LOVE DATA WEEK @ EPFL
February 12th - 25th, 2018
Empowered by EPFL Library

EPFL Library is happy to announce the 3 winners of Love Data Week 2018 *"Tell your data story"* contest!

1 - Michaël Defferrard

At some point, we happened to sit on a load of collected audio recordings. That is more than 100k tracks and a year of music under Creative Commons! Though it was not my main research direction, it was too great and useful to let it there. It should go out in the world we thought! After much more work than envisioned --- pretty typical as far as research planning goes --- we published a paper [1] in the premier music conference, and shared all the data online [2]. In the spirit of open science, we released the code to reproduce it as well [2]. As if that was not enough, we then organized a challenge at the Web Conference with the EPFL crowdAI platform to promote open data and benchmarking [3]. That was an hurray data story! The paper was well received at the conference, many people liked it online, and the challenge gathered more than 550 submissions. :D

[1] https://arxiv.org/abs/1612.01840
[2] https://github.com/mdeff/fma
[3] https://www.crowdai.org/challenges/www-2018-challenge-learning-to-recognize-musical-genre

2 - Susan Kamal

On a beautiful spring day in April 2015, sun shining across the white mountain tops, the smell of flower blossom is in the breeze, I was full of joy to start my PhD.

During the interview they told me they had a rich database of patient data ready to be analyzed. I was excited to start unraveling the mysteries behind those data, to tell the story of those patients. Except for, the database was really more of an archive of patient files on paper. I do not know why they called it a database. This linguistic mishap then shaped the future of the three coming years. I had to start that data base to be able to analyze the data!

There were about 8000 patient files, some were scanned and the paper files could no longer be accessed. I could no longer feel the spring breeze. I was smelling old paper smell all day long. Covered in piles and piles of paper, hoping to have one of those "Eureka!" moments…

Then one of the nurses comes to the office and looks at this new miserable girl at work. She says "bonjour" and her eyes look puzzled at me and wonders what my task is in the archives dungeon.

"First day at work, huh?" She asks. I nod. She resumes "I had trouble too when I first got here. I was asked to use this computer to print all those files you see here! "

My brain sees a spark of optimism and I ask: “Computer? So you use a computer system around here?”

I follow her to the office upstairs. She shows me where all the data is stored. Turns out, they had a software to enter the patient information, but they just keep the files because it is easier to use. I dig into the software documentation. I read that the company who provided the software has a tech support. Hurray!

I email and explain the situation. “Can the data be exported?” I ask, hoping the answer would be yes. I waited a day, then two, and I never got a response. On the third day I got a message. The tech guy said: “the older data that was entered 2004-2008 cannot be exported as the software was old and there was no data back-up or storage. But data starting 2009 can be exported.”

“It will take some days”, he wrote. And he has to do it over the weekend, to minimize risk of data loss when extracting the data while the software is in use.

Hurray!

The data arrived in raw form, completely unstructured. I spent the next year structuring the data. But finally, I have a database that I can analyze!

The end.

3 – Victor Janevski

World 2017: Total War

This story talks about the armed conflicts the World countries have experienced mostly in the period 1989-2016. Using the publicly available UCDP datasets, the goal is to visualise the relationships between the conflicting sides, including information about the suffered casualties and severity of the conflicts, but also to aggregate by country or by region and graphically represent the available data. Unfortunately, this is a “horror” data story, because judging by the obtained results, there is no decreasing trend on conflict activity or the number of battle-related casualties. On the contrary, the plots show extreme peaks of deaths in recent years, higher values compared to the entire period from 1989. Another proof that conflict activities are not reducing is the fact that in the centre of the largest component of the conflict graph is the node representation of IS (Islamic State), which has acted in more recent years and gained global prominence in 2014. The ultimate motivation of the data story is to raise awareness about the global conflicts and the negative consequences thereof that affect thousands of people worldwide.

Read more in depth under: https://vjan-fin.github.io./

MONDAY TOPIC: DATA QUALITY

It's all about documentation

Would you say it is qualitatively good data? What actually makes it good? Have you documented it in a specific way? How do you keep track of changes when your dataset is updated? Data quality is the ability of a given dataset to serve an intended purpose. In other words, to deliver the insights someone hopes to get out of it. For this, several aspects have to be taken into account:

- Availability; whether data is available and updated regularly
- Usability; whether the data is credible and checked regularly
- Reliability; whether data is accurate, consistent, integer and concrete
- Relevance; whether the data is actually the data described in the dataset
- Presentation quality; whether the data is clearly described and understandable.

Furthermore, at least four activities impact the quality of data:

- Modelling the world (deciding what to collect and how)
- Collecting or generating data
- Storage/access to the dataset
- Formatting/transformation of data files
- In order to have a clear and understandable follow-up, a neat and understandable documentation of is needed.

Useful resources

- If you need to know more about documentation, just take a look the data documentation part of EPFL Library Research Data Management Website.
- If you are thinking about using an Electronic Laboratory Notebook or a Laboratory Information Management System. If this is the case, take a look at this chapter.
- Goodman, A., Pepe, A., Blocker, A. W., Borgman, C. L., Cranmer, K., Crosas, M., … Slavkovic, A. (2014). Ten Simple Rules for the Care and Feeding of Scientific Data. PLoS Comput Biol, 10(4), e1003542. https://doi.org/10.1371/journal.pcbi.1003542
- Bad data issues guide https://github.com/Quartz/bad-data-guide
- Check out Kristin Briney's post on taking better notes
- Reining in your metadata – advice from an archivist
- Wilson, Greg, Jennifer Bryan, Karen Cranston, Justin Kitzes, Lex Nederbragt, et Tracy K. Teal. 2017. « Good enough practices in scientific computing ». PLOS Computational Biology 13 (6):e1005510. https://doi.org/10.1371/journal.pcbi.1005510.
- Griffin, P. C., Khadake, J., LeMay, K. S., Lewis, S. E., Orchard, S., Pask, A., … Schneider, M. V. (2017). Best Practice Data Life Cycle Approaches for the Life Sciences. BioRxiv, 167619. https://doi.org/10.1101/167619

- Data quality assessment (provides a table of various quality dimensions and their definitions): Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data quality assessment. Communications of the ACM, 45(4), 211. http://doi.org/10.1145/505248.506010
- Faulkes, Zen. 2017. « Stinging the Predators: A collection of papers that should never have been published ». https://doi.org/10.6084/m9.figshare.5248264.v4.

TUESDAY TOPIC: DATA SHARING

As open as possible, as closed as necessary!

"Managing and sharing research data as openly as possible is one of the principles of good scientific practice." The Swiss National Science Foundation writes this in its guidelines for researchers regarding data management planning. In the same document, they also propose to their researchers to share their data according to the FAIR principles whenever possible on repositories.

But, where to publish? And how to decide under which licence?

Useful resources

- There is a dedicated page for data repositories on the EPFL Library Research Data Management website.
- To find the right repository, visit re3data.org. This website gathers most of the repositories internationally available repositories. You can search repositories by discipline.
- If you are looking for sharing data which is too large for your mails, take a look at this page from EPFL IT. You can also use SwitchDrive with your AAI access from EPFL.
- Did you know that Zenodo, a repository based at CERN has a specific EPFLspace? So, do not hesitate to upload your data!

Other experiences

- This post from PLOS blog : And This is Why We Should Always Provide Our Data
- This article published in Peer J Comp Sci : Achieving human and machine accessibility of cited data in scholarly publications
- PLOS Open Data Collection
- This blog post from Ecology Bits, by Margaret Kosmala : Open data, authorship, and the early career scientist

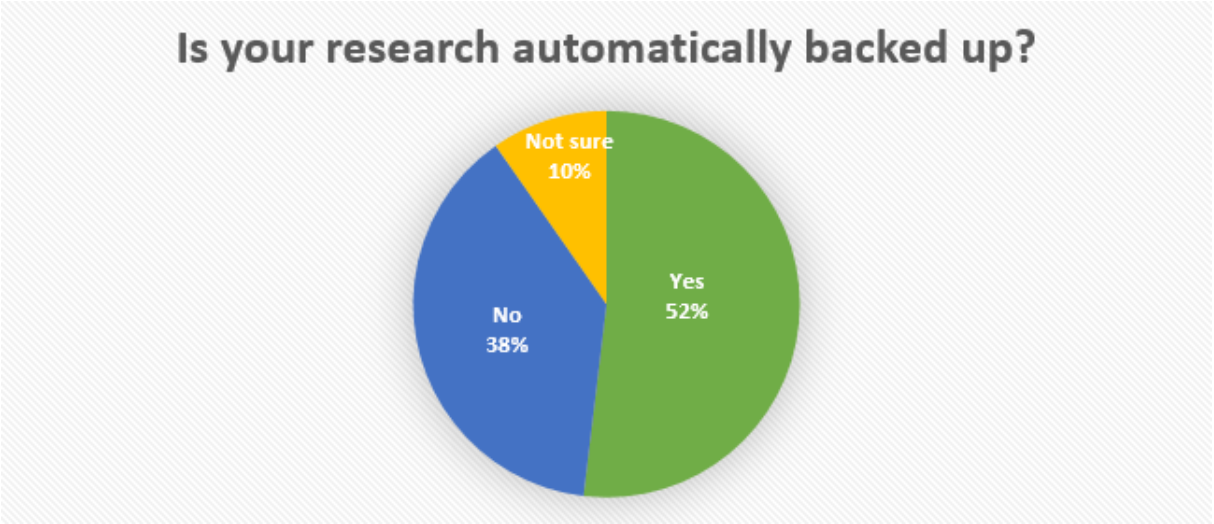
LOVE YOUR DATA

LOVE DATA WEEK @ EPFL
February 12th - 25th, 2018
Empowered by EPFL Library

WEDNESDAY TOPIC: DATA STORAGE

Is your data backed up?

Loving your data is backing it up. The survey about data management practices at EPFL at the end of 2017 shows the following results regarding automatically back-up.



What are your back-up strategies? Do they align with EPFL good practices? If you are not sure, you can take a look at the back-up chapter of the EPFL Library Research Data Management Website.

Practical tips

- Check storage and backup solutions made available on the EPFL Library Research Data website.
- You may want to consider using an Electronic Laboratory Notebook or a Laboratory Information Management System to collect and store your data. If this is the case, take a look at page about active data management.
- If you need further help to find an appropriate backup solution, contact us at: researchdata@epfl.ch

THURSDAY TOPIC: DATA MANAGEMENT PLANNING

Keep calm and fill in your DMP

LOVE YOUR DATA

LOVE DATA WEEK @ EPFL

February 12th - 25th, 2018

Empowered by EPFL Library

Data Management is not just a buzzword. An increasingly number of funders requires now to develop and implement Data Management Plans (in short, DMPs). It is the case of Horizon 2020 and Swiss National Science Foundation, for instance. To know more about the requirements funders have made clear, take a look at the comparison on the EPFL Library Research Data website.

- Why should you plan your data management?

1. To save time, resources and efforts
2. To optimize research
3. To encourage collaboration and increase research impact and visibility
4. To anticipate costs that can be reimbursed by the funding agencies
5. To comply with the funders', publishers' and institutions' requirements

The EPFL Library Data Management Team has prepared several tools – such as templates and data formats as well as data publication licences checklists. Take also a look at the screencast, which explains you how to fill in the DMP-SNSG template.

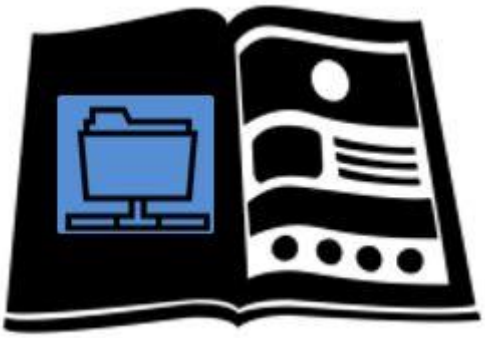
FRIDAY TOPIC: DATA PAPERS

Spotlight on your data

"Data is the new oil" has been a trendy quote for a few years now. Whether you believe in this sentence or not, the data you created in your project is certainly worth a lot. It was extracted with rigorous methods and you want to make the most of it. Why not write a data paper?

What is a data paper?

A data paper is different from a classic scientific article in the sense that the latter analyses and interprets data while the data paper thoroughly describes a dataset so that it can be understood and easily reused by other scientists.



For more information, read this case study from the University of Bristol or this instructions page from BMC about their new article type : data notes.

How to publish a data paper?

If you want to make data the subject of your next paper, check where to publish it. Contact us for personalized guidance: researchdata@epfl.ch

If you have other questions, do not hesitate to contact us: researchata@epfl.ch

#lovedata18 **@EPFLlibrary**