

01. Backup is key

Because they had forgotten their key they had to go through another year of "physical" hardships.

A team of students and academics worked on a big physics study for over a year, involving complex videos taken of micron-sized patterns, which was a very long and time-consuming process. Throughout the year, they gathered several terabytes of video. The videos were labelled systematically and the experimental info stored in a Excel file. Without this Excel file all the data is essentially meaningless for analysis. Regular and multiple backups of the video files were made, a process which took a long long time with all that data. Backing up the key file, however, was not given any thought.

The graduate student in charge of the file cleaned up his computer and somehow managed to delete the file, rendering the whole years' work meaningless. Attempts at recovery failed and the research was set back by over a year.

The story shows that a Data Management Plan makes sense right from the start. It identifies both the amount of data to be backed up and their relevance in the research process. Therefore, the DMP (hopefully) prevents central files from being overlooked during backup because they have a relatively small volume compared to others, making their backup seem comparatively trivial.

Source:

- personal communication

02. Costly Birthdays

She just couldn't finish her cost analysis, because those birthdays kept on interfering.

When she tried to create a cost analysis for electronic media, the employee was surprised when her table repeatedly showed absurd results at the end of her calculations. After a quick search, it became clear that when importing the CSV file for the usage statistics into Excel, the prices had automatically been transformed into dates. Usage statistics for electronic media are typically delivered as CSV files. If these are imported into an Excel file in the basic settings, prices are converted to dates, even though this is not the desired outcome. This prevents a meaningful interpretation of the data since no average usage costs can be calculated if a part of the prices is systematically excluded.

The example shows that when importing data into spreadsheet programs, careful attention must be paid to the correct formatting of the cells, otherwise automatic formatting changes may occur.

Source:

- personal communication

03. Lost Toys

Only because an innocent child was born Woody and cowgirl Jessi were able to make their relationship public.

Pixar came very close to losing a very large portion of Toy Story 2, because someone did an "rm *" (non geek: "remove all" command). And that's when they realized that their backups hadn't been working for a month. Fortunately, the technical director of the film realized that, because she wanted to see her family and kids (including a new baby), she had been making copies of the entire film and transferring it to her home computer. After a careful trip from the Pixar offices to her home and back, it was discovered that, indeed, most of the film could be saved.

The example shows that even with state of the art backups, data loss can occur when multiple unfortunate coincidences coincide. In general, it is very important to store files in more than one place. A good basis for securely storing data is provided by the 3-2-1 rule. It states that data should be backed up to at least 2 different storage media in three different locations, one of which should be decentralized. In addition, to ensure sufficient reliability of backups, the effectiveness of the backup should be tested regularly. In practice, this is done by retrieving central files from the backup at random times and comparing them with the originals.

Source:

- <https://kottke.org/12/05/how-pixar-almost-deleted-toy-story-2>
- <https://www.techdirt.com/articles/20120514/17243918918/how-toy-story-2-almost-got-deleted-except-that-one-person-made-home-backup.shtml>
- https://de.wikipedia.org/wiki/Toy_Story_2

04. The Sound of Silence

When he returned to his room after a long time, there was no music to be heard.

In 2019, the social networking site MySpace announced that they lost almost all music, image and video data uploaded between 2003 and 2015 when migrating the data to a new server. According to MySpace, there is no back-up for the lost data. At the outset, the online platform was primarily intended for musicians to present their works and was comprised of more than 50 million pieces of music. This includes early works by artists who started their careers on MySpace. Especially for older posts, there are often no local copies of users because they relied on storage in the cloud. Thus, the works may be lost forever.

The example shows very well that it always makes sense to store files in more than one place. A good basis for securely storing data is provided by the 3-2-1 rule. It stipulates, that data should be backed up to at least 2 different storage media in three different locations one of which should be off-site.

Sources:

- <https://www.heise.de/newsticker/meldung/Datenverlust-Myspace-verliert-riesiges-Musikarchiv-4338737.html>
- <https://taz.de/Datenverlust-bei-MySpace/!5581108/>

05. Out of the blue

After his collaborators stared into the clouds together, they dart an angry glance at him.

At the end of the semester, a group of students worked on a final report for a study course. They edited the document together in an iCloud Drive and were about to complete the assignment. One of the students deleted the report in the belief that it was an old draft. At that time, there was no data recovery feature, and the group had to rewrite the report from scratch. It is not surprising that the other group members were not too keen on their fellow student during this time.

The example shows quite well that storage in only one place is not sufficient to ensure secure data retention. In general, data should be backed up according to the 3-2-1 rule. Accordingly, there should be at least 3 copies of the data on 2 different storage media, one of which should be stored in an external location. In addition, when working together in a central location, you should be especially careful when removing/deleting files.

Source:

- personal communication

06. Gene Dating

The meta-analysis seems to imply that September 2nd plays a crucial role in the function of cells.

If files are imported or entered into spreadsheet software (such as Microsoft Excel) using the default settings, entries in cells are sometimes automatically reformatted.

A study from 2004 found that this error also occurs frequently in spreadsheets with gene names in scientific publications. The names are automatically changed to dates or floating-point numbers. Since these changes are irreversible the original information regarding the involved genes is lost.

A more recent study from 2016 revealed that the problem is still existent and no standardized solutions have been developed, yet. Roughly 20% of the investigated articles published in high-ranking journals in the field of genomics contained errors in the names of genes in spreadsheets. One example is the SEPT2 (Septin 2) gene, that plays an important role in the function of cells and is typically changed to 2 September in spreadsheets.

Since the data from such studies are a valuable resource for the scientific community and the data are usually reused by other scientists, the loss of information is extremely problematic. A spokesman of Microsoft responded to the results of the study: "Excel is able to display data and text in many different ways. Default settings are intended to work in most-day-to-day scenarios". This means that the documentation of scientific data like genetic analysis is not part of the daily business of most spreadsheet tools and that proceeding with additional care is necessary when using these programs for certain data types.

The example shows that it is important to verify the accurate formatting of cells and the proper transfer of data into spreadsheets.

Sources:

- <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-5-80>
- <https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-1044-7>
- <https://www.bbc.com/news/technology-37176926>

07. Unsure Footing

As the ship sinks into the floods, carpenters stare at their feet with concern.

The Swedish galleon Vasa was a prestige project of Swedish King Gustav II Adolf and one of the largest warships of its time. On August 10, 1628, on her maiden voyage, she sank after only about one kilometre at sea. An investigation into the cause and a subsequent trial found that the ship was inherently unsound, because it carried too much weight in the upper structure of the hull. However, no major culprit was convicted.

Archaeologists have since found that the use of different measurements for lengths contributed to the disaster. Although everything was specified in foot during the ship's planning, this information was not standardized at the time. For example, one group of carpenters used rulers in "Swedish feet" which are divided into twelve inches while the other used "Amsterdam feet" rulers divided into only eleven inches. Because the teams of builders working on the ship were using subtly different rulers but were following the same instructions, this produced parts of different sizes, contributing to the ship's asymmetry.

The example shows just how important it is to use common and well-defined standards for projects. Not only comparability and traceability depend on them, but sometimes even the overall success of the project.

Sources:

- <https://www.pri.org/stories/2012-02-23/new-clues-emerge-centuries-old-swedish-shipwreck>
- <https://www.instm.org/Festival/Why-Measurement-Matters>
- [https://de.wikipedia.org/wiki/Vasa_\(Schiff\)](https://de.wikipedia.org/wiki/Vasa_(Schiff))

08. Babylonian Confusion

Are you Japanese? This may bode ill for your reputation.

On January 1st 2020 a minor lexical revolution occurred in Japan. A new decree ordained that official documents should reverse the order of Japanese people's names when they are rendered in the Latin alphabet.

Until now in English documents Japanese names had to be written with the given name first, using the Western standard. Henceforth, the family name will come first and, to eliminate any ambiguity, may be entirely capitalized. Thus, Japanese naming conventions are now identical for Japanese and foreign characters, at the cost of breaking with the transcription standard that had been in force until then. This could result in a higher number of wrongly or non-attributed quotes, and thus a relative loss of reputation for Japanese authors. This will last until the providers of citation metrics and, more importantly, those citing Japanese authors have become accustomed to the new convention.

The example shows that it is always convenient to have a persistent identifier to work around such problems. For people, for example, there is the non-profit scheme ORCID as well as Thompson-Reuter's ResearcherID. If these PIDs are used for scientific quotations, articles are correctly attributed to their authors, regardless of changed naming conventions. Those citing others are able to more easily keep an eye on the contributions of authors despite changes in name or convention. For the authors themselves, correct attribution of articles is indispensable for their academic reputation.

Sources:

- <https://www.economist.com/asia/2020/01/02/why-japanese-names-have-flipped>
- <https://www.forschungsdaten.info/themen/bewahren-und-nachnutzen/persistente-identifikatoren/>
- <https://orcid.org/about>
- <http://www.researcherid.com/?returnCode=ROUTER.Unauthorized&Init=Yes&SrcApp=CR#rid-for-researchers>

09. The forgotten Fantasy?

Despite 20 years of waiting for the game, not all fantasies were fulfilled.

The PlayStation 1 RPG games Final Fantasy VII, VIII and IX developed by SquareEnix (at that time still Squaresoft) are considered the "golden era" of Final Fantasy history and many new games still have to compete with them. Nevertheless, SquareEnix has been silent for a long time about why the eighth installment of the series was never properly set up for PC or a current generation of consoles - although this was done with almost every other part of the series.

The answer was easy and emerged through a series of interviews: The raw data (background images, music, 3D models etc.) and the finished source code of the game had simply not been archived. In the 90s there were fewer AAA titles (large projects that ran over several years) and games were produced at shorter intervals. As with Square, this meant that space had to be freed up again and again for new projects and little attention was paid to old project data. As a result, all Final Fantasy VIII project data was lost and a new version of the game could only be released in 2019 in collaboration with other companies.

Although there are now many "remastered" versions of the Final Fantasy games at that time, most of which have been re-programmed, the lack of raw data is still a problem for the developers and the gaming community. For example, the background graphics of the games were created in high resolution, but only the compressed versions for the consoles of the time were stored. Due to today's HD displays, those are no longer up-to-date.

This example shows how important it is to archive both the raw data and the actual project data. Publishing or completing a project should not mean the end of the data life cycle. Only when the raw and result data is properly organized can follow-up projects re-use the data and previous work be appreciated.

Sources:

- <https://www.vg247.com/2018/09/14/isnt-ps4-xbox-switch-port-final-fantasy-8-preservation-may-answer/>
- <https://www.vg247.com/2019/01/09/final-fantasy-7-hd-remaster-remake-upscale/>
- <https://ffviiiiremastered.square-enix-games.com/>

10. Hello? Are you there?

It should have been routine, but suddenly nobody could talk to anyone else.

In 2009, T-Mobile was the largest mobile network provider in Germany with over 40 million customers. Nevertheless, on April 21st a service disruption occurred at around 4 p.m., which was to go down as the biggest in history. In one fell swoop, millions of customers were suddenly unable to connect to the network. Calls could not be connected, nor could SMS be sent. The reason for this was a simultaneous failure of all three home location registers. Together, these three servers form a distributed database and are a central component of every mobile network. Normally, the network could still function as long as only one of the three servers is still active. But how did all three servers suddenly crash?

The answer was made public in the press a few days later. A faulty software update was installed on all three servers at the same time. Because of this, the servers could not support each other because they were struggling with the same problem. It was only at around 8 p.m. of the same day that the software update was cleaned up and a large part of the network was restored to operation.

Since new software can sometimes react unexpectedly, it should never be installed on all critical points of a system at the same time. A step-by-step procedure is recommended, ideally following a test run in a test environment.

Sources:

- <https://www.thelocal.de/20090422/18791>
- <https://www.news-on-tour.de/13160/mobilfunk-t-mobile-netz-wieder-neu-hoch-gefahren-softwarefehler-im-sogenannten-home-location-register-hlr-gefunden/>
- <https://www.computerwoche.de/a/groesste-panne-im-t-mobile-handynetz-wird-untersucht,1893599>
- https://www.deutschlandfunk.de/kein-netz-nach-vier.676.de.html?dram:article_id=26367

11. Atlantic Lazarus

They could see his empathy, but it still felt wrong to the researchers.

A research group wanted to test the functionality of their functional magnetic resonance imaging (fMRI) device and was looking for objects with high contrast and different textures. After a pumpkin and a Cornish Game Hen did not meet the criteria, they finally tested a dead Atlantic salmon. Afterwards they presented pictures of social situations to the fish and recorded his reactions.

The data were meant for a practical course to demonstrate the analysis of fMRI data and possible error sources using an absurd example. However, the researchers were quite surprised that the device registered a response to the pictures in the brain of the dead salmon. An important step in the analysis of fMRI data is the correction for potential false-positive results. Without this correction, the measured values were wrongly interpreted as significant changes in the neuronal activity of the dead fish. At the time the study was published, a lot of publications regarding fMRI data were based on analyses without the necessary corrections which garnered a lot of attention.

The example shows that not only the calibration of instruments plays a role for the success of experiments, but also the appropriate analysis of the collected data including controls and corrections.

Sources:

- <https://teenspecies.github.io/pdfs/NeuralCorrelates.pdf>
- <https://blogs.scientificamerican.com/scicurious-brain/ignobel-prize-in-neuroscience-the-dead-salmon-study/>

12. Family Ties

Had she not insisted on sorting out the birthdays, the siblings would not have been upset.

In 2003, a researcher was responsible for coordinating the data collection among family members of the primary respondents to a study. Among those that they had to approach for participation in the survey were randomly selected siblings for which the primary respondents had provided contact information.

The siblings' contact information was stored in an Excel file in order to make stickers for envelopes and questionnaires that were sent out to them. The targeted siblings were informed that their [brother / sister] born on [dd.mm.yyy] had given the researchers their contact details. However, right before printing the stickers, the researcher in charge sorted the data by date of birth, or so they thought. In fact, they had only sorted the column with the birth dates and nothing else. The result was that thousands of people received a letter that their sibling with a completely wrong date of birth had registered them for the study which resulted in a lot of upset calls and emails. People thought they had a sibling they didn't know about or that their father had another family.

The researchers had to reconstruct and manually correct the data using the original paper questionnaires and then re-sent all the letters together with a letter of apology.

This example shows that master data files should never be overwritten and that extra care needs to be taken when handling personal information!

Sources:

- <https://www.lcrdm.nl/horror-family-stress>

13. Household matters

Busy Mills helpful friend Carlyle has to revive the French revolution twice because of a maids negligence.

John Stuart Mill, a friend of the historian Thomas Carlyle, found himself caught up in other projects and unable to meet the terms of a contract he had signed with his publisher for a history of the French Revolution. Mill proposed that Carlyle produce the work instead. Mill even sent his friend a library of books and other materials concerning the Revolution and by 1834 Carlyle was working furiously on the project. When he had completed the first volume, Carlyle sent his only complete manuscript to Mill. While in Mill's care the manuscript was destroyed, according to Mill by a careless household maid who mistook it for trash and used it as fire kindling. Carlyle then rewrote the entire manuscript, achieving what he described as a book that came "direct and flamingly from the heart."^[1]

This was not helped by his usual mode of working, which involved tearing his notes apart after he had finished using them for the intended chapter. He voluntarily destroyed the closest thing to a backup that was available at the time.

This story shows two things: first research data management and its mishaps are not a new topic and second whatever material form your manuscripts, it is vital to have more than just one copy of them in case something happens to the original.

Source:

- https://en.wikipedia.org/wiki/The_French_Revolution:_A_History

14. Wrong Direction

She not only cleaned up the basement, but demolished the whole house.

Tracy Teal was a student who studied computational linguistics as part of a master's degree in biology from the University of California at Los Angeles. She had spent months developing and implementing simulation software when she was finally ready to start her analysis. The first step before the analysis was to organize all important data and delete all unnecessary data. For the deletion process, she used the typical routine command "rm -rf *", which deletes all data in the current directory and in the subdirectories. The only problem was that she did not execute the command on the directory where the disposable data was located, but on the root directory of her project. Since this command, when executed in Unix systems, does not first put the files in the recycle bin, as it does in Windows or Macintosh, all project data was deleted in one fell swoop.

Tracy was lucky because an automated backup saved her work. To retrieve them, all she had to do was kindly ask her department's IT helpdesk whether they could restore her files. Tracy Teal is now Executive Director at The Carpentries, a non-profit organization that provides basic knowledge of coding and data science to researchers worldwide. Nevertheless, Tracy looks back at this situation in shame, because she herself had worked for the IT helpdesk before the accident. For her it was like "the lifeguard who needs to be rescued".

This story shows that even experienced scientists can make mistakes when dealing with data. A versioning or backup system should always be used for important data, so that complex data is not accidentally lost.

Source:

- <https://www.nature.com/articles/d41586-019-01040-w>

15. No good deed goes unpunished

Had he not picked it up, he wouldn't have been attacked.

Researchers of the University of Illinois performed an experiment to investigate if people are plugging USB devices that they find on the street into their computers.

The results of the study show that the majority of the sticks were picked up quickly and in many cases the data were accessed. A file on the sticks enabled the researchers to see which files on the devices were opened. Although curiosity was the reason to pick up the USB stick for many people, the majority wanted to find out the identity of the owner in order to return the device. However, the authors found that only a few people took safety measures before opening the files on the stick. It is easy to imagine that a method similar to the harmless attack by the researchers could be used by real criminals.

The example shows that it is important to be cautious when using unfamiliar external storage devices in order to prevent attacks on computers. In addition, the study shows that lost storage devices will often be searched by the finder and sensitive data can become accessible to non-authorized persons. Thus, external storage devices (e.g. USB stick, external hard drive) which contain sensitive data should always be password-protected and/or encrypted.

Source:

- <https://elie.net/publication/users-really-do-plug-in-usb-drives-they-find/>

16. Look, don't touch

Had she taken the saying "Look, don't touch!" more seriously, she would have spared herself the additional work.

A scientist saved experimental data on her computer. One day, she opened the original raw data in Microsoft Excel and the formatting of some columns was changed automatically. In doing so, the original values were permanently re-formatted and were lost for any further analyses. Since there was no additional copy of the raw data, the experiment had to be repeated.

The example shows that the long-time preservation of raw data requires particular attention. They should always be backed up and only be accessed on a read-only basis. The analyses should be done in a separate file, so that the original raw files are not accidentally altered.

Source:

- <https://www.nature.com/articles/d41586-019-01040-w>

17. Intervention from up-high

A fallen president killed his laptop.

An engineer's office was shaken by a small earthquake — hardly a surprise in California. A picture of former US president, and one-time client, Gerald Ford fell off the wall and hit his laptop, shattering the screen. After that, the man gave much more thought to what can happen to devices and data if such an event re-occurred.

The story exemplifies that disaster preparedness is part of efficient research data management. This can be accomplished by using well-functioning backup systems and storing several copies of important data on different mediums and in at least one other location.

Source:

- <https://www.nature.com/articles/d41586-019-01040-w>

18. Bad Recycling

The university saves resources, but now his quotations are a mess.

A researcher begins his career enthusiastically. For this purpose, he is provided with a new laptop and a new email address and shown the way to the coffee machine. He quickly got to work and soon he had published his first scientific publication! To see if his publication is already listed in search engines, he checked in Google Scholar whether he found his work there. In fact, the website had listed his publication ... unfortunately, however, this was not the only work he found there related to his "new" email address. What had happened?

The email address assigned to the researcher in question had previously been used by another scientist. After this person with the same first and last name changed employers, the address was "recycled". The existing email traffic was deleted and the address released for the new researcher. The university, or rather the responsible data center, probably did not expect publications and other services to be connected to this address. It was easy for the new researcher to get a new address, but it was very difficult to have his incorrect citations corrected on Google Scholar. On the other hand, the researcher was able to access all registered services with the recycled email address, which was a security risk for the previous user.

Although the blame was clearly not with the researcher himself, the story shows how one has to be careful when reusing addresses, because at first glance it is impossible to see what they are associated with. If it is unclear whether your own email address is unique, for important registrations you should always use a persistent identifier such as ORCID or ResearcherID, which is related to a single unique person. This has the further advantage that it remains unchanged, even if the person concerned moves or changes their name or affiliation. In addition, it should be good practice to document your own registrations and to switch the registered services to the new mail address after changing institutions to minimize your security risks.

Source:

- <https://www.lcrdm.nl/horror-address-recycled>

19. Antiques

Wanting to return to the beginning of her scientific career, she spent a lot of time at the flea market.

At the beginning of her career, Leslie Vosshall stored her data on floppy disks the contemporary local storage devices. After that, the data was neither accessed nor migrated to a modern memory format (CDs, DVDs, USB sticks, or external hard drives). The disks are still there and present and nicely labelled. Even if the data is still uncorrupted and complete today, her hardware for reading the disks simply no longer exists. So in order to get the appropriate hardware, she would have to be very lucky and find a device which is still functional: at a flea market, a storage room at the university or, in the worst case, at a museum.

The most important step to avoid this problem is to document the storage media used for past projects and continuously monitor the availability of complementary hardware. At the latest, when manufacturers announce that they will no longer install ports or drives in the newer generations of their devices, the data should be migrated to ensure bitstream preservation.

Upfront you can avoid trouble of this sort, if data is not stored exclusively on one medium, but, in keeping with the 3-2-1 rule, on several media types. The obsolescence of one type of memory device will then no longer directly result in an inability to access the data.

Source:

- <https://www.nature.com/articles/d41586-019-01040-w>

20. Don't judge a book by its cover

After she was informed about the determined identity, she was on the brink of tears.

A PhD student used a DNA sample for her experiments that she received from a colleague. Since she trusted it, she didn't verify the identity of the sample.

She worked on the sample for several months, but couldn't get meaningful results. Finally, the DNA sample was sequenced to see whether there was something wrong with the sample. It turned out that the sample was mixed up due to a mistake in labelling. Months of hard work had been completely in vain.

The example shows that correct and clear labelling is important in the handling of physical samples. If there is uncertainty regarding the identity of a sample, it should be checked before it is further used in experiments. In addition, samples and the corresponding information should be documented in a physical and/or digital inventory.

Source:

- <https://www.lcrdm.nl/horror-wrong-sample>

21. Unsatisfying boots

Only after he had the chance to recover, he was able to go into the mountains.

A researcher went on a field trip to Tajikistan to do a study in the Pamir mountains. When he arrived, he noticed that his laptop, where he stored all the materials he wanted to use on the field trip, was not properly booting up.

Since he did not have the material on an additional hard drive or as a printed version and the internet connection was too bad to get access to his online storage, the whole field trip was in danger. Luckily, he was able to find the only service center that had a contract with the company that produced his computer and the laptop was repaired.

The story shows the importance of backups, especially in the planning of field work. Since there might not be adequate opportunities for repair and purchase of equipment at the location, possible complications need to be considered at the onset. Online storage is only suitable as a backup option if a steady and high-quality connection to the internet can be ensured.

Source:

- <https://www.lcrdm.nl/horror-crash-on-field-trip>

22. Command line troubles

When she tried to erase the errors, things got really bad.

During the first year of her computational biology PhD, a young researcher was still getting used to working with the command line. The supercomputer that all analyses were run on generated two files from each analysis job – a ".o" standard output file and a ".e" standard error file.

While trying to clean up the thousands of error files that had accumulated, the researcher typed "rm *e*" into the command line and forgot the all-important dot. As a result, she lost every file that had an "e" in the file name. This included all of the evolutionary tree files she had spent months on, some of which took weeks to generate on the supercomputer.

However, the problems didn't stop there. To her horror, she discovered that she had not backed up many of those files. This small typo set her back weeks in her PhD.

The moral of the story is to always backup your work every day and if you do a data cleaning, make sure you are deleting the correct files.

Source:

- <https://www.lifehacker.com.au/2016/06/file-error-your-nightmare-data-loss-stories/>

23. Sharing is Caring

After sharing his files, his colleagues got mad at him.

A researcher worked in a team of six to write an encyclopedia of 800,000 words. After some deliberation, the team decided that all work should be shared on a common work area. There was already a server for this, which was used collaboratively for the project data by some members. When he put the result of his work on the shared memory, suddenly, the others were angry with him.

The problem was that both the researcher's own files and the other colleagues' existing files were given generic names. The absence of an explicit, content-based naming convention meant that many existing files were overwritten when the local files were copied to the shared memory. The researcher did not worry too much about it since the old files should have been safely restored by a backup - only, unfortunately, this was more than 1 month ago and the recent backup had only covered the new files due to bad timing.

Obviously, the research group did not place much importance on using a versioning system or a more thorough backup system. What is even more interesting here is the organizational structure, which apparently was never properly determined. Thus, all folders and files kept their default names that were created by the respective working environment (such as "index.html" on the web, "main.tex" in latex or simply "document.docx" in Microsoft Word) instead of well-defined names with the date, subject or author being part of the file name. This helps files to be recognized and retrieved more easily and reduces the risk of overwriting other files when copying.

Source:

- personal communication

24. Questions of Calendar

1908: Late Russians do not shoot.

In the year 1908, the Olympic Games lasted 6 whole months. And yet, the Russian team managed not to be present for the shooting competition on July 11th. The problem was that they were still using the Julian calendar instead of the Georgian Calendar like the rest of Europe. This only changed with the Russian revolution in 1917. When the Russian delegation finally arrived, a month and a half of the Games was already over. They were able to compete in some other disciplines and win some medals. This, however, was a small comfort for the shooting team.

The 1908 Olympics, however, were not only remarkable because of this calendric confusion. Documentation left much to be desired, as well. For instance, it's still debatable whether Turkey was represented at the games at all!

The story shows that establishing agreed conventions and standards is vital - be it for the smooth implementation of international competitions or in science. Without a clear definition of the standards used, data from third parties might not be interpreted or replicated correctly, if laboratory notebooks or memos are misunderstood.

Source:

- <https://www.rbth.com/history/331074-russia-late-for-olympics-1908>

25. On fire

If the day had been a bit warmer, he wouldn't have been in so much trouble.

A box containing original documents was standing in a researcher's living room. On a particularly cold day, his daughter's babysitter looked for material to light a fire. She mistook the papers for fire kindling since she assumed that the filled-in questionnaires could only be scratch paper.

This story shows that storing documents containing personal or sensitive data in locked furniture within locked rooms is crucial for data protection, but also to preserve the documents as such. While the damage from the document destruction would have been significantly reduced if there had been digital copies, this is not always possible for all documents. Thus, particular care must be taken with irretrievable materials.

Source:

- personal communication

26. If it ain't broke, don't fix it

Since they no longer spoke the old languages, no money could be distributed to the people in need.

In early 2020, the COVID-19 disease, caused by the coronavirus SARS-CoV2, broke out globally, which led to the closure of many shops and businesses for quarantine reasons. The result, especially in the USA, was a large number of unemployed people who urgently needed money for their next rent payment, food or other expenses. As a consequence, the government decided to set up a relief package for anyone who registers as unemployed - but why didn't the money get to the people?

The reason for this was the overload of critical systems on which COBOL is still running. COBOL is a programming language that was developed in the late 1950s to control commercial applications. From today's perspective, the programming language is very outdated and no longer taught in the training of programmers. That is why there was no personnel to take care of the systems when they collapsed. Unfortunately, many applications with the outdated programming language are still running in the business sector. In order to solve this urgent problem, the Trump administration is now desperately attempting to get previous COBOL programmers to come out of "retirement".

This example shows that even if a system (supposedly) runs well, it might, especially in the IT area, not be a horse you should be backing. Innovation and evolution are important in computer science. Existing systems should be questioned, since requirements can change and established habits can lead to problems from today's perspective. For example, at some point data might no longer be able to be called up or might exist in formats that are increasingly difficult to be processed.

Sources:

- https://youtu.be/PpV_5-tCS-c?t=310
- <https://www.theverge.com/2020/4/14/21219561/coronavirus-pandemic-unemployment-systems-cobol-legacy-software-infrastructure>
- <https://fortune.com/2020/04/15/how-to-get-unemployment-benefits-coronavirus-extra-600-dollars/>
- <https://www.datacenter-insider.de/cobol-eine-programmiersprache-wird-uns-alle-ueberleben-a-865219/>
- <https://www.linkedin.com/pulse/never-change-running-system-warum-diese-weisheit-im-zeitalter-welsch-1e/>

27. Trouble down the line

Even though he found the line in the inventory, all his work was for nothing in the end.

During his PhD, a researcher needed to use a certain cell. He searched for the cell line in an inventory shared with the whole department and started growing the cells. He performed an expensive mass spectrometry experiment using the cell line, just to find out afterwards that the cell line was not the one it was supposed to be due to mislabeling. Since his PhD contract was about to end, there unfortunately was no time to solve the problem.

This could be avoided by correctly labelling samples and verifying the sample before using it for any experiments. The management of physical samples, i.e. their verification, correct labeling and entry into a physical and digital inventory is an essential part of research data management and it is very important that these steps are done properly and documented well.

Source:

- <https://www.lcrdm.nl/horror-mislabelling>

28. Expensive Mice

Transport stress made the planned experiment not only expensive, but unworkable.

A doctoral student in Germany ordered mice from a certain KnockOut series directly from the company in the USA that held the patent. In KnockOut mice, one or more genes are specifically deactivated by genetic manipulation in order to be able to investigate the biological mechanisms regulated by them. In addition, such animals are suitable as a model for human diseases and for pharmacological issues. Because they were ordered in the US the transport costs of the mice (several thousand Euros for animal transport by air) was a large multiple of their value (approx. 3.50 Euros per mouse). To add insult to injury the animals were useless for the planned stress experiment. The transport had already inflicted too much stress on them. A later order from a European licensee for the patent was possible without any problems, but all this cost the doctoral student valuable time and her institute unnecessary expenses. Probably from the point of view of most employees in the laboratory it was "general knowledge" that not only patent holders can deliver the right mice, but also licensees in Europe. This seemed so self-evident that there was no reason to talk about it and the doctoral student would have needed to inquire actively to get this information.

The documents were checked to ensure that the correct genotype mouse was ordered. The rather mundane question of where the mice came from was probably not on the supervisor's priority list, considering the order was confirmed without the choice of supplier being questioned.

This story shows how important it is to write implicit knowledge down. Had there been a list of possible suppliers of lab animals of different gene lines, the doctoral student would surely have noticed that the US company is only one of several eligible sources. If not only the individual suppliers but also the selection criteria had been recorded, an informed, non-accidental decision could have been made on the source of supply. Such a systematic approach to selection is useful not only for lab animals, but also for software, hardware, measuring instruments and consumables.

Source:

- personal communication

29. There is always an option

They didn't think of all their options when it was time to submit.

A group of research assistants worked on a project that was close to the deadline. So that everyone was able to access to the project files, all data was stored on a Microsoft SharePoint server, which displayed the current status of the documents and allowed collaborative work. In the final phase, the report for the project needed to be finished and all old and unnecessary files were to be deleted. One of the team members deleted the report, because he thought it was an old draft. However, this was not supposed to be a problem since SharePoint versions files in the background and the file can easily be restored. Why, then, didn't it go as planned?

It was the settings. Even though SharePoint is one of many systems that can version and restore project files, this is not specified in the factory settings of the software, because this requires more storage space. Only via clicking through some menus, you can enable file versioning and determine backup intervals. However, the team had not done so and the report was lost forever and had to be completely rewritten in a very short amount of time.

This story shows that the functionality of various backup, cloud and other software solutions for project data should not be taken for granted. It is important to find out beforehand how data backup works and, best of all, to test it on the application. Many functions first have to be activated or adapted to the needs of the project.

Source:

- personal communication

30. Sunken treasures

Sandy's stormy temper was not helping in the efforts to understand the little pests.

Leslie Vosshall, a neurobiologist at the Rockefeller University in New York, stored her research data on a server in the basement of her home. During hurricane Sandy in 2012, her basement was flooded and she almost lost all of her data belonging to a mosquito genome project.

The story shows that unexpected events and catastrophes like hurricanes or fires can damage data storages at a location. Specialized buildings (e.g. IT centers) are generally better protected against such events than residential buildings and are therefore preferable for data storage. In general, data should be stored in accordance with the 3-2-1 rule: 3 copies on two different storage media and one copy should be kept offsite. This reduces the risk of the destruction of all copies due to disasters at one location.

Source:

- <https://www.nature.com/articles/d41586-019-01040-w>

31. Undocumented

The data existed, but could not be re-used despite their best efforts.

When a young researcher started on his PhD, he was told to work on unpublished data that had been collected 3 years prior. He received various folders that were full of data. After going through them, he found that there were several datasheets with duplicate names but different contents, scripts that nobody knew what they did or why and column names that were unclear and ambiguous. Moreover, the exact equipment and/or settings used for the experiments were unknown in some cases. Since it had been several years, not even extensive talks with the manufacturers of the used equipment or the data authors could make the data usable. In the end, the data were not able to be re-used.

This shows how essential describing and documenting the data collection and analysis process really is for data re-use. Although it takes quite some time to document data, it takes even more time and frustration to try and figure out poorly documented data from years ago. Even though many people think they know their data, it is very likely that they will forget almost all details in a matter of a few years. Therefore, documentation should always be as concise, detailed, precise and easy to understand for third parties as possible.

Source:

- <https://www.lcrdm.nl/horror-lack-of-documentation>

32. It's the small things that matter

If his name had been easier to spell, his accomplishments would have been recognized more.

Small changes in the spelling of names can have severe consequences for the researchers' careers. Many alternative versions can be found especially for names that contain special characters. This can lead to a systematically lower recognition of citations.

The researcher Terje Tüür-Fröhlich shows this effect in her work using the well-known sociologist Pierre Bourdieu as an example. She found 85 mutations for his name in public research databases. A flawed record of citations can be an issue in particular in the beginning of an academic career since citation indexes are used as a measure for scientific accomplishments and play a role in the allocation of funding or job applications.

The use of persistent identifiers (PIDs) for persons represents a possible solution for these issues. Since authors are identified by the PID, mistakes in the spelling of the name are less problematic. Common PIDs for researchers are the non-profit scheme ORCID or ResearcherID by Thompson Reuters. By now, PIDs for persons have become somewhat common and a number of publishers ask for them when a manuscript is submitted.

Sources:

- <https://www.heise.de/tp/features/Auch-Pierre-Bourdieu-ist-ein-Indexierungsopfer-3727711.html>
- https://lisa.gerda-henkel-stiftung.de/fehler_in_zitationsdatenbanken_sind_nicht_zufaellig_verteilt?nav_id=7314

33. Employee with an invisibility cloak

He just seemed to keep slipping through the fingers of the human resources department.

Although the HR department made several entries for Steve Null, he repeatedly disappeared from the database. The system took "Null" (german for zero) literally, and interpreted the entry as a missing date. For the database Steve did not exist. His non-existence was proclaimed by his name. Before processing the request for Steve, the system first checked whether any data had been entered at all. Modern systems prevent requests sent without content, which happens quite a lot, from unnecessarily burdening the system. Unfortunately, the side effect of this "search_term!= NULL" is that people named Null cannot be found in such systems, even though the corresponding entry exists. The search is simply aborted too early.

This story shows that it may be useful to take a close look at the limits of the database system used.[1] Are there rules within the system that cause certain entries to be systematically not found or interpreted in a way that is not intended? In such a case, a look at the (hopefully existing) documentation of the software or in the case of transferred data at conventions used for missing values and the like helps. In addition, it is always useful to create appropriate documentation for all self-generated data, so that what was taken for granted yesterday, can still be traced in the future.

[1] For any programmers thinking there is no way this is still a problem, Parker 2020 p. 257 refers to an XMLencoder problem in Apache Flex. Check out bug report FLEX-33644.

Source:

- Matt Parker (2020): Humble Pi - When Math Goes Wrong in the Real World, p. 259.

34. False Alarm

Burn it did not – but the data were gone nevertheless.

The head engineer at a data recovery firm once lost all of his personal possessions in a wildfire, so it was a little ironic what happened to one of his clients. The client had stored a rack of 96 hard disks underneath a fire-control sprinkler. One day, the sprinkler went off, probably due to a false alarm, and the disks were inundated with water. Most of that data was gone, since it had not been backed up.

This story illustrates that you should always store your data according to the 3-2-1 rule (at least 3 copies on at least 2 different storage media and 1 of those in a different location. Also, regular and reliable backup is key.

Source:

- <https://www.nature.com/articles/d41586-019-01040-w>

35. In the enemy base

When the agent infiltrated the base, they responded with hardware disposal.

In 2008, the US military decided to dispose of all removable USB storage devices at all military bases and to stop using USB devices. What led to this decision?

When a USB stick was discovered and analyzed during a Middle East mission, it later turned out that it contained malware from a foreign secret service organization. The data with the malicious code went unnoticed into the internal network of the US military. Ironically, the malicious malware program was called "Agent.btz". When the program and the vulnerability were discovered, the Pentagon decided to immediately dispose of all removable devices that run via the USB interface. To date, this has been one of the largest security breaches in US military history. However, devices that use the USB port remain a security risk to this day. Because of their universal purpose, they can quickly gain control of a system (for example, by pretending to be a keyboard) or just inject dangerous data.

The moral to be learned from this story is that it is particularly important for research institutions and companies to never simply connect third-party USB devices to critical systems. An additional hardware adapter, virtual work environment or at least a virus scanner should always be switched on as a software solution in order to detect malware on a USB device at an early stage.

Source:

- <https://www.computerworld.com/article/2514879/infected-usb-drive-blamed-for--08-military-cyber-breach.html>
- https://www.vice.com/en_us/article/7xy5ky/the-american-military-sucks-at-cybersecurity

36. Lost in Translation

Since they couldn't agree on common standards, he strayed from the right path.

The Mars Climate Orbiter (MCO) was part of a NASA program to gather information on Mars. On December 11, 1998 the MCO was launched. The aim was for MCO to circuit Mars in a spherical orbit and perform measurements regarding the atmosphere and the climate of the planet. However, there was a problem during the maneuver and the space probe came too close to Mars and was lost.

The navigation problems resulted from the use of different units for calculations by the involved institutions. While the navigations team used the metric system, Lockheed Martin Astronautics, the American company that had produced the probe, used Anglo-American units of measurement. The conversion of the units (e.g. Newton-seconds vs. pound-seconds) was not always taken into account, thus leading to errors in course corrections.

In fact, NASA had made it clear in its "software interface specification" that the metric system should be used. The course correction program SM_FORCES by Lockheed Martin Aeronautics was not written in accordance with the official specifications and caused the loss of the spacecraft.

The story shows that the use of well-established standards is a major pre-requisite for successful projects. This is especially important in projects with international partners from countries that use a system of measurement that is not based on the International System of Units (SI units), e.g. USA. The use of standards is also important for the comparability and traceability of projects. The use of a double-check system also helps to ensure consistent implementation of specifications.

Source:

- <https://www.simscale.com/blog/2017/12/nasa-mars-climate-orbiter-metric/>
- https://de.wikipedia.org/wiki/Mars_Climate_Orbiter#Verlust

37. Null Island

Statistically, the area around the police station was the most dangerous place of all.

The online crime map of the Los Angeles Police Department showed that between October 2008 and March 2009, over 1,380 entries came from the area surrounding the police station itself. That accounts for almost 4% of all recorded crimes in this city during this period. It was only when the LA Times (which base is also in the neighbourhood) complained that the police station noticed the error in the system. But what had happened?

All police reports were written by hand at the time and the majority of them were automatically entered into the database. It also happened often that the location of the crime was not recognized. In this case, the location of the police station itself was simply taken as the default value. This was not checked, which led to a major falsification of crime statistics. The LAPD corrected the error by setting missing location information to a "null" value (information for missing value in computer science). Of course, null values can also render certain parts of data records unusable if the values have to be used for certain visualizations or calculations. One, therefore, speaks of "Null Island - where bad data goes to die".

From this story one can learn how important it is to correctly determine the attributes of tables and databases, especially if they might also have missing values. If you simply set a value that appears to be logically readable for machines (such as "Null" as comment text or (0.0,0.0) as location), the data is not interpreted correctly and can falsify the subsequent results. You should always ensure that all possible values of a data set are well documented and that programs can recognize exceptional cases if necessary.

Source:

- "When Good Data Turns Bad" from the book "Humble Pi: A Comedy of Maths Errors" p. 253

38. In deep "shit"

A less special feed would have allowed a more representative census among the beetles.

Because they are cheap and ubiquitous, biologists have used human faces to take stock of fecal insects. Since some species find this food significantly more attractive than others, this procedure might have distorted the monitoring of local biodiversity, as studies by the Oxford zoologist Elizabeth Raine show. Alternative methods are currently being evaluated. The previous standard might be suboptimal and the concerted development of a new standard is the next logical step.

History shows that traditional approaches should be questioned in terms of their impact on the results and that it can be important to develop a measurement procedure that pinpoints better what actually needs to be counted. At the same time, it becomes clear that the use of standards facilitates uniform error corrections and a consistent introduction of new procedures.

Source:

- <https://www.economist.com/science-and-technology/2020/01/09/dung-beetles-prefer-human-faeces-to-those-of-wild-animals>

39. Easy come, easy go

If the visitor had come a bit later, he wouldn't have lost his points.

A researcher performed a lengthy experiment regarding the properties of plasma in a Plasma Enhanced Chemical Vapour Deposition reactor. He hadn't saved the most recent data of the measurement as a theoretician of the project came to visit the lab and to discuss his models. The visitor pressed a key on the keyboard and deleted the latest test measurement data by accident. Fortunately, only one measurement point of the measurement series was lost so that they were still usable for the study. The lost data, however, could not be restored.

The story shows that data should be saved and backed-up as soon as possible to prevent data loss. In addition, distractions should be avoided during the documentation and saving of data to prevent errors.

Source:

- <https://www.lcrdm.nl/horror-erroneous-keyboard-key>

40. Return to sender

Without sorting, delivery would have been much easier.

The Association of German Librarians (VDB) needed three attempts in 2019 to deliver its yearbook to its members. Names had been assigned to wrong addresses (a typical error if only one column in an Excel cell is sorted and the original line alignment is broken), so that the delivery of the books generated many returns. The fact that the delivery took place over the Christmas holidays made communication with VDB-members more difficult. During holidays, both those responsible for the association and the members had other priorities. Even after the members were informed that they should accept the books sent, even though they were not the correct addressees, not all yearbooks could be delivered in the first round. In the second attempt, a programming error occurred in the reorder form. Those who did not provide a membership number could not be served and had to reorder their volume in February a second time. Overall, the accumulation of shipping problems likely caused irritation among members and considerable costs for the VDB.

The story clearly shows that special care is required when sending out mail. In any case, address data should be saved in read-only form, backed up several times and checked for integrity before dispatch. Software that is used in “customer” contact should be thoroughly tested for functionality before it is used. In this way, errors can be avoided internally and do not have to be ironed out live.

Source:

- personal communication

41. The end is nigh

How Excel almost sabotaged Wiki-Leaks once.

When Wiki-Leaks founder Julian Assange handed a file with over 92.000 field reports from the Afghanistan war over to journalists from The New York Times and The Guardian in 2010, the data abruptly ended in April 2009 even though there should have been data from the entire rest of the year. What had happened?

The journalists had opened the data in Excel which, at that point, had a size limitation of 65.536 lines so that the file exceeded the spreadsheet's maximum capacity. All data after line 65.536 had simply been cut off.

Even though the maximum size has been increased to 1.048.576 lines since then, this story still neatly illustrates that Excel is not an adequate substitute for a professional database. This is especially true for research projects that expect to generate a large amount of data. In such cases, it is very important to consider suitable alternatives and associated costs early on.

Source:

- "When Good Data Turns Bad" from "Humble Pi: A Comedy of Maths Errors", pages 244-245

42. Too good to be true

The materials were too good to be true, which led to a great disgrace.

Beginning in 1998, Bell Laboratories staff published a number of notable articles in a short sequence on the discovery of new carbon-based materials. However, other scientists from the discipline failed to replicate the results.

Despite the strong interest, it took another 3 years until other researchers noticed that the numbers in the publications were unusual. Some graphics were simply too "beautiful" to represent real world systems. Ironically, a young German scientist named Jan Hendrik Schön co-authored all the dubious publications and was involved in the work on them. An independent committee of experts was set up. Shockingly, it concluded that in at least 16 out of 25 cases the data underlying the publications had never existed. Schön's explanations that he deleted the primary data due to lack of space and that used storage media no longer functioned or were thrown away seemed more than doubtful to the committee.

A solid commitment to open data would have revealed the fraud much faster. Ideally, it would have been impossible from the outset. On the institutional level at least the case had consequences: it led Bell Laboratories to introduce new data retention guidelines, co-author responsibility, and review primary data prior to publication.

Sources:

- On Being a Scientist. A Guide to Responsible Conduct in Research: Third Edition (2009), <https://doi.org/10.17226/12192>

43. A cat's a cat, and that's that.

If they had had a dog, the home office might have been a safer place.

Due to Corona, a scientist had to work in the home office. This was unproblematic, as he was able to organize and process all work steps and documents from home. Nevertheless, he hadn't expected one thing that would cause him problems for the next few days.

At home he kept several cats and they had previously shown little interest in all the cables that were part of the technical equipment of the home office. But this was to change one night. When the scientist got up one morning, the power cable of the work laptop was suddenly gnawed through. Luckily, he still had a spare laptop, and since the work data was continuously backed up via the university cloud service, he was able to quickly set up his virtual workstation again and keep all his appointments.

This story shows that (true to Murphy's adage that "Anything that can go wrong will go wrong") you should be prepared for any problems. Even if the research data on your laptop seems safe, there should always be a backup to make the data recoverable in case of a problem. In addition to replacement hardware, cloud solutions that can quickly synchronize files on different devices are particularly recommended.

Sources:

- personal communication

44. Surveying Nirvana

Less confidence in professional tools would have saved a lot of work.

One researcher was extremely annoyed when she noticed that the commercial survey tool had not saved her last version of her questionnaire. However, both the company's in-house support and customer service were unable to provide a satisfactory explanation for the incident. The unsaved version had to be recreated. Thanks to the researcher's existing local backup this was possible with reasonable effort.

Even if commercial products offer an integrated backup, it is necessary to test its functionality at regular intervals. Running your own automated backup regime can also increase system redundancy and reduce the likelihood of data loss. The connection between the online tool and the server should not be interrupted to ensure data transfer. Common reasons for interruptions are application errors (e.g. session time out) or internet connection problems. An alternative to working online can be working locally before uploading content.

Sources:

- personal communication

45. Finding a good match

She excelled and slipped right through the grid.

An expert in artificial intelligence (AI) applied for a position in this field. Because the required internships did not appear in her CV and the AI did not accept the existing work experience abroad instead, she was not invited for an interview. She asked the company why they had decided not to invite her. This demonstrated both her commitment and her understanding of how artificial intelligence works. Her suggestion that her CV did not conform to the "successful" CV pattern in the company was not far from the truth. She had not been singled out because she had an unusual and feminine first name, as she first suspected. The issue was an 'objective' criterion, the requirement for traineeships. The software simply did not recognize equivalent or higher-quality alternatives.

Since artificial intelligence makes pattern-based decisions, two points are essential for successful use: a suitable training data set must be available and the statements must be checked with test data in order to detect and correct undesirable effects in later use as early as possible.

Sources:

- <https://www.spiegel.de/wissenschaft/technik/ki-forscherin-ueber-algorithmen-sind-wir-menschen-wirklich-so-simpel-a-00000000-0002-0001-0000-000172493030>

46. Not only set in stone

Thanks to his predecessors, he was able to solve the puzzle.

During Napoleon's expedition to Egypt, the French officer Pierre François Xavier Bouchard discovered the Rosetta Stone in the Nile River Delta in 1799. It is the fragment of a high stele on which a decree is carved in three different languages (ancient Greek, demotic script and Egyptian hieroglyphics). Immediately after the discovery, French scientists made numerous transcripts of the inscriptions on site. After the defeat by the British, the Rosetta Stone fell into British possession and the researcher Thomas Young began to study the texts. Fortunately, the French had made copies, so that in 1822 Jean-François Champollion succeeded in deciphering the demotic script and hieroglyphics using the ancient Greek language he knew. After his discovery was published, further hieroglyphics were deciphered and the basis for modern Egyptology was laid.

The story shows that, thanks to the transcripts, two ideas of research data management were implemented. On the one hand, there were copies in case the original was lost (backup) and, on the other hand, other scientists were given access to the texts and conduct research on them (open data).

Sources:

- <https://www.youtube.com/watch?v=TDnuTzAyCss>
- https://de.wikipedia.org/wiki/Stein_von_Rosette

47. The debt brake stands on feet of clay

A more conventional weighting would have done less damage to their reputation.

Researchers Kenneth Rogoff and Carmen Reinhart of Harvard University postulated in 2010 that exceeding the national debt by 90% of economic output has a negative impact on a state's economic growth. This basic assumption is both the basis of the German debt brake and the austerity requirements of the Euro rescue policy. However, an attempt to replicate the result based on the underlying data did not succeed. In fact, the 2010 study did not include data for specific years, weighted some cases abnormally high, and inadvertently failed to include several countries.

The media echo that followed the failed replication significantly damaged the reputation of the two economists. Nevertheless, in the replication study as well as in a follow-up study of the Harvard researchers, the fundamental connection remained, although the reduction in growth rates was less severe than originally calculated.

This example shows that individual decisions in the course of data analysis must be well documented and mentioned in the associated publications in order not to raise the suspicion of data manipulation in favor of particularly spectacular or significant results. Pre-registration of studies in a journal is also a good step towards having research questions, research design and implementation independently assessed.

Sources:

- <https://www.spiegel.de/wirtschaft/panne-mit-excel-tabelle-rogoeff-und-reinhart-haben-sich-verrechnet-a-894893.html>
- <https://www.spiegel.de/wirtschaft/soziales/excel-panne-von-kenneth-rogoeff-das-war-ein-massaker-a-929248.html>

48. The small difference

On this magnetic track, the system started to lurch.

In the spring of 2020, it emerged that hardware vendors such as Western Digital had sold unmarked HDD hard disks in SMR (Shingled Magnetic Recording) format instead of conventional hard disks in CMR (Conventional Magnetic Recording) format in various backup systems. This subtype of HDD hard drives saves space by overlapping magnetic tracks (similar to roof shingles) instead of just laying side by side, but the read and write speed is slower. As a result, backup systems simply failed, because they could not perform their data backup routines according to schedule.

History shows that using the wrong or outdated hardware can lead to errors. Even if in this example it was not the researchers' fault, it should still be ensured that, as the amount of data increases (keyword: Big Data), the right conditions for working with the appropriate software are in place. The hardware should always be tested before it is put into routine use.

Sources:

- <https://arstechnica.com/gadgets/2020/04/caveat-emptor-smr-disks-are-being-submarined-into-unexpected-channels/>
- <https://arstechnica.com/gadgets/2020/04/caveat-emptor-smr-disks-are-being-submarined-into-unexpected-channels/>
- https://de.wikipedia.org/wiki/Shingled_Magnetic_Recording

49. Clone Wars

More openness about the ancestry would have saved a lot of effort and money.

At the beginning of the 2000s, stem cell research was the great hope for the development of new therapies. When, in 2004 and 2005, a relatively unknown Korean laboratory under the leadership of Woo-Suk Hwang announced in two publications in the prestigious journal *Science* that it had extracted a total of 11 stem cell lines from cloned human embryos, it was a sensation. At first, a new era of stem cell research seemed to be imminent, and researchers around the world were trying to replicate the Hwang method. Unfortunately, to no avail.

A good year after the publication of the second paper, inconsistencies and striking similarities were noticed in the articles' illustrations. A commission was set up to analyze the primary data. The results prompted it to conduct tests on DNA samples. It turned out that none of the stem cell lines came from cloned embryos. All the information and representations in the publications were invented. By then, millions of grant money around the world had flown into replicating the results - for nothing.

If the journal in which the articles were published insisted that Hwang and his co-authors provide the primary data as a supplement, the fraud attempt would most likely have flared up immediately. Insistence on open research data could have created transparency and prevented the waste of funding. Following the well-publicized case, major journals changed their publication policies and research funding organizations began to pay greater attention to openly accessible research data.

Sources:

- On Being a Scientist. A Guide to Responsible Conduct in Research: Third Edition (2009), <https://doi.org/10.17226/12192>

50. Promising developer personality

As the captain of the women's chess club, she was in a bad position to be selected.

Amazon developed a computer program from 2014 onwards to evaluate the CVs of applicants. The aim was to automate the search for suitable candidates for any vacant positions. The program used artificial intelligence and rated the suitability of the application with one to five stars. However, in 2015 the company found that the program did not make gender-neutral selections for applications for software development or other technical positions. The online recruiting program simply didn't like women.

This was due to the training of the computer model, which was fed with CVs of applicants from the last 10 years. Most of the applications came from men, a reflection of male dominance across the tech industry. The system taught itself that male applicants should be given preference and downgraded applicants if the word "woman" appeared in their resumes, for example "captain of the women's chess club". After the case became known, Amazon claimed that the program was never used by recruiters to evaluate candidates. However, insiders said that the AI-based recommendation system had been used, but Amazon's recruiters never relied solely on those rankings.

The example shows that data quality is of crucial importance in machine learning. According to the motto "Garbage In, Garbage Out", an algorithm can only be as good as the data set it is fed for training by humans.

Sources:

- <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G>

51. Burning Cloud

Before they knew it their game burst into flames.

On March 10th 2021, 3.6 million websites were suddenly offline, including the French government portal – what had happened? For as of yet unknown reasons one of cloud provider OHV's 5-story server buildings burned down completely while a second building was severely damaged. The incident led to the destruction of 12.000 servers and to the sites hosted on the servers becoming unavailable.

Beyond the servers, the company's hosted private cloud service, which had stored data from many big private corporations had also fallen victim to the destruction. Shortly after the blaze, the video game company FacePunch was forced to announce that all its data for the popular online game "Rust" had been lost, since FacePunch had no local backup.

This story illustrates that cloud solutions can only be one part of an effective backup strategy. Following the 3-2-1 rule, at least two other local copies should always be made to avoid complete data loss like in this ScaryTale.

Sources:

- <https://www.faz.net/aktuell/feuilleton/medien/groesstes-rechenzentrum-europas-brennt-komplett-nieder-17241629.html>

52. The table is full

Excel shuts down and Corona has to wait.

In Great Britain, Corona case numbers were underreported for several days due to a data mishap that occurred when using the Microsoft Excel program. An overfilled Excel sheet is said to have been the reason for the loss of the test data. How did this happen?

In Great Britain, the "test and trace" unit of the NHS (National Health Service) is responsible for sending the data for those who test positive for Covid-19 to the responsible health authority PHE (Public Health England). As the British Guardian reported, the authority had compiled and managed the case numbers in Excel spreadsheets since the beginning of the pandemic.

Excel can only be used as a database to a limited extent, as the number of available rows in the software is limited. In the current Excel version, up to 1,048,576 rows can be entered. In this case, the limit was only 65,000 rows due to the use of an outdated Excel format. As a result, when importing the CSV file sent by the NHS with the 16,000 Corona cases submitted, the excess rows were deleted for a total of 15,841 cases. This resulted in the UK corona statistics looking better for a few days as well as infection chains not being followed up on.

The example shows that when importing data, it is important to ensure that both the data format and the software can handle the type and scope of the data correctly. If possible, the latest version of software and open formats should always be used.

Sources:

- https://www.t-online.de/digital/id_88701962/england-peinliche-excel-panne-unterschlaegt-fast-16-000-infizierte.html
- <https://www.spiegel.de/netzwelt/web/corona-in-grossbritannien-excel-panne-behoerde-verschlampt-tausende-positive-tests-a-685a1d0a-7022-4dc2-82a7-1150fb0ec85e>
- <https://www.faz.net/aktuell/feuilleton/pandemie-datenpanne-in-england-die-tabelle-ist-voll-16989088.html>
- <https://www.bbc.com/news/technology-54423988>

53. Technical Revolution

Russian red guards receive militant assistance from the future.

During a history exam on the Russian Revolution of 1917, the painting "Storming the Winter Palace on October 25, 1917" by Nikolai Kochergin was supposed to be shown as a visual aid for interpretation. However, the examiners probably relied somewhat inattentively on Google Image Search and did not integrate the original image, but an edited image of the revolution that had been ranked highly by hits. In this alteration of the image, the revolutionaries were assisted by a large battle robot in storming the palace.

This probably rather irritating or even amusing faulty image selection, probably did not prevent the students from passing their history exam, but it demonstrates clearly the importance of ensuring the provenance of the source to ensure authenticity when working with third-party data. In the case of image sources in particular, it is also important to pay attention to licensing regulations, otherwise things could get expensive as well as embarrassing.

Sources:

- <https://www.smh.com.au/education/history-transformed-in-vce-exam-20121114-29ce7.html>
- <https://9gag.com/gag/a5W3W4r>

54. Chatty confetti shower

At the carnival parade even the confetti was talkative.

At a carnival parade in 2016 in a small town in Thuringia, not only small colorful round snippets fell out of the confetti cannon, but also snippets that were inscribed with sensitive personal information and thus spoke for themselves.

A hospital had used the carnival to improperly dispose of its old patient files, and the shredded files were scattered on the street. Unfortunately, the pieces of confetti were still big enough that names, addresses and telephone numbers were partially readable. The incident caused an outcry and drew the attention of the Thuringian data protection commissioner who issued an administrative order and a fine.

This scary tale illustrates the importance of deleting data properly. Personal data must be permanently erased when the purpose for which it was processed has ceased to exist and there are no retention periods. In this context, deletion is understood as the complete rendering of this data unrecognizable.

Sources:

- <https://www.datenschutz-notizen.de/konfettiregen-aus-dem-krankenhaus-recycling-von-patientenakten-0413884/>
- <https://www.welt.de/regionales/thueringen/article152088666/Untersuchungen-zu-Konfetti-aus-Patientenakten-dauern-an.html>

55. Who am I?

Is the person wearing inconspicuous glasses and a trench coat?

A public institution wanted to measure the health of its employees and designed a detailed survey with questions about their experience at work, mental load and satisfaction. The survey participants were guaranteed anonymity. Many employees took advantage of the offer and experienced an unpleasant surprise when their supervisor asked them about specific answers they had given in the survey. What happened?

The survey did not ask for names, but it did ask for a lot of detailed information and demographic data. Participants' responses were then analyzed and reported in such small groups/organizational units that derivable characteristics such as gender and job function made it possible to accurately match individuals to their responses. Thus, the promised anonymity was not maintained and many complaints were made to the responsible data protection officer, not to mention awkward conversations with the respective managers.

Particularly when evaluating small samples, attention must be paid to "derivable" personal data. Even if the name is not recorded, a combination of different data can make a person identifiable to others. In this case, the survey is no longer anonymous and the EU Data Protection Regulation must be observed - which requires, among other things, different declarations of consent and processing steps/technical and organizational measures than anonymous data.

Sources:

- personal communication

56. Blurred Man on the Moon

This incident left a lot of room for interpretation

In 2006 the Goddard Center's Data Evaluation Laboratory was scheduled to close. It was the last institute with equipment still able to read the original moon landing recordings in SSTV format. The closure prompted a search for the original tapes in NASA's archives. However, as it turned out in 2009, these had either disappeared or even been dubbed over by the early 1980s, as was common practice at the time. Only a few other photos or videos could be found. The modern known recordings of the moon landing mostly originate from the conversion into the improved NTSC format for television sets. However, in the conversion from SSTV they simply filmed a screen in most cases, which resulted in strong losses in contrast, brightness and resolution. The fact that the original tapes have not been found to this day, caused a wave of conspiracy theories.

This story illustrates the importance of properly archiving meaningful research data. Characteristics of archiving are complete transfer, immutability, access restriction, and documentation of access and relocation of the data. If necessary, the data must be transferred to new media if the old ones have become obsolete.

Sources:

- https://en.wikipedia.org/wiki/Apollo_11_missing_tapes
- <https://www.zeit.de/online/2009/30/mondlandung-aufnahmen-verschwunden>
- https://de.wikipedia.org/wiki/Verschw%C3%B6rungstheorien_zur_Mondlandung#Verschwundene_Filmaufnahmen

57. Binding Contracts

With the concluding list they simply promised too much

In this example, informed consent for a study prepared with model declarations used wording that enumerated the permitted uses and promised restriction to those uses. Specifically, the text contained the statement "data will be processed exclusively in previously described ways." Since the (anonymized) publication was missing from the list, the data could not be published. The very strict wording was binding for both parties.

It makes sense to have all legally binding texts proofread by third parties with relevant experience before they are used, in order to check that the legal consequences of the wording actually correspond to the desired effects. The post-use conditions of the collected research data should be formulated in such a way that they do not stand in the way of publication or further use of the data in other projects for the purpose.

Sources:

- <https://doi.org/10.18450/dataman/98>

58. Data Wildfire

Climate change hits the forest twice

The scientific monitoring of the reforestation of burned forest areas in Brandenburg was massively set back in 2022 by renewed fires in the area already affected in 2018. Even though the cause of the fire has not yet been clearly determined, the fact that a complete clearance of nearby World War II ammunition had been waived for cost reasons certainly played a role. The fire rendered the measuring instruments and any unexported data collected by them unusable.

Data collections in risk areas should be designed so that data can be exported and backed up at the earliest possible moment, preferably using the 3-2-1 back up rule. Risk assessment in advance of studies is essential.

Sources:

- <https://podcasts.apple.com/de/podcast/forschung-aktuell-deutschlandfunk/id79538418?i=1000567397808>
- <https://www.geo.de/natur/oekologie/waldbrand-in-brandenburg--wo-der-forst-in-flammen-stand-31973428.html>

59. Strong as Iron

Two myths in one story

Since the early 1900s, Popeye the Sailor, who became particularly strong by eating spinach, has been held up to children as a role model. It has been suggested that consuming spinach meets our iron needs and thus contributes to good health. But if you look into the original literature, you discover that both were myths. Both Popeye, who is said to have become strong from the iron in spinach, and the high iron value of spinach itself. In the cartoon by EC Segar, Popeye himself says "Spinach is full of Vitamin A. An' tha's what make hoomans strong an' helty!".

With the myth about the iron content of spinach, it's a bit more complicated, here one myth replaces another. The British Medical Journal published in 1981 that when the iron content of spinach was determined in the 1930s, the decimal point had accidentally slipped one place to the right. This decimal place myth is still present today in many publications on the subject of iron content in spinach. In fact, the excessively high iron content of spinach is due to the failure to take into account the difference between dried and fresh spinach.

In the case of supposedly generally valid facts as well as assumptions, one should always critically examine the available literature and, if possible, refer to the original or primary sources.

Sources:

- [https://web.archive.org/web/20111001101111/http://www.internetjournalofcriminology.com/Sutton Spinach Iron and Popeye March 2010.pdf](https://web.archive.org/web/20111001101111/http://www.internetjournalofcriminology.com/Sutton%20Spinach%20Iron%20and%20Popeye%20March%202010.pdf)
- [http://irep.ntu.ac.uk/id/eprint/30230/1/7987 Sutton.pdf](http://irep.ntu.ac.uk/id/eprint/30230/1/7987_Sutton.pdf)
- <https://www.mcgill.ca/oss/article/food-health-news-quirky-science/setting-facts-straight-about-iron-spinach>

60. Units of measurement are no laughing matter

He has put his foot in it, twice

The program "More or Less: Behind the stats" on BBC 4 has been dedicated to the understanding and sometimes debunking of figures circulating in the British media for more than 20 years. On the June 1, 2022 show Tim Harford quipped "Nautical Miles - like ordinary miles only wetter". He did so knowing that nautical miles and miles are indeed defined differently. Listener feedback vehemently pointed this difference in definitions out to him. The following week, a correction was broadcast. Ironically enough, it was from 2014. The joke had already met with criticism at the time.

This shows it is important to clearly define the units of measurement used to ensure the interpretability of the data. The criticism of the joke was therefore quite justified. In addition, the repetition of the slip-up could have been avoided. A good error culture aims to document problems that have occurred and solutions to avoid them in the future.

Sources:

<https://podcasts.apple.com/de/podcast/more-or-less-behind-the-stats/id267300884?i=1000565616932>

61. Cheat Code

If you don't take it too seriously with the AI.

The hype surrounding artificial intelligence (AI) has led to researchers in fields ranging from medicine to sociology using AI without a full understanding of the technology and its limitations. This has resulted in a wave of spurious AI-generated results. A number of publications have described astonishing results using machine learning, which is the foundation of modern AI. Machine learning involves feeding an algorithm with data from the past in order to attune it to future data that has not yet been seen. However, in several papers, researchers failed to cleanly separate the pools of data used for training vs. testing. This is a mistake that resulted in testing a system with data it has already seen.

Before using a new technique or software such as machine learning for data evaluation, one should critically examine its handling and limitations and, if necessary, also have the data evaluation checked by experts for the method used.

Sources:

- <https://www.wired.com/story/machine-learning-reproducibility-crisis/>

62. This thing can fly?

This mistake significantly accelerated the race.

The Western world was surprised when the Soviet Union launched the first artificial Earth satellite into orbit on October 4, 1957. This event is historically referred to as the "Sputnik shock." The CIA, the foreign intelligence agency of the United States, had assumed that achieving this feat would require a rocket with a thrust exceeding 1000 tons to reach orbit, which was considered unrealistic at the time. These estimates were based on their own projects. However, three years prior, the CIA had learned that the USSR was developing rockets with approximately 677 tons of thrust as part of the "Operation Dragon Return" project. The CIA chose to ignore this information. Nonetheless, the Sputnik shock had the immediate consequence of significantly intensifying the United States' efforts in the "space race," leading to the establishment of NASA.

This story illustrates the importance of ensuring that, before embarking on a research project, one should ascertain what research results are already available or what research data has been collected. It is crucial to monitor these findings throughout the course of a project. Reviewing secondary data is essential to avoid unintentional redundant research and to save time.

Sources:

- <https://de.wikipedia.org/wiki/Sputnikschock>

63. A or B?, B or A? It's all one, isn't it?

It's not that easy to become a Brit.

In the run-up to the British census in 2021, the response options to questions were optimized using empirical methods. For the question concerning one's own national identity, this was done via a card sorting pretest. The resulting order was British, English, Welsh, Scottish, Northern Irish, and Other. Compared to the 2011 census, the first two options swapped places. The results of the survey then showed that the values for British had risen sharply, while those for English had fallen. This resulted in intensive commentary both in the scientific community and the press. The 2021 data are not comparable to the 2011 data due to the sequence sorting effect which led to a discontinuity in the time series with all its troubling implications.

In principle, it makes sense to avoid structural breaks in panel data as far as possible. Therefore, it is advisable to keep both the questions and the order of answers constant for each wave of a survey. For interoperability with third-party data sets, detailed documentation of the procedure or the use of standardized variables is also advantageous. The fact that the U.K. Office of National Statistics both documented the pretest procedure and published warnings regarding the interpretability of the data along with it represents a best practice in dealing with the problem of discontinuity in the time series.

Sources:

- <https://www.ons.gov.uk/peoplepopulationandcommunity/culturalidentity/ethnicity/bulletins/nationalidentityenglandandwales/census2021>
- <https://www.bbc.co.uk/sounds/play/m001hx2z>

64. That's me, indeed!

How educational commitment became the community's undoing.

Danielle documented her transition to another gender through hormone replacement therapy via a video journal and published it on YouTube under the Standard YouTube License.

Images from her videos, along with those of other trans individuals, were used to train a facial recognition software, and they were featured in the associated scientific publication. The video data from this publication was made available to other researchers through a Dropbox link.

Danielle was not informed about the use of her video material in this study and its dissemination, and she would have never consented to it. She and the trans community fear significant discriminatory consequences if trans individuals could be identified through facial recognition in the future.

In research projects involving socially sensitive topics, and data material or personal data, an ethics committee should be consulted in the planning phase. This committee will conduct an impact assessment and, among other things, provide recommendations for extensive measures to protect involved individuals and the concerning communities, or even advise against a specific study.

The reuse, publication and sharing of copyrighted data requires a usage/license agreement between the copyright holder and the data users.

When processing personal data, informed consent must be obtained from the individuals concerned, and comprehensive data protection regulations must be observed. To the extent possible, this data should be anonymized or aggregated and made accessible to third parties only through suitable access restrictions, such as certified and specialized repositories by the means of data usage agreements.

Sources:

- <https://doi.org/10.1109/BTAS.2013.6712710>
- <https://www.theverge.com/2017/8/22/16180080/transgender-youtubers-ai-facial-recognition-dataset>
- <https://doi.org/10.1177/20539517221113772>
- <https://algorithmwatch.org/de/wenn-datensaetze-grundrechte-verletzen/>

65. Metadata against oblivion

More than a side note?

In the Duke August Library in Wolfenbüttel, the theologian and church historian Prof. Ulrich Bubenheimer is researching the history of the Reformation. In the process, he also looked through a collection volume containing several chronicles. This volume is indexed in the library database with only a few descriptive data. Marginal notes that occur in various parts of the book are not evaluated in any detail. Prof. Bubenheimer was surprised when during his research on the Reformation he came across annotations that originate from the pen of Martin Luther. The manuscript was quickly analyzed and provided insights into topics that Luther dealt with during his time in Erfurt. After evaluating the handwritten marginal notes, the chronicle volume was described with the corresponding metadata and can thus be found by "Luther researchers" from now on. In the meantime, the entire book has been digitized and can be viewed online, enabling further evaluations.

Finding the marginal notes from Luther's pen shows how important it is to deposit meaningful and evaluable metadata in databases. Only in this way can the potential of research data and objects be made visible and tapped into. The digitization of analog materials also offers the opportunity to reach a larger audience and to network research results.

Sources:

- <https://idw-online.de/de/news547606>
- <https://www.zeit.de/news/2013-08/19/literatur-notizen-von-martin-luther-entdeckt-19163807>
- <http://opac.lbs-braunschweig.gbv.de/DB=2/SET=3/TTL=1/MAT=/NOMAT=T/CLK?IKT=12&TRM=766213218>

66. Deadly Checkbox

That's one way to achieve incomparable numbers.

Between 2003 and 2017, maternal mortality in the USA appears to have doubled. A shocking trend, given the fact that the indicator has been in global decline throughout the 20th century. What happened?

Originally the US counted cases of maternal death based on the cause of death listed on death certificates. However, the procedure overlooked relevant cases. The WHO recommended the introduction of a "pregnant" checkbox on the death certificate to correct this "underreporting". Since 2003, all deaths with the box ticked have been included in the nationwide maternal mortality statistics. States rolled out the new procedure one by one over a period of 15 years.

The staggered introduction of the new recording methodology means that data on maternal mortality in the US cannot be meaningfully compared to previous data points in the time series or to data from other countries.

Only content-identical indicators may be summarized across different geographical units. If a coding scheme (especially that of a long-running international data collection) is changed, it is essential to clearly communicate the application of new elements. In particular, structural breaks should be implemented at the same time within an observation area. Alternatively, the old and new measurement methods can be applied in parallel for a transitional period to simplify aggregation between different geographical units.

Sources:

- <https://www.bbc.co.uk/programmes/p0j74zfs>
- <https://www.who.int/publications/i/item/9789240068759>
- Vgl auch: Trends in maternal mortality 2000 to 2020: estimates by WHO, UNICEF, UNFPA, World Bank Group and UNDESA/Population Division. Geneva: World Health Organization; 2023. Licence: CC BY-NC-SA 3.0 IGO

67. Clean Science

The rocky road to the truth could have been avoided.

In 2018, behavioral scientist Zoé Ziani challenged the results of Francesca Gino's study on feelings of moral and physical impurity after career networking events in her dissertation. As Gino was a successful professor at Harvard Business School, Ziani's criticism was judged to be inappropriate, self-righteous and inflammatory and she had to rewrite her paper. It was only after much effort from data forensics and lawyers that Ziani was able to prove in 2023 that Gino's findings were tantamount to data falsification. Data had been falsified in some places and good values had been replicated to fill in blank spots in others. Consequences of this incident included the demand for reproduction of old results, journals requiring data sanity checks, pre-registrations placing value on negative results too as well as open data becoming more established as a standard.

Researchers should document the collection and processing of the data as well as the analysis process. This enables a critical reception of the results and the original data and further research based on them.

Sources:

- <https://www.youtube.com/watch?v=X5MI9mrFwqE>
- https://de.wikipedia.org/wiki/Francesca_Gino
- News zur Ursprungsstudie: <https://www.spiegel.de/wissenschaft/mensch/karriere-teilnehmer-netzwerk-treffen-fuehlen-sich-dreckig-a-991230.html>
- Ähnliche Fälle in Deutschland 1: https://de.wikipedia.org/wiki/Hans-Ulrich_Wittchen#Fälschungsvorwürfe
- Ähnliche Fälle in Deutschland 2: https://de.wikipedia.org/wiki/Jens_Förster#Kontroverse

68. The good old programming script

Who was Svenja?

For years, a script titled `collate_samples_SvenjaMax.py` was used in a working group for data analysis. The script had been brought into the group years ago by a doctoral student named Max and had been repeatedly expanded and adapted. In order to follow the guidelines of good scientific practice, the plan was to publish the script together with the next publication.

To do this, it first had to be clarified whether all authors agreed to this, as the script had not yet been provided with a license. The former employees were painstakingly identified and asked for permission. However, they were only unable to identify Svenja since not even Max could remember who she was.

Source code/script code also counts as research data and should be documented, archived and published with an appropriate license. Dedicated versioning systems such as GitLab as well as classic commenting of source code can be used for documentation. It is essential that it is clear who created or changed the script and when.

Quellen:

- persönliche Kommunikation

69. The Robbed Thief

How a Call of Nature turned into a Criminal Case.

Hans was working on a study about a drug for geriatric health issues. All data (contact information, interviews, health records, medication records) from nursing home residents were stored at the research institute under strict security protocols and could only be processed on-site. With a bit of trickery, Hans managed to bypass the security measures and made local copies of all the files on his laptop.

One day, Hans was working in a café. While he used the restroom, he left his laptop unattended and without a screen lock on the table. A person, who specialized in spying on data in public spaces, took the opportunity to copy the data, including personally identifiable information (names, addresses) in order to sell the information to criminal third parties.

This story highlights extremely concerning researcher behaviour with regards to data protection: the careless handling of sensitive data in public spaces, the storing of personal data together with other research data, and the bypassing of security measures.

A key element for handling personal data safely is establishing a security-focused research culture. Researchers should be thoroughly educated and trained in data protection, especially regarding their responsibilities and obligations. This includes ensuring the security of local data, for example, through encryption and access control, as well as being fully informed about the risks of working in public spaces. If personal data is essential for analysis, pseudonyms should be used to reduce the risk of re-identification. To increase the commitment to agreed principles and measures, these can be formally documented.

Sources:

- Zellhöfer, D. & Weber-Wulf, D. (2023). Identitätsdiebstahl. In Class, C. B., Coy, W., Kurz, C. et al. (Eds). *Gewissensbisse - Fallbeispiele zu ethischen Problemen der Informatik*. Edition Medienwissenschaft. transcript Verlag. S. 91-94. DOI: 10.14361/9783839464632
- Zellhöfer, D. & Weber-Wulf, D. (2013). Gewissensbits – wie würden Sie urteilen? Fallbeispiel: Identitätsdiebstahl. *Informatik Spektrum* 36 (3): 333-335. DOI: 10.1007/s00287-013-0709-9

70. Getting ahead of the curve

The sudden change gave the virus an unfair advantage.

Germany in April 2021: in the midst of one of the strongest Covid-19 waves in Germany, the Robert Koch Institute decided to change the data format for the output of daily vaccination figures without prior notice. This made waves, especially among science and data journalists, as it meant that web scraping programs set up to use the old output format no longer worked overnight. As a result, the RKI data could no longer be crawled automatically and produced nothing but error messages. In this case, the sudden change in format and content led to considerable extra work for the journalists affected and to outages on websites and dashboards that millions of people were using to find out about the pandemic.

Data providers should be proactive and provide advance notice of upcoming format changes. For example, publications can be made available in the old and new formats for a transitional period so that data users can adapt. Good planning of the data schema and the data formats to be used in advance also helps to avoid unnecessary major changes in the course of the research process.

Quellen:

- Open Data Anti Patterns - Hase und Igel: <https://github.com/transportkollektiv/opendata-antipatterns/blob/main/patterns/formataenderung.md>
- <https://x.com/datentaeterin/status/1380203124858699778>