Data Sharing Best Practices

Katie Mika
katherine_mika@harvard.edu
DataFest 2021
https://bit.ly/2KBtk1f

What is **Data Sharing**?

How to share your data

Data Repositories

Other things to think about

Resources

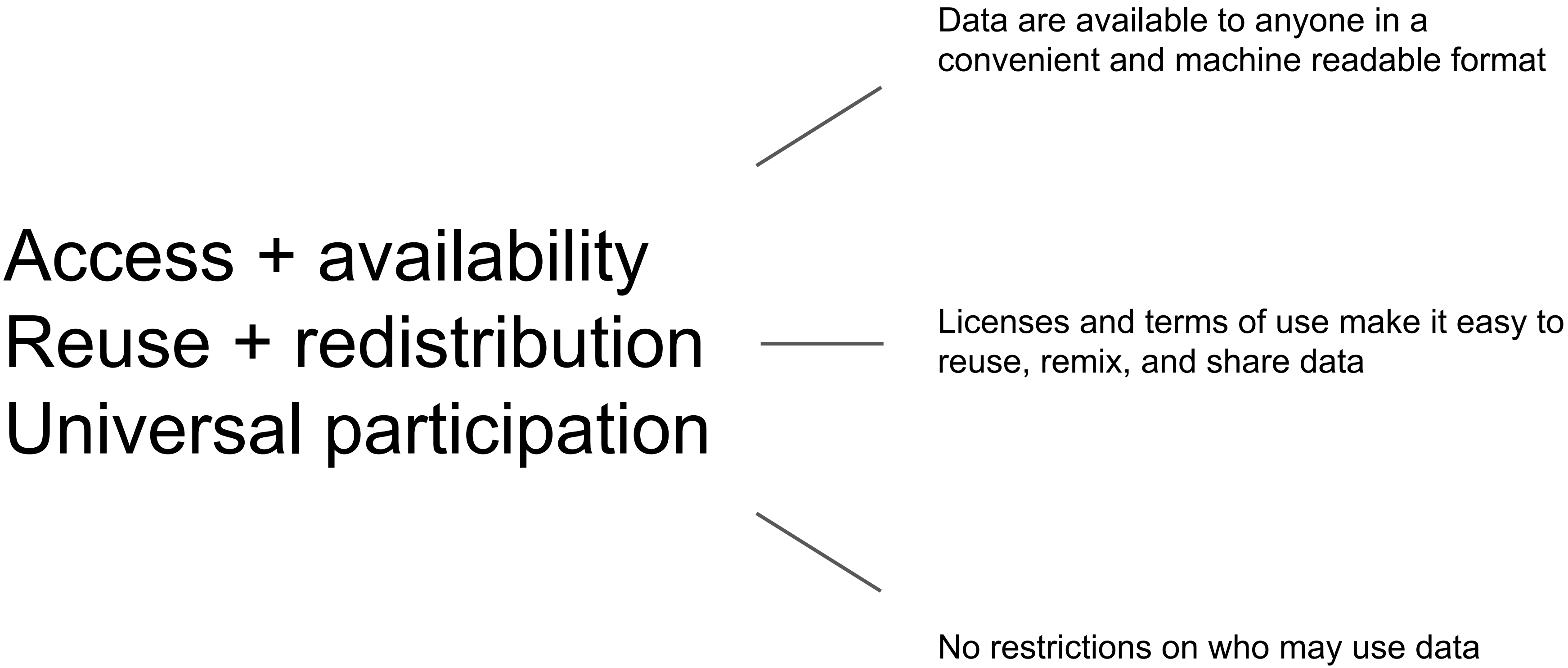
What are data?

“Data refers to entities used as evidence of phenomena for the purposes of research or scholarship.”

“Data are not pure or natural objects with an essence of their own. They exist in a context, taking on meaning from that context and the perspective of the beholder.”

— Borgman, C. (2015). Big data, little data, no data: scholarship in the networked world. Cambridge, Massachusetts: MIT Press.

If you love your data, set it free



Access + availability

Reuse + redistribution

Universal participation

Data are available to anyone in a convenient and machine readable format

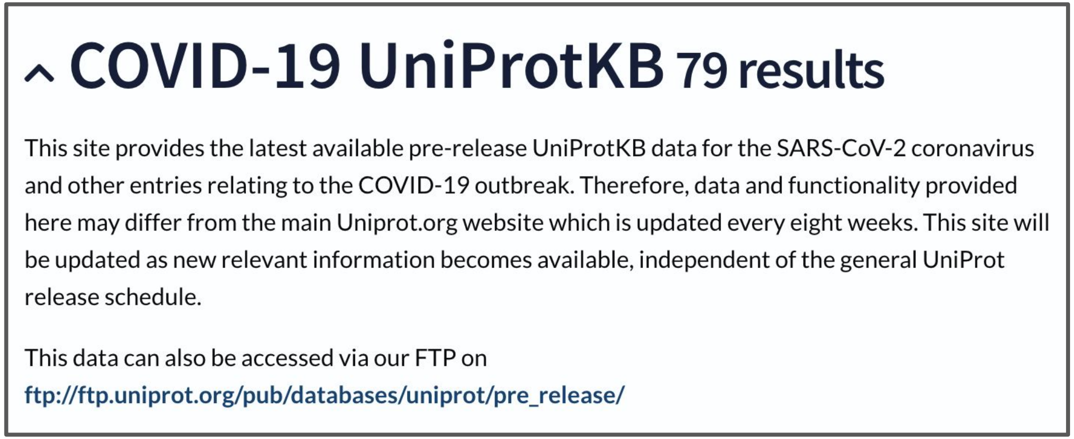
Licenses and terms of use make it easy to reuse, remix, and share data

No restrictions on who may use data

Open Knowledge Foundation: https://okfn.org/opendata/

Benefits of Data Sharing

1. Increase the pace of research and reduce duplication of effort



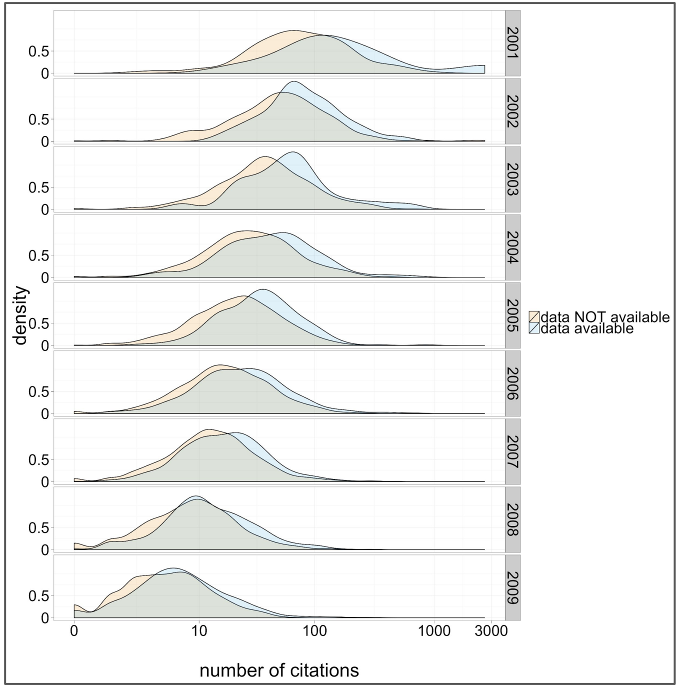
^ COVID-19 UniProtKB 79 results

This site provides the latest available pre-release UniProtKB data for the SARS-CoV-2 coronavirus and other entries relating to the COVID-19 outbreak. Therefore, data and functionality provided here may differ from the main Uniprot.org website which is updated every eight weeks. This site will be updated as new relevant information becomes available, independent of the general UniProt release schedule.

This data can also be accessed via our FTP on ftp://ftp.uniprot.org/pub/databases/uniprot/pre_release/

Benefits of Data Sharing

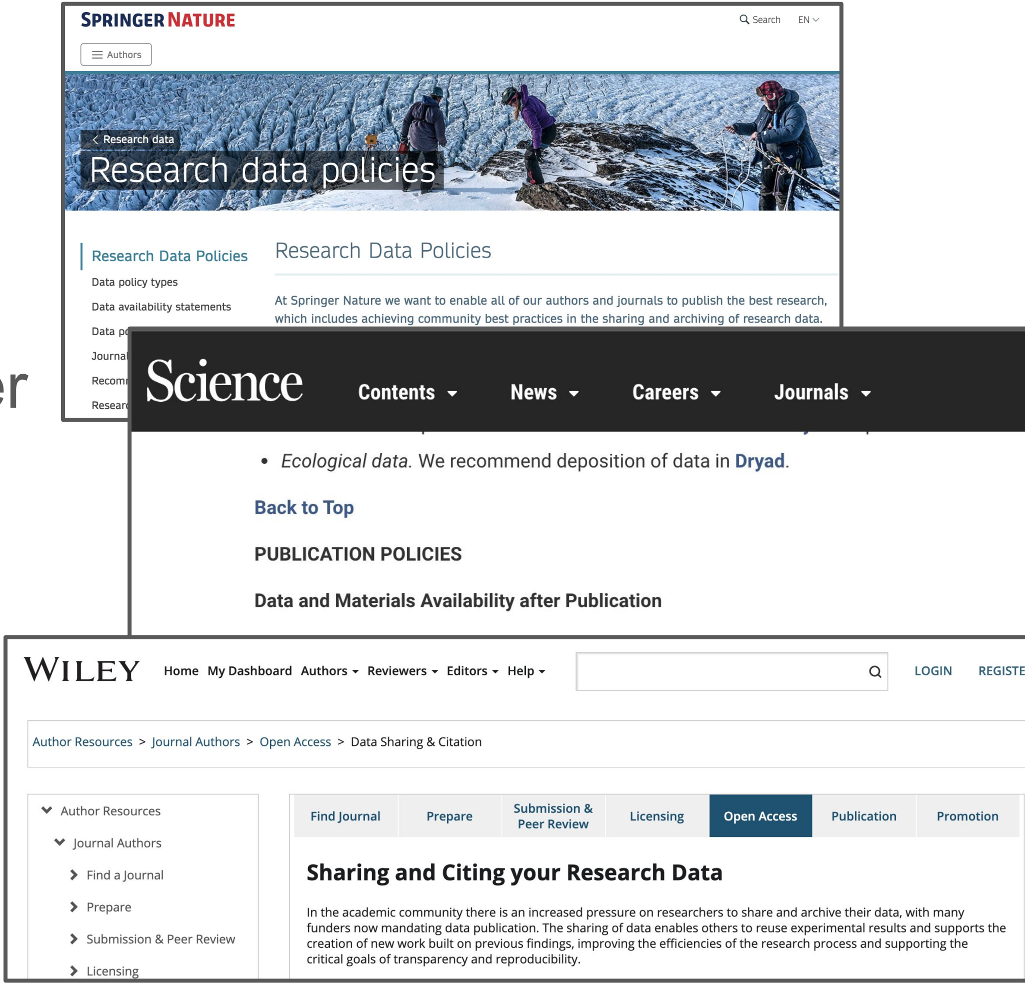
1. Increase the pace of research and reduce duplication of effort
2. Increase scholarly impact



From: Piwowar HA, Vision TJ. 2013. Data reuse and the open data citation advantage. PeerJ 1:e175 https://doi.org/10.7717/peerj.175.

Benefits of Data Sharing

1. Increase the pace of research and reduce duplication of effort
2. Increase scholarly impact
3. Meet publisher and funder requirements



Research data policies

Research Data Policies

At Springer Nature we want to enable all of our authors and journals to publish the best research, which includes achieving community best practices in the sharing and archiving of research data.

- *Ecological data.* We recommend deposition of data in **Dryad**.

Back to Top

PUBLICATION POLICIES

Data and Materials Availability after Publication

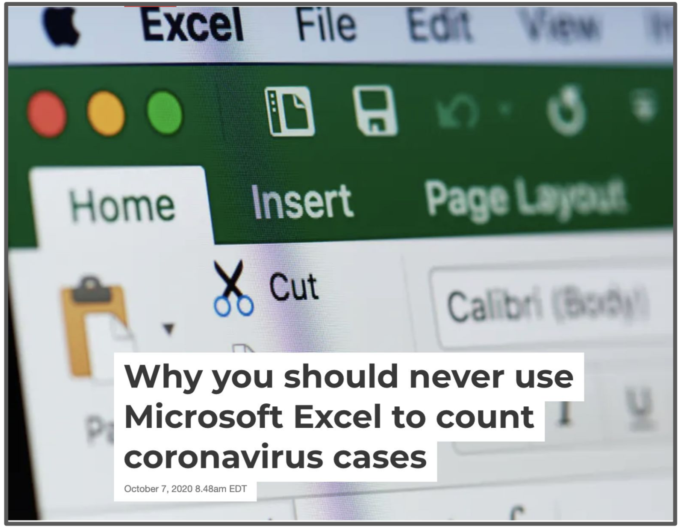
Author Resources > Journal Authors > Open Access > Data Sharing & Citation

Sharing and Citing your Research Data

In the academic community there is an increased pressure on researchers to share and archive their data, with many funders now mandating data publication. The sharing of data enables others to reuse experimental results and supports the creation of new work built on previous findings, improving the efficiencies of the research process and supporting the critical goals of transparency and reproducibility.

Benefits of Data Sharing

1. Increase the pace of research and reduce duplication of effort
2. Increase scholarly impact
3. Meet publisher and funder requirements
4. Preserve valuable and unique data



Why you should never use Microsoft Excel to count coronavirus cases

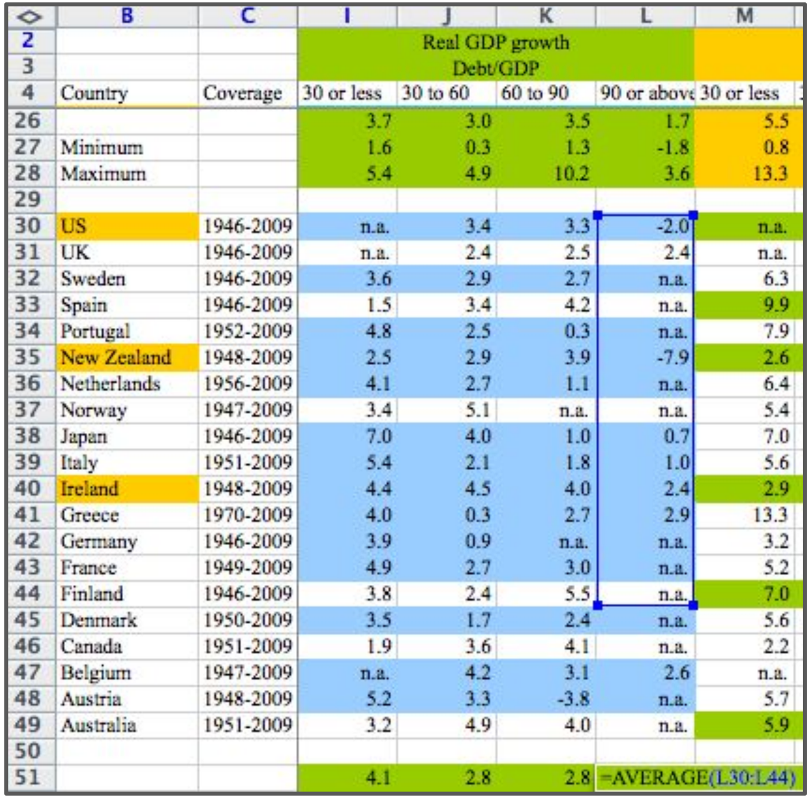
October 7, 2020 8.48am EDT

Benefits of Data Sharing

1. Increase the pace of research and reduce duplication of effort
2. Increase scholarly impact
3. Meet publisher and funder requirements
4. Preserve valuable and unique data
5. Encourage collaboration across disciplines

Benefits of Data Sharing

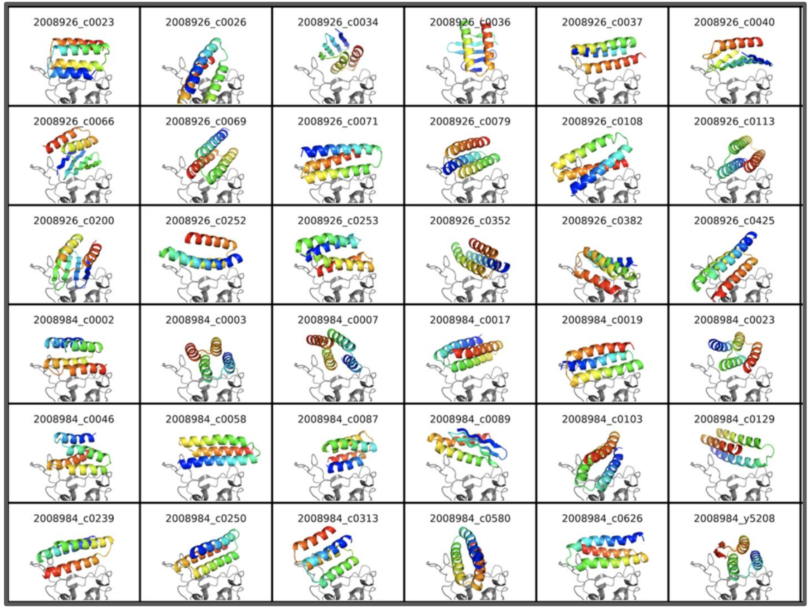
1. Increase the pace of research and reduce duplication of effort
2. Increase scholarly impact
3. Meet publisher and funder requirements
4. Preserve valuable and unique data
5. Encourage collaboration across disciplines
6. Maintain accountability and integrity in results



	B	C	I	J	K	L	M
2			Real GDP growth				
3			Debt/GDP				
4	Country	Coverage	30 or less	30 to 60	60 to 90	90 or above	30 or less
26			3.7	3.0	3.5	1.7	5.5
27	Minimum		1.6	0.3	1.3	-1.8	0.8
28	Maximum		5.4	4.9	10.2	3.6	13.3
29							
30	US	1946-2009	n.a.	3.4	3.3	-2.0	n.a.
31	UK	1946-2009	n.a.	2.4	2.5	2.4	n.a.
32	Sweden	1946-2009	3.6	2.9	2.7	n.a.	6.3
33	Spain	1946-2009	1.5	3.4	4.2	n.a.	9.9
34	Portugal	1952-2009	4.8	2.5	0.3	n.a.	7.9
35	New Zealand	1948-2009	2.5	2.9	3.9	-7.9	2.6
36	Netherlands	1956-2009	4.1	2.7	1.1	n.a.	6.4
37	Norway	1947-2009	3.4	5.1	n.a.	n.a.	5.4
38	Japan	1946-2009	7.0	4.0	1.0	0.7	7.0
39	Italy	1951-2009	5.4	2.1	1.8	1.0	5.6
40	Ireland	1948-2009	4.4	4.5	4.0	2.4	2.9
41	Greece	1970-2009	4.0	0.3	2.7	2.9	13.3
42	Germany	1946-2009	3.9	0.9	n.a.	n.a.	3.2
43	France	1949-2009	4.9	2.7	3.0	n.a.	5.2
44	Finland	1946-2009	3.8	2.4	5.5	n.a.	7.0
45	Denmark	1950-2009	3.5	1.7	2.4	n.a.	5.6
46	Canada	1951-2009	1.9	3.6	4.1	n.a.	2.2
47	Belgium	1947-2009	n.a.	4.2	3.1	2.6	n.a.
48	Austria	1948-2009	5.2	3.3	-3.8	n.a.	5.7
49	Australia	1951-2009	3.2	4.9	4.0	n.a.	5.9
50							
51			4.1	2.8	2.8	=AVERAGE(L30:L44)	

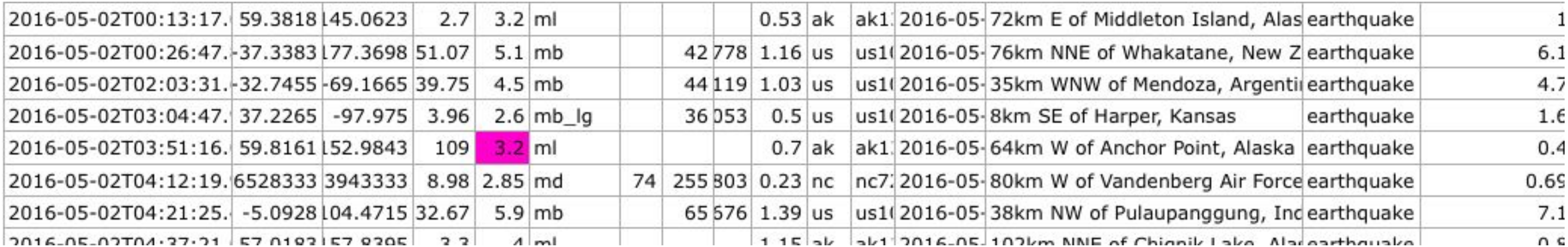
Benefits of Data Sharing

1. Increase the pace of research and reduce duplication of effort
2. Increase scholarly impact
3. Meet publisher and funder requirements
4. Preserve valuable and unique data
5. Encourage collaboration across disciplines
6. Maintain accountability and integrity in results
7. Encourage public engagement with research

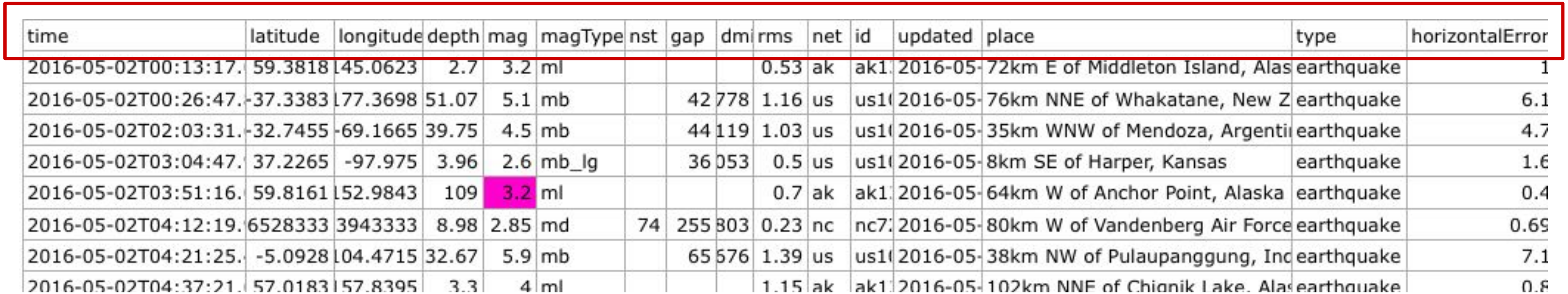


Releasing your data into the wild

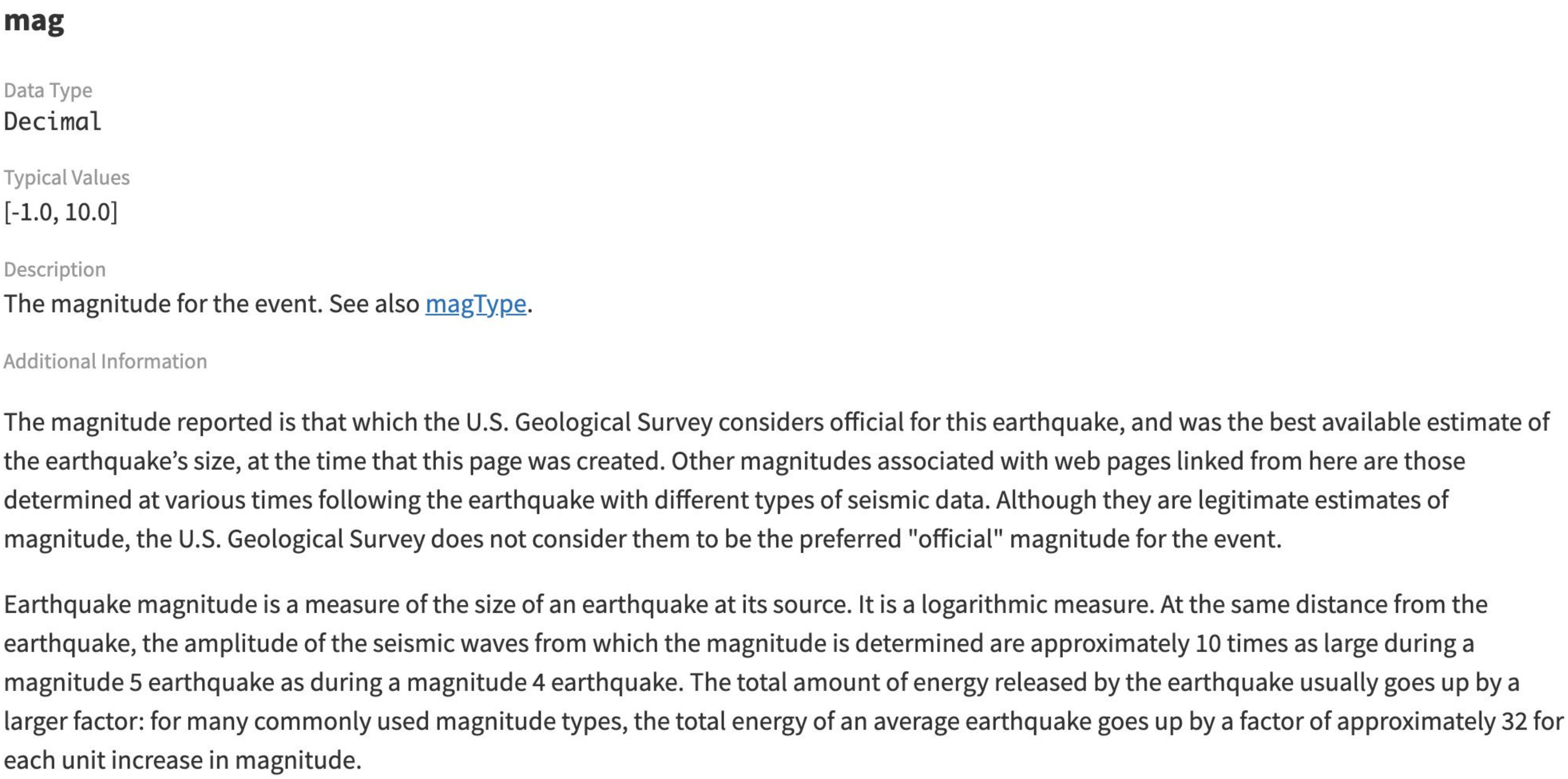
3.2



2016-05-02T00:13:17.	59.3818	145.0623	2.7	3.2	ml				0.53	ak	ak1	2016-05-	72km E of Middleton Island, Alas	earthquake		1
2016-05-02T00:26:47.	-37.3383	177.3698	51.07	5.1	mb		42	778	1.16	us	us1	2016-05-	76km NNE of Whakatane, New Z	earthquake		6.1
2016-05-02T02:03:31.	-32.7455	-69.1665	39.75	4.5	mb		44	119	1.03	us	us1	2016-05-	35km WNW of Mendoza, Argenti	earthquake		4.7
2016-05-02T03:04:47.	37.2265	-97.975	3.96	2.6	mb_lg		36	053	0.5	us	us1	2016-05-	8km SE of Harper, Kansas	earthquake		1.6
2016-05-02T03:51:16.	59.8161	152.9843	109	3.2	ml				0.7	ak	ak1	2016-05-	64km W of Anchor Point, Alaska	earthquake		0.4
2016-05-02T04:12:19.	6528333	3943333	8.98	2.85	md	74	255	303	0.23	nc	nc7	2016-05-	80km W of Vandenberg Air Force	earthquake		0.69
2016-05-02T04:21:25.	-5.0928	104.4715	32.67	5.9	mb		65	676	1.39	us	us1	2016-05-	38km NW of Pulaupanggung, Inc	earthquake		7.1
2016-05-02T04:37:21.	57.0183	157.8395	3.3	4	ml				1.15	ak	ak1	2016-05-	102km NNE of Chignik Lake, Ala	earthquake		0.8



time	latitude	longitude	depth	mag	magType	nst	gap	dmi	rms	net	id	updated	place	type	horizontalError
2016-05-02T00:13:17.	59.3818	145.0623	2.7	3.2	ml				0.53	ak	ak1	2016-05-	72km E of Middleton Island, Alas	earthquake	1
2016-05-02T00:26:47.	-37.3383	177.3698	51.07	5.1	mb		42	778	1.16	us	us1	2016-05-	76km NNE of Whakatane, New Z	earthquake	6.1
2016-05-02T02:03:31.	-32.7455	-69.1665	39.75	4.5	mb		44	119	1.03	us	us1	2016-05-	35km WNW of Mendoza, Argenti	earthquake	4.7
2016-05-02T03:04:47.	37.2265	-97.975	3.96	2.6	mb_lg		36	053	0.5	us	us1	2016-05-	8km SE of Harper, Kansas	earthquake	1.6
2016-05-02T03:51:16.	59.8161	152.9843	109	3.2	ml				0.7	ak	ak1	2016-05-	64km W of Anchor Point, Alaska	earthquake	0.4
2016-05-02T04:12:19.	6528333	3943333	8.98	2.85	md	74	255	303	0.23	nc	nc7	2016-05-	80km W of Vandenberg Air Force	earthquake	0.69
2016-05-02T04:21:25.	-5.0928	104.4715	32.67	5.9	mb		65	576	1.39	us	us1	2016-05-	38km NW of Pulaupanggung, Inc	earthquake	7.1
2016-05-02T04:37:21.	57.0183	157.8395	3.3	4	ml				1.15	ak	ak1	2016-05-	102km NNE of Chignik Lake, Alas	earthquake	0.8



mag

Data Type

`Decimal`

Typical Values

[-1.0, 10.0]

Description

The magnitude for the event. See also magType.

Additional Information

The magnitude reported is that which the U.S. Geological Survey considers official for this earthquake, and was the best available estimate of the earthquake’s size, at the time that this page was created. Other magnitudes associated with web pages linked from here are those determined at various times following the earthquake with different types of seismic data. Although they are legitimate estimates of magnitude, the U.S. Geological Survey does not consider them to be the preferred "official" magnitude for the event.

Earthquake magnitude is a measure of the size of an earthquake at its source. It is a logarithmic measure. At the same distance from the earthquake, the amplitude of the seismic waves from which the magnitude is determined are approximately 10 times as large during a magnitude 5 earthquake as during a magnitude 4 earthquake. The total amount of energy released by the earthquake usually goes up by a larger factor: for many commonly used magnitude types, the total energy of an average earthquake goes up by a factor of approximately 32 for each unit increase in magnitude.

Source: USGS, https://earthquake.usgs.gov/earthquakes/feed/v1.0/csv.php#earthquakes

What to share?

1. Files

- Data, in open formats
- Code
- Figures + output files

What to share?

1. Files
2. Documentation

- Codebooks and data dictionaries
- A README file for high level documentation & metadata
- Information about data collection, equipment, and software used
- Software documentation, especially for custom packages/libraries

What to share?

1. Files
2. Documentation
3. Metadata

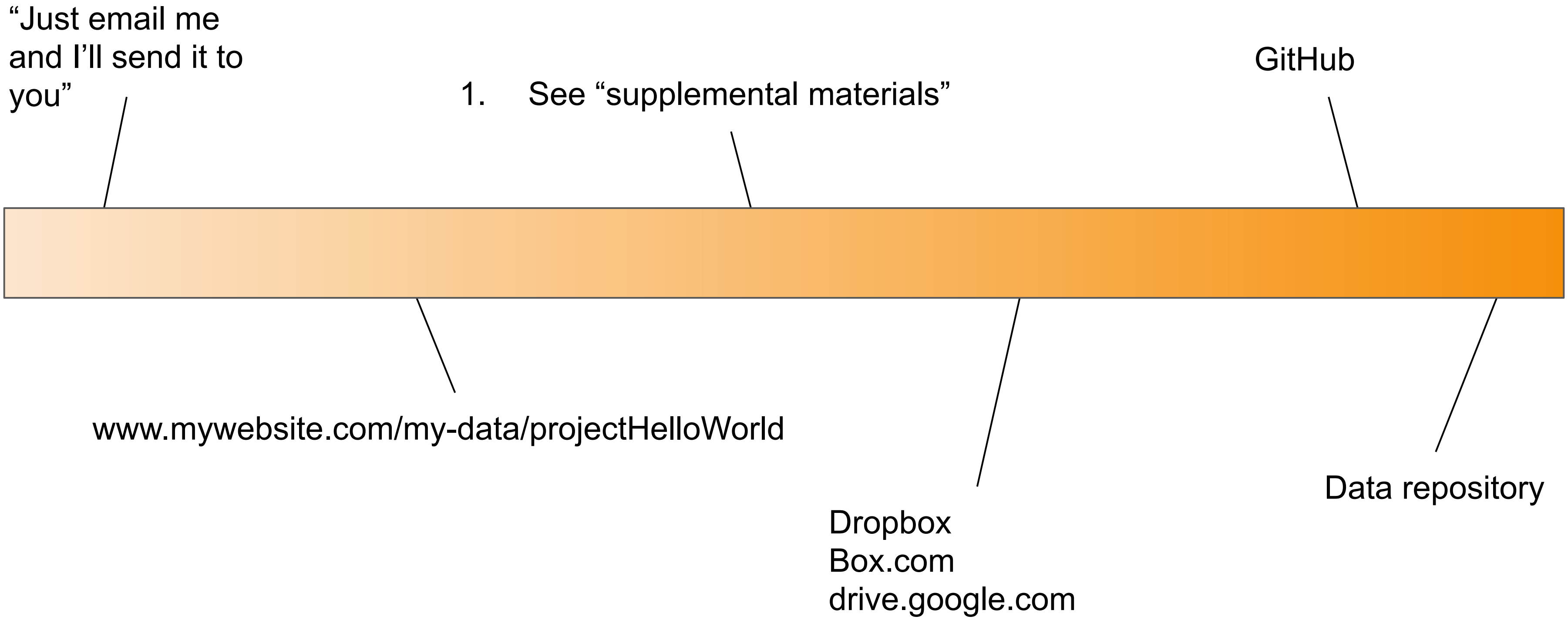
- Title
- Author(s)
- DOI
- General description
- Publication citation
- Discipline specific information

What to share?

1. Files
2. Documentation
3. Metadata
4. License

- Can you share it?
  - Copyright
  - IP
  - Funder mandates
  - Privacy + ethical considerations
- Use a permissive license
  - Creative Commons
  - Be aware of license stacking

How to share your data



“Just email me
and I’ll send it to
you”

www.mywebsite.com/my-data/projectHelloWorld

1. See “supplemental materials”

Dropbox
Box.com
drive.google.com

GitHub

Data repository

Choosing a data repository

1. Does it need to be discipline specific?
2. How much storage do you need and what does it cost?
3. Is there any support or guidance?
4. What kind of files do they take?
5. Does it support code and software?
6. Can you get a DOI?
7. Can you choose a license?
8. What is the preservation & archiving policy?
9. Does it support multiple versions?
10. Will you be the data owner?

Data Sharing Best Practices

- Use open file formats
- Put it in a repository
- Make sure it has a DOI
- Add sufficient metadata
- Include documentation
- Apply a permissive license

Other issues...

- When *not* to share data
    - Personally Identifying Information (PII)
    - Be careful about reidentification
    - Other types of potential harm
        - Endangered species
        - Archeological dig sites
        - Other discipline specific ethical considerations
- What to do with big & unstructured data?
    - What type of data is it? Text, IoT, data mining, etc.
    - Streaming or constantly updated?
    - Who will use these data?
    - Even bigger ethical considerations (algorithmic bias, privacy, etc.)

Data Sharing Resources

- Getting Started with Dataverse: https://support.dataverse.harvard.edu/getting-started
- The Open Data Institute: https://theodi.org/, especially the Data Ethics Canvas: https://theodi.org/service/consultancy/data-ethics/
- Open Science MOOC: https://opensciencemooc.eu/
- Australian National Data Service guides: https://www.ands.org.au/working-with-data/publishing-and-reusing-data/data-reuse
- Open Science, Open Data, Open Source: https://pfern.github.io/OSODOS/gitbook/
- COVID Data Sharing: https://www.nature.com/articles/d41586-020-01516-0
- “Beyond Open Data” Talk and GitHub repo: https://github.com/saverkamp/beyond-open-data, especially the Data Packaging Guide: https://github.com/saverkamp/beyond-open-data/blob/master/DataGuide.md
- More information on licensing research data: https://www.dcc.ac.uk/guidance/how-guides/license-research-data