Introduction to Research Data Management

Qi Wang

Department of Plant Sciences

What is "Data"?

- Raw instrument readings
- Processed/Analysed data
- Microscopic photos
- Western blot images
- Videos
- Measurements
- Spreadsheets
- Metadata
- Surveys and interviews
- Field notes
- Maps
- Lab books
- Physical samples
- Protocols
- Software
- Graphs/Figures

It is anything you produce in the course of your research and is the 'bed-rock' of your findings!

Data Types Recommended by UK Data Archive

Type of data	**Recommended formats**
Tabular data with extensive metadata variable labels, code labels, and defined missing values	SPSS portable format (.por) delimited text and [illegible] (SPSS, Stata, SAS, etc.) structured [illegible] metadata information, e.g. DDI X[illegible]
Tabular data with minimal metadata column headings, variable na[illegible]	[illegible] values (.csv) [illegible] file (.tab) delimited text with SQL data [illegible]
Geospatial data vector and raster data	ESRI Shapefile [illegible] .sbn optional) geo-ref[illegible] [illegible] [illegible] attribute data Geography Markup Language (.gml)
Textual data	Rich Text Format (.rtf) plain text, ASCII (.txt) eXtensible Mark-up Language (.xml) text according to an appropriate Document Type Definition (DTD) or schema
Image data	TIFF 6.0 uncompressed (.tif)
Audio data	Free Lossless Audio Codec (FLAC) (.flac)
Video data	MPEG-4 (.mp4) OGG video (.ogv, .ogg) motion JPEG 2000 (.mj2)
Documentation and scripts	Rich Text Format (.rtf) PDF/UA, PDF/A or PDF (.pdf) XHTML or HTML (.xhtml, .htm) OpenDocument Text (.odt)

Open and non-proprietary

Information loss during conversion

https://ukdataservice.ac.uk/learning-hub/research-data-management/format-your-data/recommended-formats/

Why we need data management?

“Where did I put that file??”

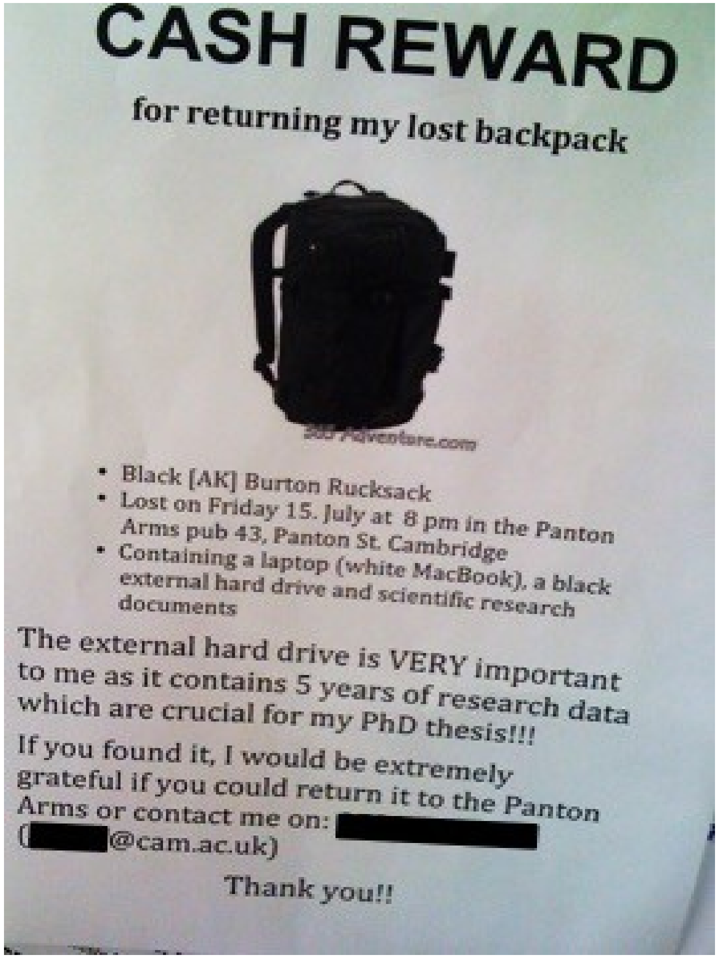
Why we need data management?

"I'm asked to continue the project of a previous student/postdoc, but ..."

Area of Data Management

- creation and reuse
- **storage and backup**
- **organization**
- sharing

Why we need backup?



CASH REWARD

for returning my lost backpack

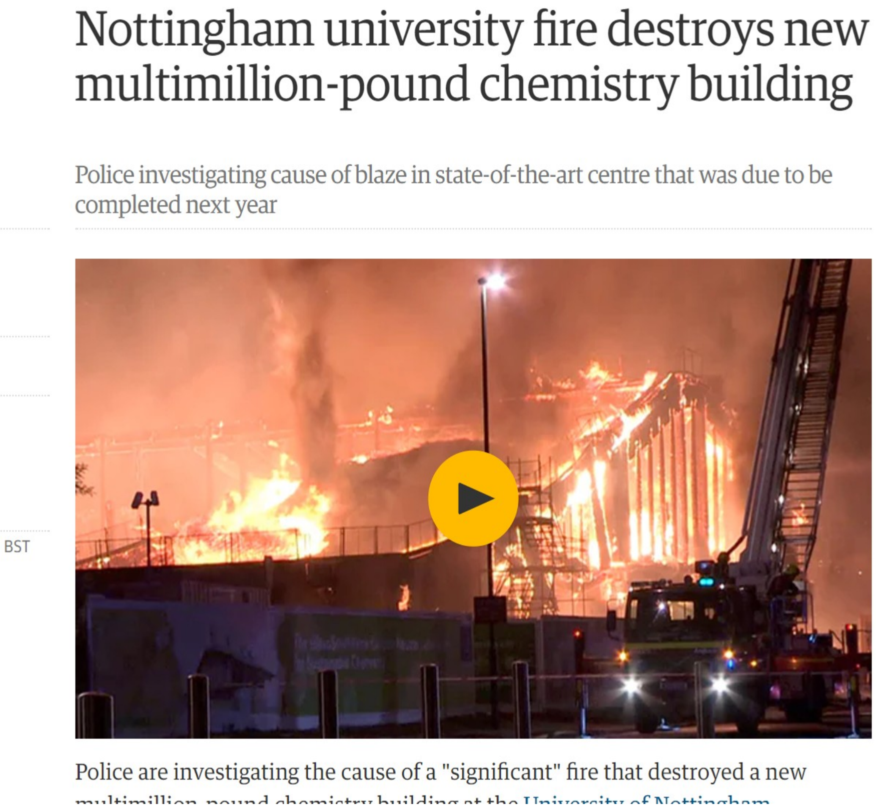
360Adventure.com

- Black [AK] Burton Rucksack
- Lost on Friday 15. July at 8 pm in the Panton Arms pub 43, Panton St. Cambridge
- Containing a laptop (white MacBook), a black external hard drive and scientific research documents

The external hard drive is VERY important to me as it contains 5 years of research data which are crucial for my PhD thesis!!!

If you found it, I would be extremely grateful if you could return it to the Panton Arms or contact me on: [redacted] ([redacted]@cam.ac.uk)

Thank you!!



Nottingham university fire destroys new multimillion-pound chemistry building

Police investigating cause of blaze in state-of-the-art centre that was due to be completed next year

Police are investigating the cause of a "significant" fire that destroyed a new multimillion-pound chemistry building at the University of Nottingham

Credit: Peter Murray-Rust, http://blogs.ch.cam.ac.uk/pmr/2011/08/01/why-you-need-a-data-management-plan/, August 2011,CC-BY

https://www.theguardian.com/uk-news/2014/sep/13/nottingham-university-fire-police-investigate-significant-blaze

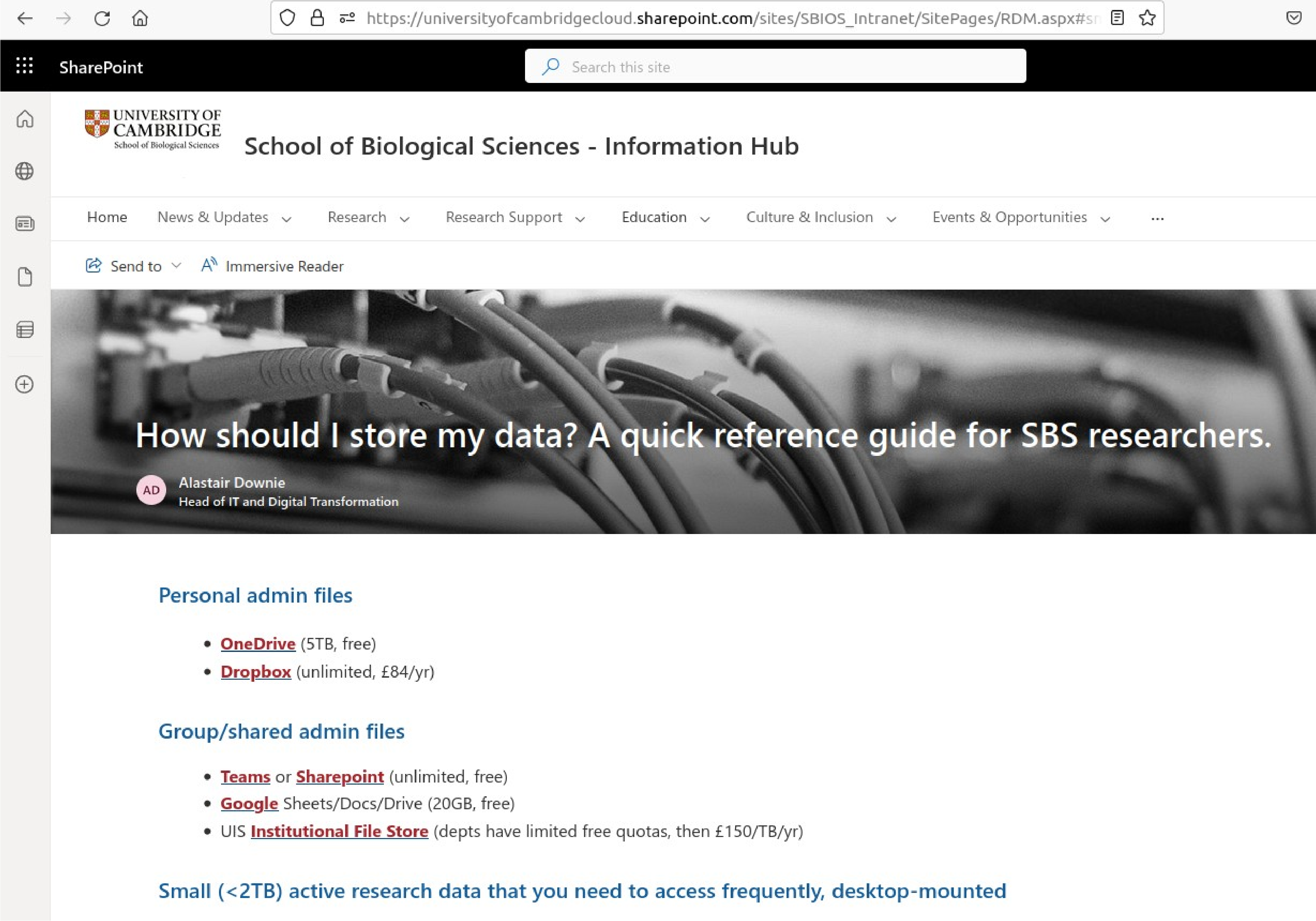
Backup Strategy

Cloud

Departmental server
(hydrogen)

External Disks

At least 2 backups, at 2 different locations
Accountable backup frequency



How should I store my data? A quick reference guide for SBS researchers.

Alastair Downie
Head of IT and Digital Transformation

Personal admin files

- **OneDrive** (5TB, free)
- **Dropbox** (unlimited, £84/yr)

Group/shared admin files

- **Teams** or **Sharepoint** (unlimited, free)
- **Google** Sheets/Docs/Drive (20GB, free)
- UIS **Institutional File Store** (depts have limited free quotas, then £150/TB/yr)

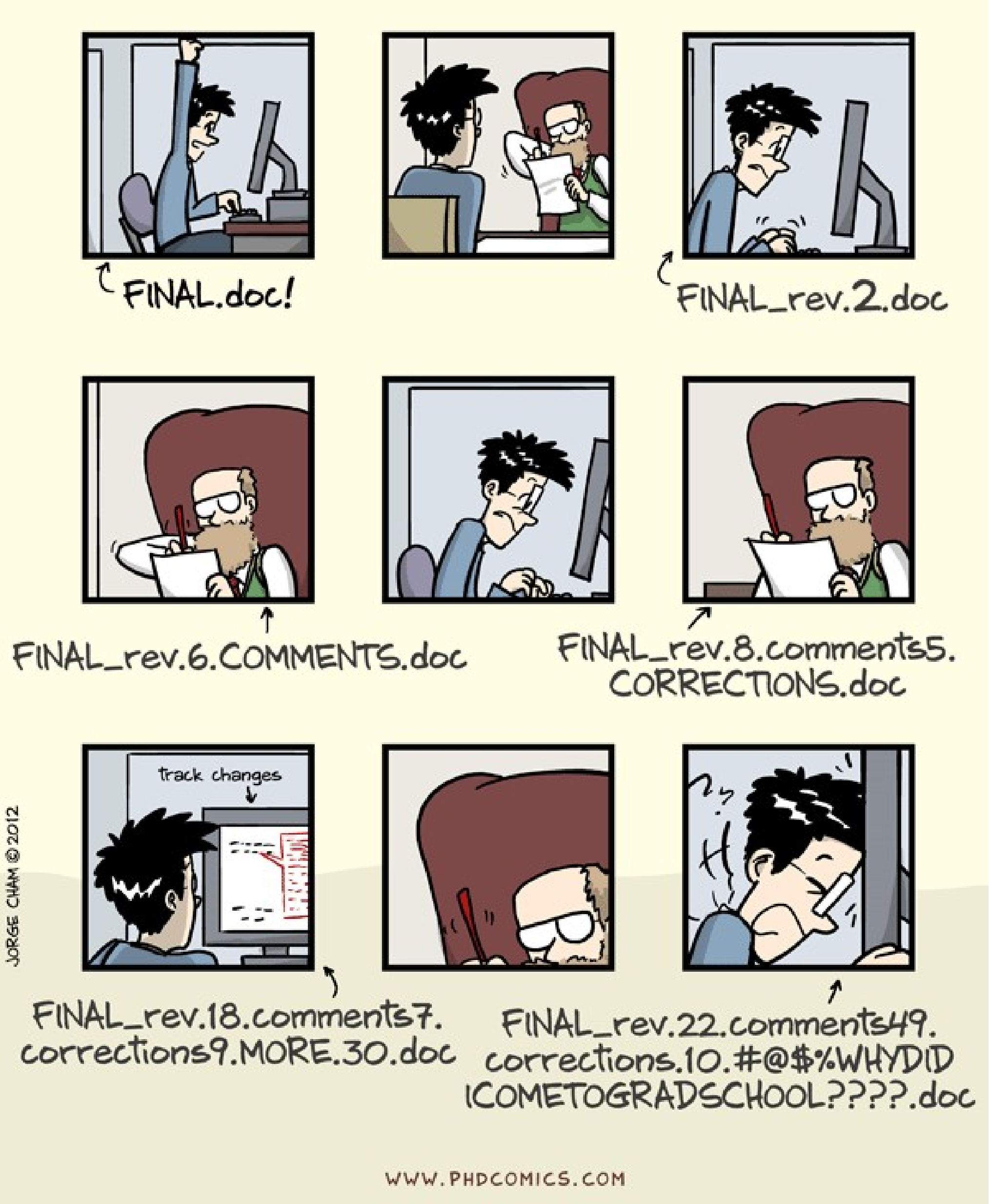
Small (<2TB) active research data that you need to access frequently, desktop-mounted

https://universityofcambridgecloud.sharepoint.com/sites/SBIOS_Intranet/SitePages/RDM.aspx

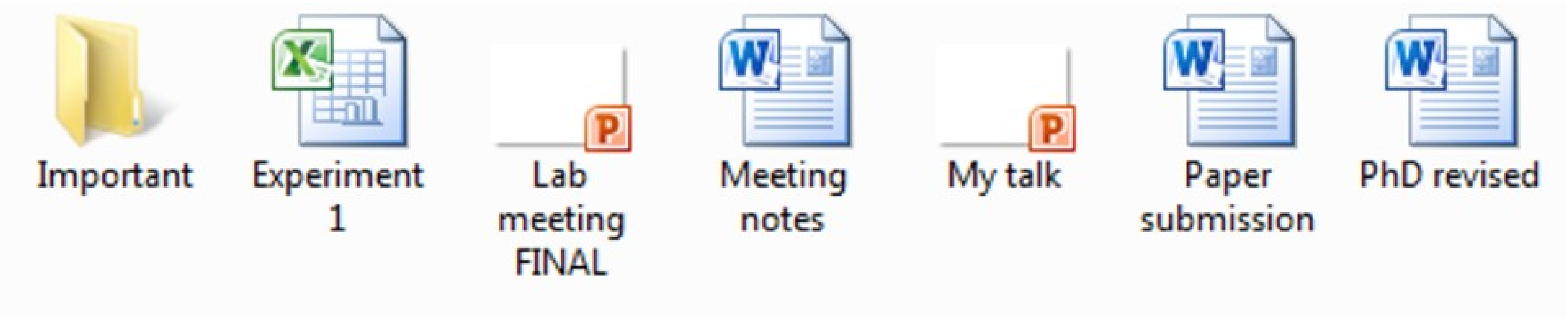
Area of Data Management

- Creation and reuse
- Storage and backup
- **Organization**
    - File naming
    - File Organization
    - Metadata
- sharing

File Naming – does it matter?



File Naming – does it matter?



In 3 years' time would you know what these are?

Sacha Jones (www.data.cam.ac.uk)

File Naming – 3C principles

Criteria: *Can your collaborator (or yourself 5 years from now) identify the content without opening the file?*

- **Clear**
  - Objective: ~~my~~, ~~current~~, ~~latest~~, ~~final~~
  - Meaningful: He?
- **C**oncise
  - ~~the~~, ~~and~~
- **C**onsistent
  - *qPCR_batch1_20190130.csv & batch2_qPCR_1201.csv*
  - *[Date]_[Run]_[SampleType]*

https://www.jisc.ac.uk/guides/managing-information/good-file-name

File Naming – Other Tips

- Use **underscores** “_ “ to separate elements
  - avoid spaces “ “ and special characters, *e.g.*, “@”
  - Periods “. “ only before the file extension

- *e.g.*, compare:

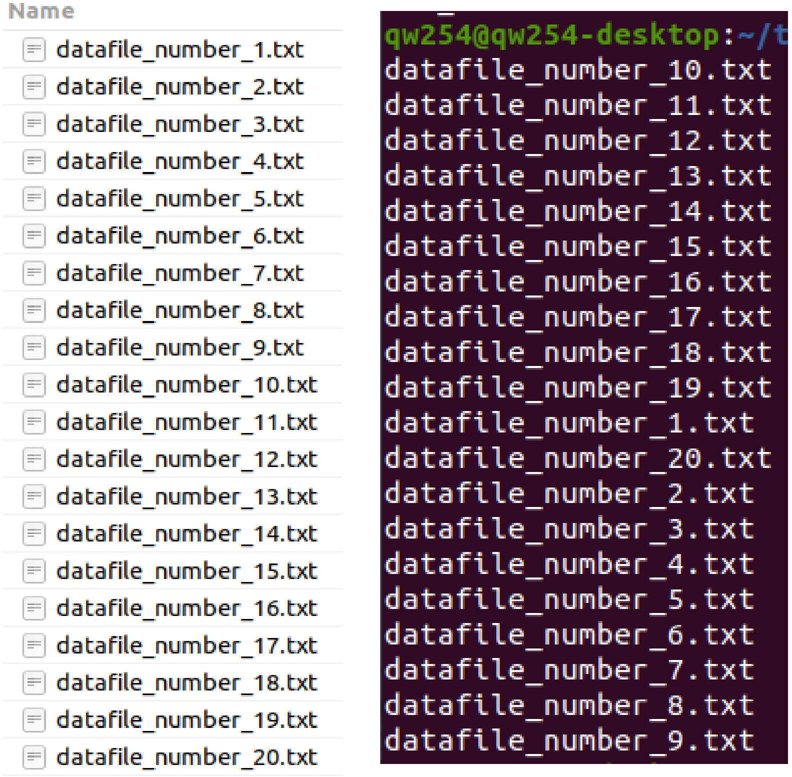
  averagetrendclusterearlyonly.png 😓

  average_trend_cluster_early_only.png 😌

File Naming – Other Tips

- Use **underscores** “_ “ to separate elements
  - avoid spaces “ “ and special characters, *e.g.*, “@”
  - Periods “. “ only before the file extension
- Use **leading zero** for consistent sorting

Without Leading Zero

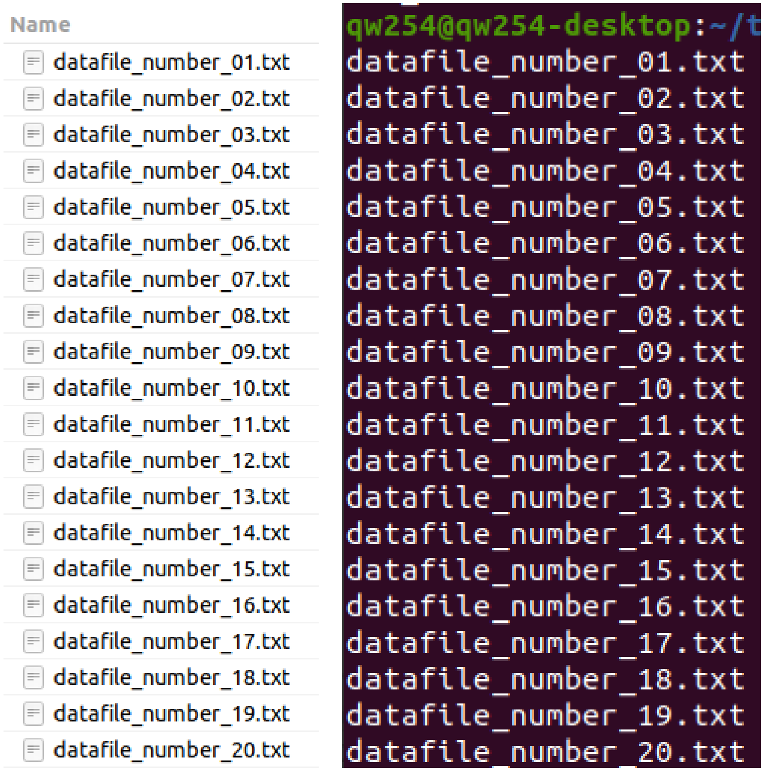


```
Name
datafile_number_1.txt
datafile_number_2.txt
datafile_number_3.txt
datafile_number_4.txt
datafile_number_5.txt
datafile_number_6.txt
datafile_number_7.txt
datafile_number_8.txt
datafile_number_9.txt
datafile_number_10.txt
datafile_number_11.txt
datafile_number_12.txt
datafile_number_13.txt
datafile_number_14.txt
datafile_number_15.txt
datafile_number_16.txt
datafile_number_17.txt
datafile_number_18.txt
datafile_number_19.txt
datafile_number_20.txt
```

```
qw254@qw254-desktop:~/t
datafile_number_10.txt
datafile_number_11.txt
datafile_number_12.txt
datafile_number_13.txt
datafile_number_14.txt
datafile_number_15.txt
datafile_number_16.txt
datafile_number_17.txt
datafile_number_18.txt
datafile_number_19.txt
datafile_number_1.txt
datafile_number_20.txt
datafile_number_2.txt
datafile_number_3.txt
datafile_number_4.txt
datafile_number_5.txt
datafile_number_6.txt
datafile_number_7.txt
datafile_number_8.txt
datafile_number_9.txt
```

Not consistent sorting.

With Leading Zero



```
Name
datafile_number_01.txt
datafile_number_02.txt
datafile_number_03.txt
datafile_number_04.txt
datafile_number_05.txt
datafile_number_06.txt
datafile_number_07.txt
datafile_number_08.txt
datafile_number_09.txt
datafile_number_10.txt
datafile_number_11.txt
datafile_number_12.txt
datafile_number_13.txt
datafile_number_14.txt
datafile_number_15.txt
datafile_number_16.txt
datafile_number_17.txt
datafile_number_18.txt
datafile_number_19.txt
datafile_number_20.txt
```

```
qw254@qw254-desktop:~/t
datafile_number_01.txt
datafile_number_02.txt
datafile_number_03.txt
datafile_number_04.txt
datafile_number_05.txt
datafile_number_06.txt
datafile_number_07.txt
datafile_number_08.txt
datafile_number_09.txt
datafile_number_10.txt
datafile_number_11.txt
datafile_number_12.txt
datafile_number_13.txt
datafile_number_14.txt
datafile_number_15.txt
datafile_number_16.txt
datafile_number_17.txt
datafile_number_18.txt
datafile_number_19.txt
datafile_number_20.txt
```

Consistent!

Batching Renaming Tools

Windows:

- Ant Renamer: http://www.antp.be/software/renamer
- Bulk Rename Utility: http://www.bulkrenameutility.co.uk/
- PSRenamer: http://www.powersurgepub.com/products/psrenamer.html

Mac:

- PSRenamer: http://www.powersurgepub.com/products/psrenamer.html
- Renamer4Mac : http://renamer4mac.com/
- Name Mangler: http://manytricks.com/namemangler/

Linux/Unix:

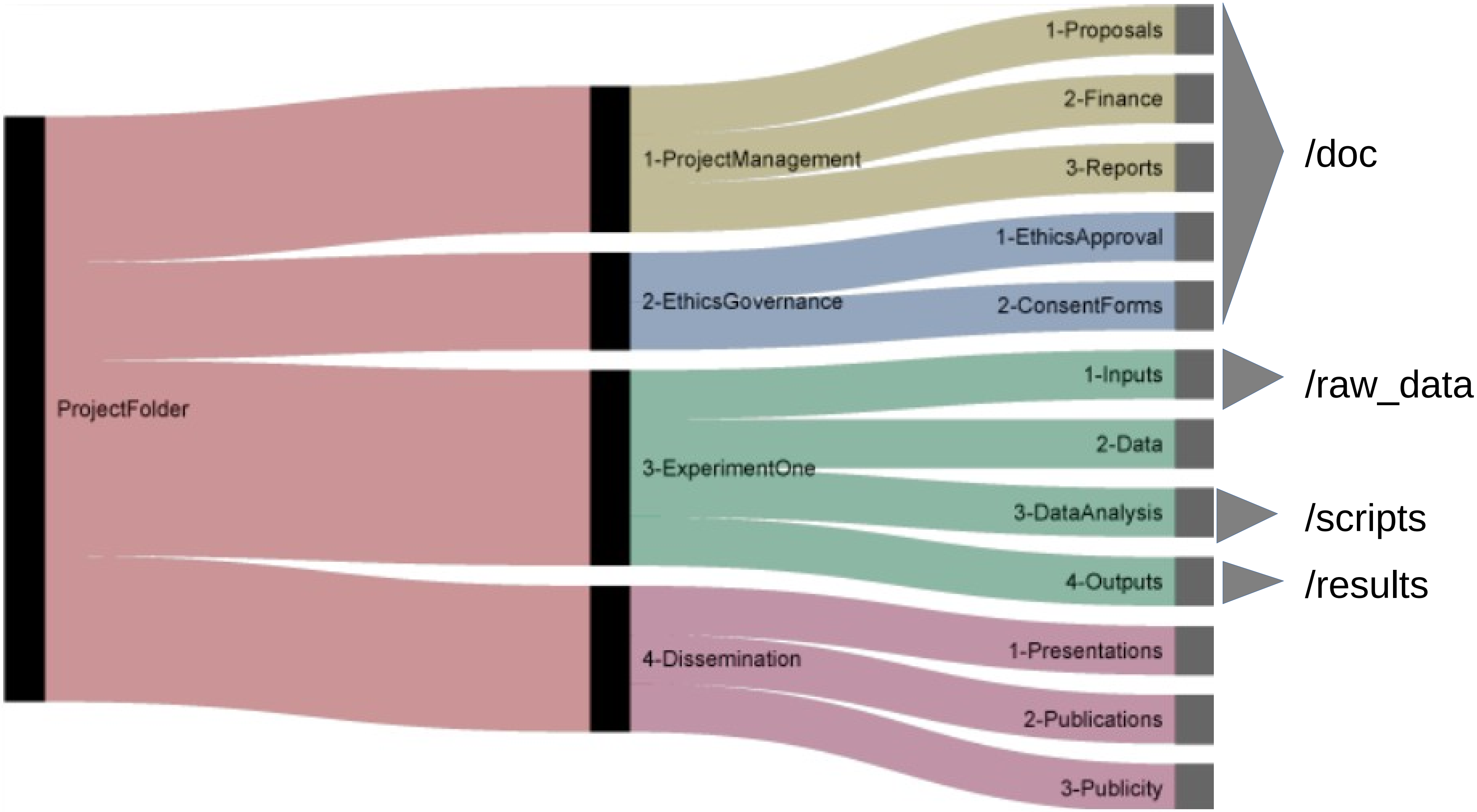
- GNOME Commander: http://www.nongnu.org/gcmd/
- PSRenamer: http://www.powersurgepub.com/products/psrenamer.html
- Use *grep, sed and awk* to search for and change

More information:
https://libraries.mit.edu/data-management/store/organize/
Batch file renaming tools handout (pdf)

Area of Data Management

- Creation and reuse
- Storage and backup
- Organization
  - File naming
  - **File Organization**
  - Metadata
- sharing

Clear Folder Structure (Example)



- Balance between breadth and depth
- to_be_sorted

http://nikola.me/folder_structure.html

Area of Data Management

- Creation and reuse
- Storage and backup
- Organization
  - File naming
  - File Organization
  - Metadata
- Sharing

Why metadata?

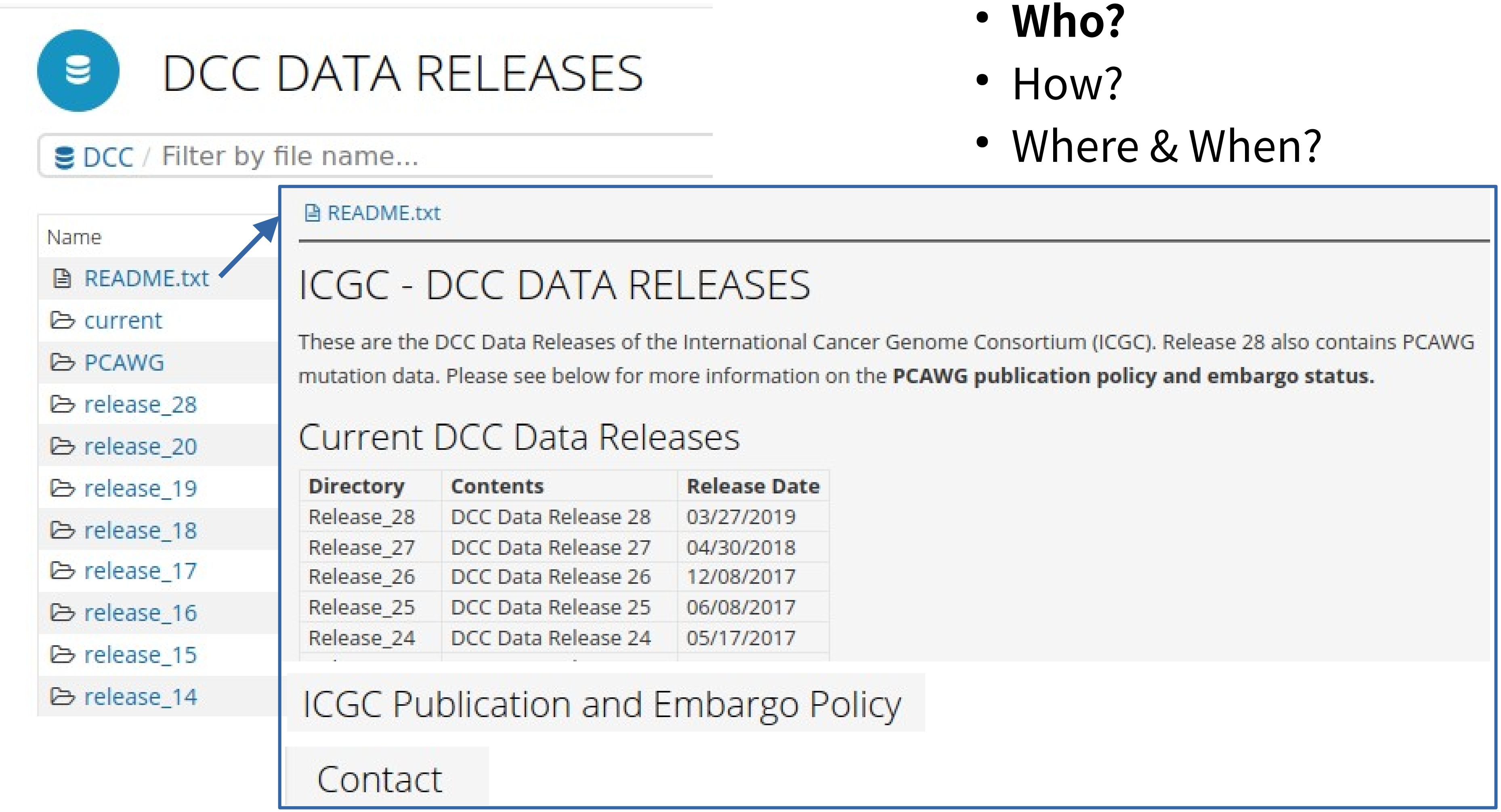
- What is metadata?
  - Description that helps someone else understand the contents and organization of your files *in your absence*

- What should metadata include?
  - What?
  - Who?
  - Where & When?
  - How?

@ Project-level @ Data-level @ File-level

What Is in Metadata? - @Project-Level

- **What?**
- **Who?**
- How?
- Where & When?



DCC DATA RELEASES

DCC / Filter by file name...

Name
README.txt
current
PCAWG
release_28
release_20
release_19
release_18
release_17
release_16
release_15
release_14

README.txt

ICGC - DCC DATA RELEASES

These are the DCC Data Releases of the International Cancer Genome Consortium (ICGC). Release 28 also contains PCAWG mutation data. Please see below for more information on the **PCAWG publication policy and embargo status.**

Current DCC Data Releases

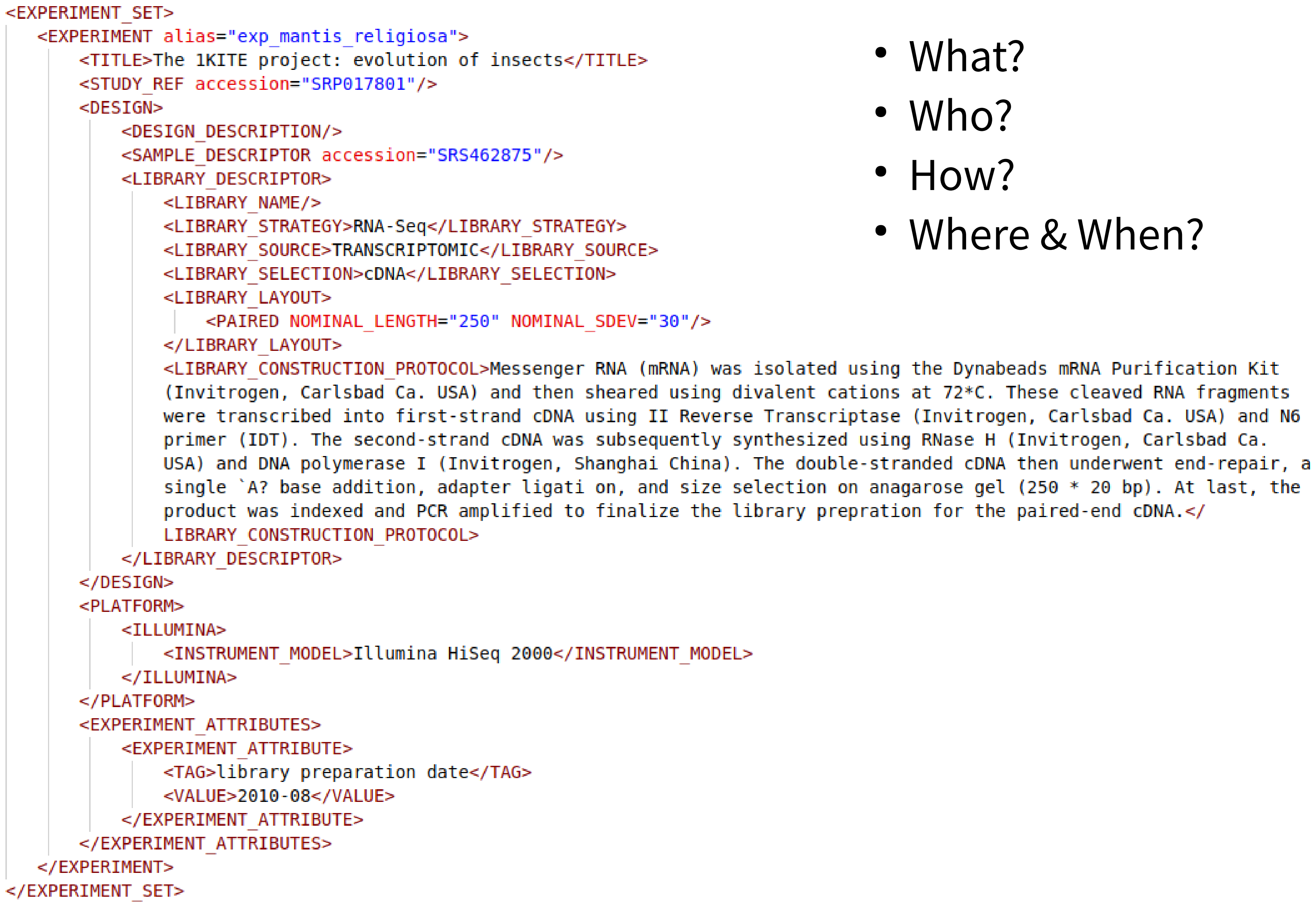
Directory	Contents	Release Date
Release_28	DCC Data Release 28	03/27/2019
Release_27	DCC Data Release 27	04/30/2018
Release_26	DCC Data Release 26	12/08/2017
Release_25	DCC Data Release 25	06/08/2017
Release_24	DCC Data Release 24	05/17/2017

ICGC Publication and Embargo Policy

Contact

https://dcc.icgc.org/releases

What Is in Metadata? - @Data-Level

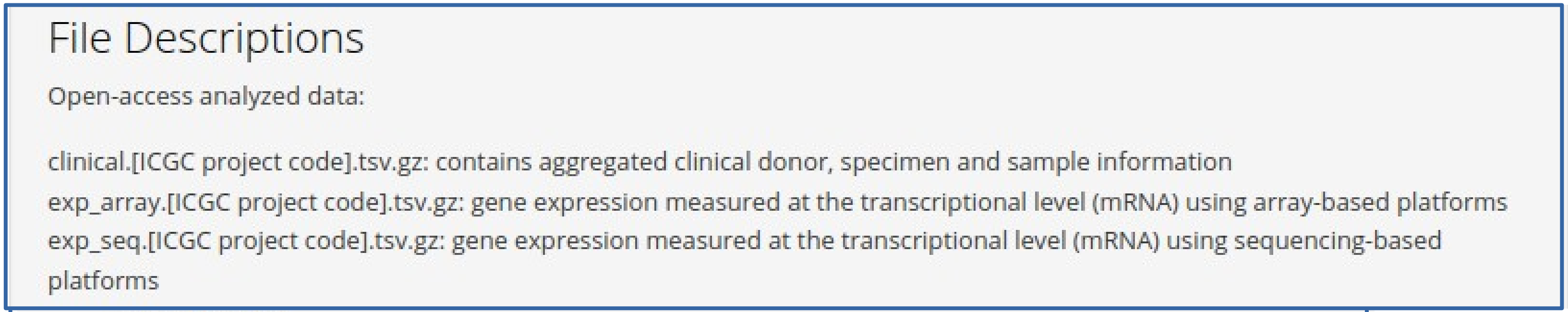


```
<EXPERIMENT_SET>
   <EXPERIMENT alias="exp_mantis_religiosa">
       <TITLE>The 1KITE project: evolution of insects</TITLE>
       <STUDY_REF accession="SRP017801"/>
       <DESIGN>
           <DESIGN_DESCRIPTION/>
           <SAMPLE_DESCRIPTOR accession="SRS462875"/>
           <LIBRARY_DESCRIPTOR>
               <LIBRARY_NAME/>
               <LIBRARY_STRATEGY>RNA-Seq</LIBRARY_STRATEGY>
               <LIBRARY_SOURCE>TRANSCRIPTOMIC</LIBRARY_SOURCE>
               <LIBRARY_SELECTION>cDNA</LIBRARY_SELECTION>
               <LIBRARY_LAYOUT>
                   <PAIRED NOMINAL_LENGTH="250" NOMINAL_SDEV="30"/>
               </LIBRARY_LAYOUT>
               <LIBRARY_CONSTRUCTION_PROTOCOL>Messenger RNA (mRNA) was isolated using the Dynabeads mRNA Purification Kit
               (Invitrogen, Carlsbad Ca. USA) and then sheared using divalent cations at 72*C. These cleaved RNA fragments
               were transcribed into first-strand cDNA using II Reverse Transcriptase (Invitrogen, Carlsbad Ca. USA) and N6
               primer (IDT). The second-strand cDNA was subsequently synthesized using RNase H (Invitrogen, Carlsbad Ca.
               USA) and DNA polymerase I (Invitrogen, Shanghai China). The double-stranded cDNA then underwent end-repair, a
               single `A? base addition, adapter ligati on, and size selection on anagarose gel (250 * 20 bp). At last, the
               product was indexed and PCR amplified to finalize the library prepration for the paired-end cDNA.</
               LIBRARY_CONSTRUCTION_PROTOCOL>
           </LIBRARY_DESCRIPTOR>
       </DESIGN>
       <PLATFORM>
           <ILLUMINA>
               <INSTRUMENT_MODEL>Illumina HiSeq 2000</INSTRUMENT_MODEL>
           </ILLUMINA>
       </PLATFORM>
       <EXPERIMENT_ATTRIBUTES>
           <EXPERIMENT_ATTRIBUTE>
               <TAG>library preparation date</TAG>
               <VALUE>2010-08</VALUE>
           </EXPERIMENT_ATTRIBUTE>
       </EXPERIMENT_ATTRIBUTES>
   </EXPERIMENT>
</EXPERIMENT_SET>
```

- What?
- Who?
- How?
- Where & When?

https://ena-docs.readthedocs.io/en/latest/submit/reads/programmatic.html

What Is in Metadata? - @File-Level



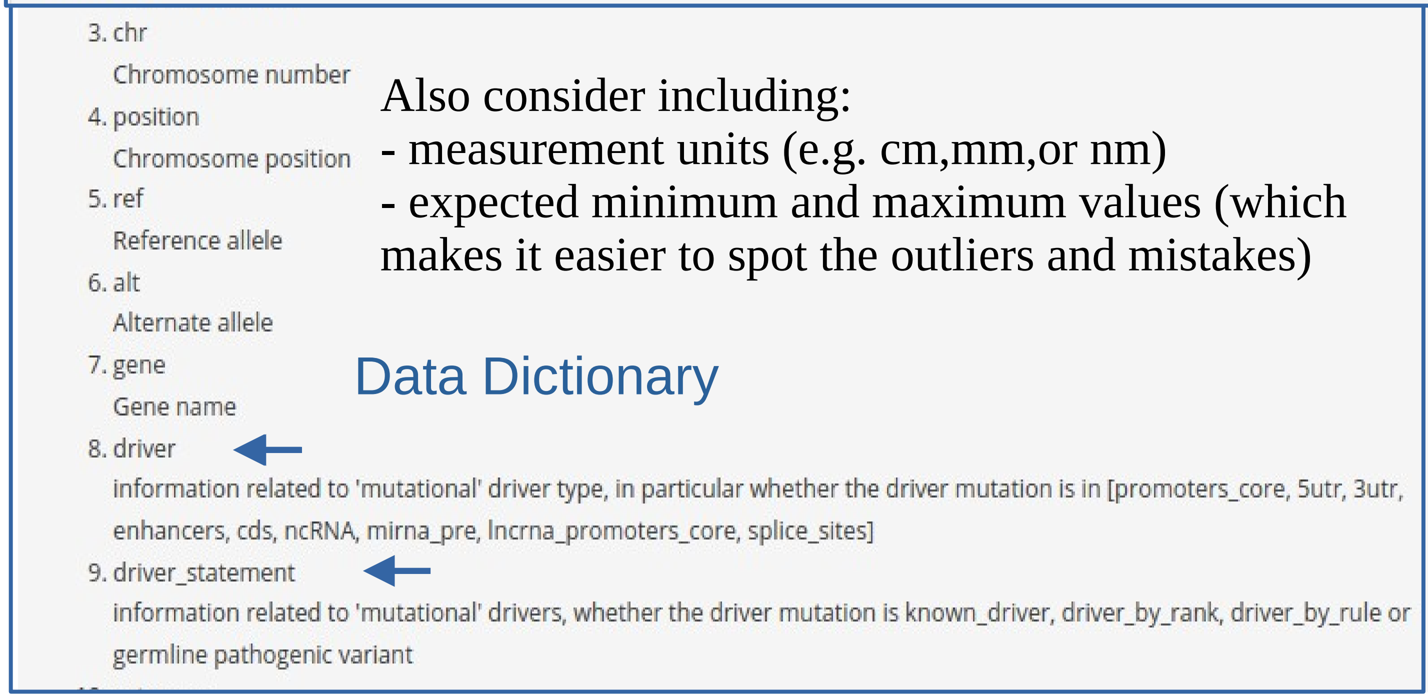
File Descriptions

Open-access analyzed data:

clinical.[ICGC project code].tsv.gz: contains aggregated clinical donor, specimen and sample information

exp_array.[ICGC project code].tsv.gz: gene expression measured at the transcriptional level (mRNA) using array-based platforms

exp_seq.[ICGC project code].tsv.gz: gene expression measured at the transcriptional level (mRNA) using sequencing-based platforms



3. chr
   Chromosome number
4. position
   Chromosome position
5. ref
   Reference allele
6. alt
   Alternate allele
7. gene
   Gene name
8. driver
   information related to 'mutational' driver type, in particular whether the driver mutation is in [promoters_core, 5utr, 3utr, enhancers, cds, ncRNA, mirna_pre, lncrna_promoters_core, splice_sites]
9. driver_statement
   information related to 'mutational' drivers, whether the driver mutation is known_driver, driver_by_rank, driver_by_rule or germline pathogenic variant

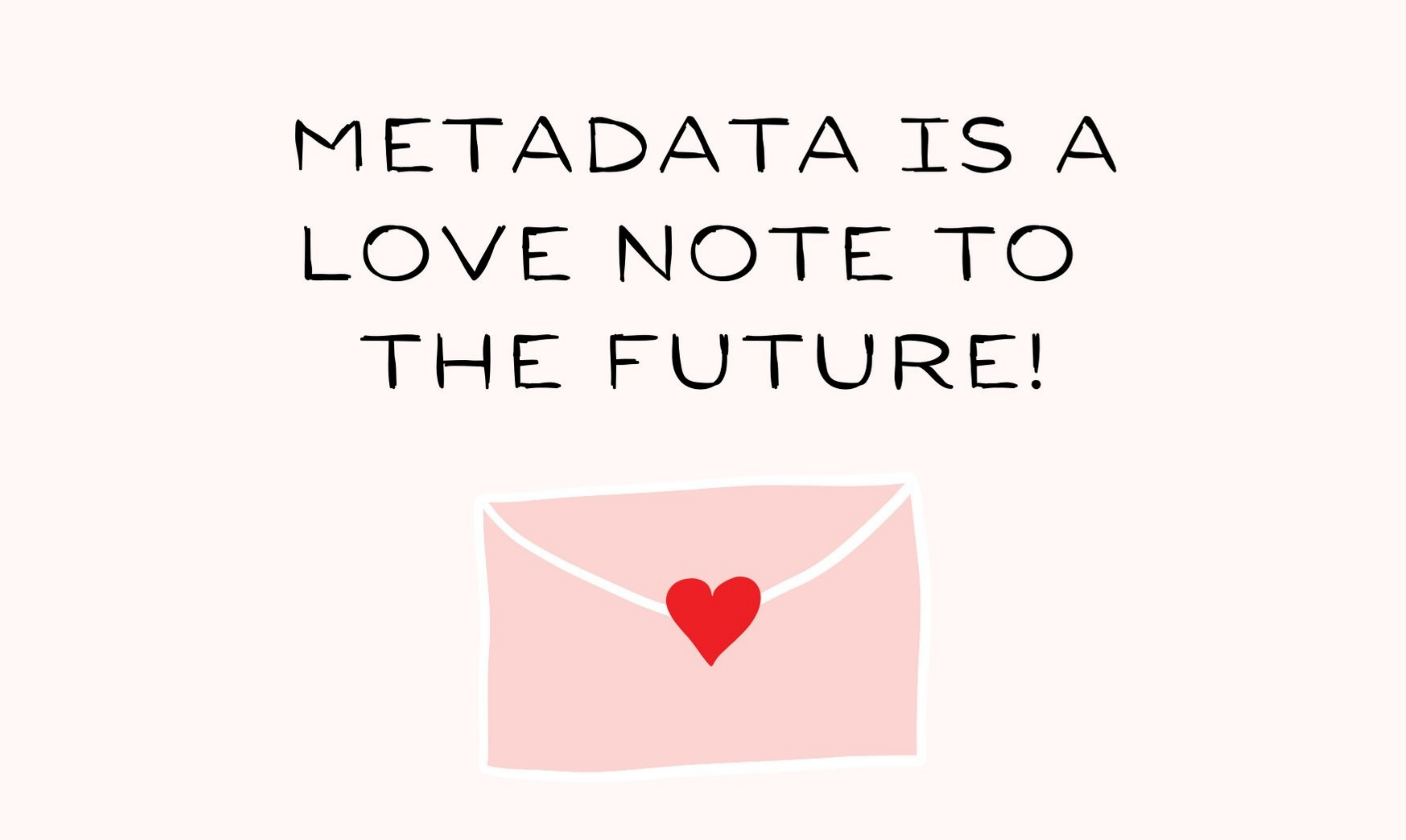
Also consider including:
- measurement units (e.g. cm,mm,or nm)
- expected minimum and maximum values (which makes it easier to spot the outliers and mistakes)

Data Dictionary

- **What?**
- Who?
- How?
- Where & When?

More on data dictionary: http://kbroman.org/dataorg/pages/dictionary.html

https://dcc.icgc.org/releases



METADATA IS A LOVE NOTE TO THE FUTURE!

https://twitter.com/hlrdmp/status/1360263691061633024/photo/1

No one has
perfect data management habits,
but adopting even a few
goes a long way.

Photo by Jesse Bowser on Unsplash

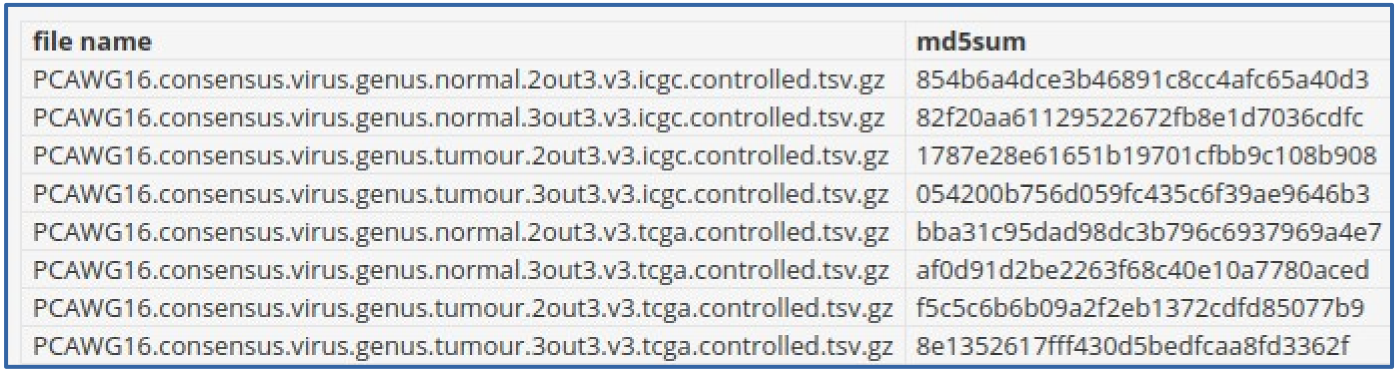
What if your file transfer

got **interrupted**

without any warning message?

What Is in Metadata?

- Avoid pitfalls in data transfer using md5sum check



file name	md5sum
PCAWG16.consensus.virus.genus.normal.2out3.v3.icgc.controlled.tsv.gz	854b6a4dce3b46891c8cc4afc65a40d3
PCAWG16.consensus.virus.genus.normal.3out3.v3.icgc.controlled.tsv.gz	82f20aa61129522672fb8e1d7036cdfc
PCAWG16.consensus.virus.genus.tumour.2out3.v3.icgc.controlled.tsv.gz	1787e28e61651b19701cfbb9c108b908
PCAWG16.consensus.virus.genus.tumour.3out3.v3.icgc.controlled.tsv.gz	054200b756d059fc435c6f39ae9646b3
PCAWG16.consensus.virus.genus.normal.2out3.v3.tcga.controlled.tsv.gz	bba31c95dad98dc3b796c6937969a4e7
PCAWG16.consensus.virus.genus.normal.3out3.v3.tcga.controlled.tsv.gz	af0d91d2be2263f68c40e10a7780aced
PCAWG16.consensus.virus.genus.tumour.2out3.v3.tcga.controlled.tsv.gz	f5c5c6b6b09a2f2eb1372cdfd85077b9
PCAWG16.consensus.virus.genus.tumour.3out3.v3.tcga.controlled.tsv.gz	8e1352617fff430d5bedfcaa8fd3362f

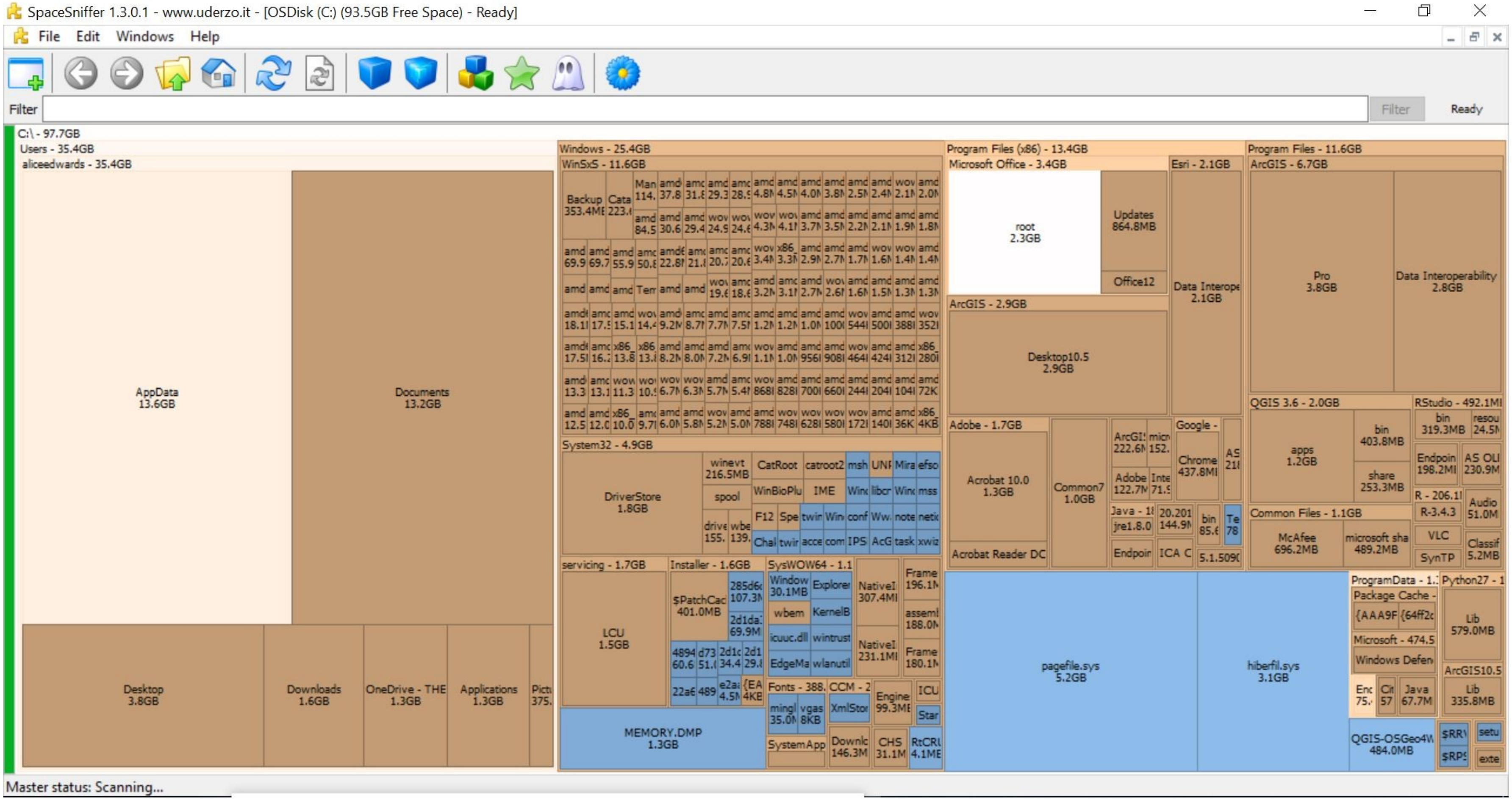
- Md5sum output are “**fingerprints**” to files. They are hash values derived using the whole file as input.
- Changes to a file will cause md5sum output to change. Conversely, if md5sum outputs are the same the files are identical.

Note : If you are worried that the data is maliciously altered instead of accidental corruption, there are more advanced options: SHA-256 (sha256sum), SHA-512 (sha512sum) or BLAKE2(b2sum).

https://dcc.icgc.org/releases

Running out of Space – for Windows

SpaceSniffer *http://www.uderzo.it/main_products/space_sniffer/index.html*



SpaceSniffer 1.3.0.1 - www.uderzo.it - [OSDisk (C:) (93.5GB Free Space) - Ready]

Master status: Scanning...

Abigail Edwards

Running out of Space – for Mac & other Linux

For Mac:

Disk Inventory X

http://www.derlien.com/

Linux command line (bash):

```
du -sh # shows you how much disk space the current folder takes

du -h -d 1 | sort -h # sort all folders in the current directory by size
```

Summary

- What data is & data format
- **-** Storage & backup
- Organization
  - File naming (3C)
  - File Organization
  - Metadata (**W W H** + **W**here and **W**hen)
- Avoid pitfalls in data transfer using md5sum check
- Running out of storage space

Not Covered Here

- Electronic Lab Notebooks

- Version Control (protocols, manuscripts, code, etc.)

- Data Sharing
  - FAIR principle (**F**indable, **A**ccessible, **I**nteroperable, and **R**eusable)
  - Repositories (ENA, Apollo, etc.)

- License (*e.g.* CC-BY**)**

- Data Formatting (tabular data)

- Data Management Plan

Where to find support

People

Data team at the Office of Scholarly Communication

Your departmental librarian

Data Champions

Resources

Data management libguide – reminders, videos and further readings

DMPOnline – Data Management Plan template

RDM policy framework – expectations at Cambridge

Homework

?