averse organizational culture
- Risk/fear of violating legislation: e.g. violating privacy rights (if identifiable data could be included)
- Unclear ownership: e.g. if data were created by several organisations/individuals
- Lack of publishing policy
- Lack of incentives

[1] Janssen, M. et al. (2012), Benefits, Adoption Barriers and Myths of Open Data and Open Government. Information Systems Management 29, 4.
[2] Zuiderwijk, A., Janssen, M. (2014), The negative effects of open government data - Investigating the dark side of open data. Proceedings of the 15th Annual International Conference on Digital Government Research.

Practical issues: certification

By the Open Data Institute

→ Automatic via standardized data publishing platforms (e.g. ckan.org)

	Bronze	Silver	Gold	Platinum
Legal				
Openly licensed & legally reusable (= 'open')	✓	✓	✓	✓
Clear rights statement, detailing any copyrights		✓	✓	✓
Privacy issues addressed		✓	✓	✓
Machine readable rights statement			✓	✓
Practical				
Accessible on the web	✓	✓	✓	✓
Discoverable (linked to from other web pages)		✓	✓	✓
Data is timestamped or up to date		✓	✓	✓
Data will be available for at least a year		✓	✓	✓
Guaranteed timeliness (data always up to date)			✓	✓
Regular backups of data			✓	✓
Quality issues documented			✓	✓
Technical				
Data uses a machine readable format		✓	✓	✓
Data published in content appropriate formats		✓	✓	✓
Data uses open standard machine readable formats			✓	✓
Single consistent URL for downloading data			✓	✓
Machine readable provenance documentation				✓
URLs used as identifiers within data				✓
Social				
Data is documented		✓	✓	✓
Contact details for people to provide feedback and ask questions		✓	✓	✓
Machine readable metadata (documentation)			✓	✓
Social media accounts used to promote data			✓	✓
Forum or mailing list for users			✓	✓
Dedicated comms team building user community				✓

[1] Images by Open Data Institute, "Certificate badge levels", CC BY-SA 2.0 UK, accessed 2019/10/17

Incentives & business models (selection)

Incentives from provider's perspective

- Legal obligations:
  - EU Directive 2019/1024 on open data, to be implemented by 2021
  - EU Horizon 2020 research program [1] mandates open access to research data
- Community building & advancing science:
  - E.g. researchers, Kaggle

Incentives from users' perspectives

- Science:
  - Build on previous research
  - Avoid duplication of effort
  - Speed up innovation

Business models

- App development
- Platforms
  - Figshare.com: research data repository
- Research
  - Kaggle.com: open datasets to advance core business
- Data-related services
  - Consulting, aggregation, cleanup, add-ons
- Content selling
- Journalism

[1] European Commission, "Guidelines to the Rules on Open Access", accessed 2019/10/20

Journalism and open data

Increasing importance of open data in journalism [1]

- E.g. to expose data behind stories, provide key data (selection)

Examples

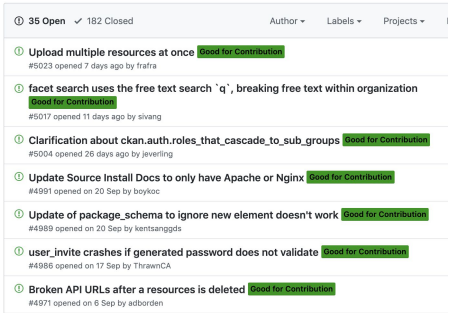
- International Consortium of Investigative Journalists (ICIJ):
  - Offshore Leaks [2] and Panama Papers published/licensed under Open Database License and Creative Commons
- The Guardian Open Platform:
  - free for non-commercial access; usage-based pricing
  - **But:** restricts number of calls per second/day/etc.; restricts commercial use, etc.
- ProPublica:
  - free and non-free data sets (commercial/non-commercial licenses)
  - **But:** restrictive licenses (can't re-publish raw data, distribute, modify, sub-license, etc.)

[1] The Guardian (2012/09/20), "Open data journalism", accessed 2019/10/20
[1] ICIJ, "Offshore Leaks Database", accessed 2019/10/18

Open data project ideas

- Organize an open data hackathon/datathon:
  - Ideas by Open Knowledge Austria (in German) [1]
- Export or request your own data (GDPR!) from a company:
  - E.g. Billa/JÖ, Google, Runtastic, etc. and analyze/visualize the data in new ways
- Build a web app or mobile app based on one of the data sources mentioned:
  - E.g. “Woody” app developed by TU Wien students in 2013 [2]
- Contribute to ckan (open-source data portal platform):
  - Implement new features or see open issues
  - Check out contributing guidelines

See course website for example descriptions!



[1] Open Knowledge Österreich, “Gute Taten für gute Daten”, accessed 2019/10/20
[2] Os, “Woody”, accessed 2019/10/20

An upcoming open data project from Vienna

[1] Shadowmap, [“Shadowmap. Helping Humans Find The Sun."](), accessed 2019/10/20

Outlook

Open research data is a significant aspect of:

- open science/research (lecture #5)
- open access (lecture #6)

An increasing number of funding bodies mandate open access/data:

- E.g. EU Framework Programme “Horizon 2020” defaults all projects to its “Open Research Data Pilot” [1]
- Requires all research data for publication to be open access
- Requires enabling of further use of the data
- **But:** projects may still opt out

[1] FFG, "OLD EN: Open Data in Horizon 2020", accessed 2019/10/20

Upcoming tasks

- Next lecture: Open Content/Open Educational Resources:
  - **Tuesday, October 29**, 17:00–19:00, Argentinierstraße 8, Seminarraum/Bibliothek 194-05
- Project group forming and topic selection:
  - **Friday, October 25**, via email to **both** lecturers
  - See project ideas at https://free-and-open-technologies.github.io
- First project meeting (45 min., discussion of your project idea):
  - **Friday, November 8, 13:00–18:00**, Argentinierstraße 8, **project room**
  - Will send out slots via termino.gv.at
- Paper group forming and topic selection:
  - **Friday, November 29**, via email to **both** lecturers

Literature and resources

Open Knowledge Foundation: Open Data Handbook

European Data Portal: E-learning program on open data

The Governance Lab @ NYU: Open Data's Impact

The Open Data Institute: Guides, reports, etc.

Verhulst, S., Young, A. (2016), "The Global Impact of Open Data", O'Reilly Media