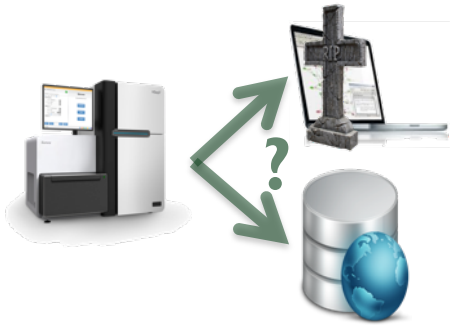
Project Data Management

Niclas Jareborg, NBIS
niclas.jareborg@bils.se

Introduction to NGS course, 2016-12-02

Why manage research data?

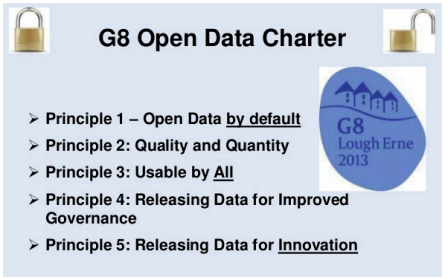
- To make your research easier!
- To stop yourself drowning in irrelevant stuff
- In case you need the data later
- To avoid accusations of fraud or bad science
- To share your data for others to use and learn from
- To get credit for producing it
- Because funders or your organisation require it



Well-managed data opens up opportunities for re-use, integration and new science

Open Access to research data

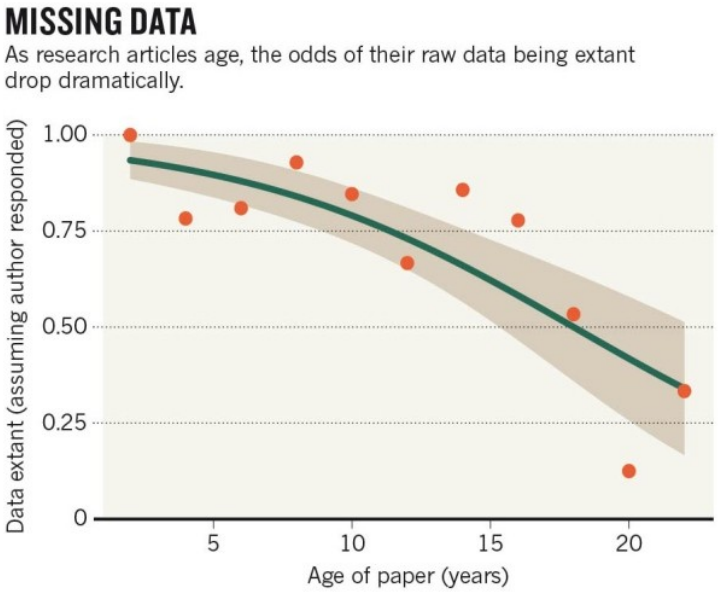
- *The practice of providing on-line access to scientific information that is free of charge to the end-user and that is re-usable.*
  - Does not necessarily mean unrestricted access, e.g. for sensitive personal data
- Strong international movement towards Open Access (OA)
- European Commission recommended the member states to establish national guidelines for OA
  - Swedish Research Council (VR) submitted proposal to the government Jan 2015
- Research bill 2017–2020 – *28 Nov 2016*
  - "*The aim of the government is that all scientific publications that are the result of publicly funded research should be openly accessible as soon as they are published. Likewise,* ***research data*** *underlying scientific publications should be* ***openly accessible*** *at the time of publication.*"
    [my translation]



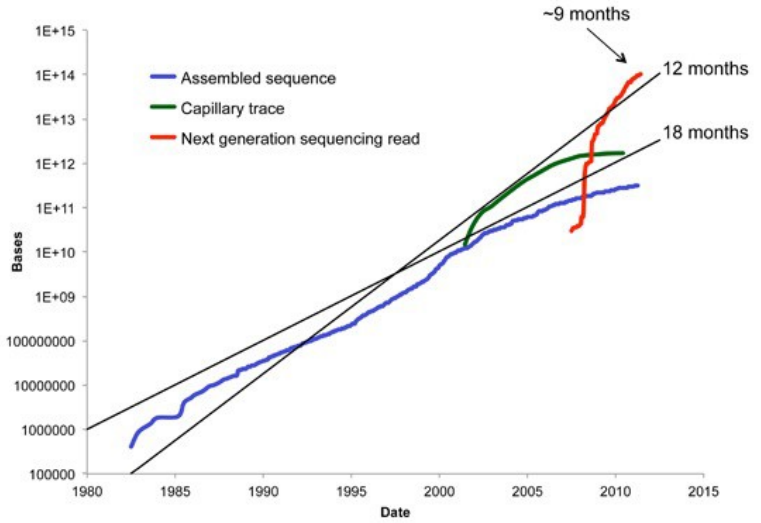
Why Open Access?

- Democracy and transparency
  - Publicly funded research data should be accessible to all
  - Published results and conclusions should be possible to check by others
- Research
  - Enables others to combine data, address new questions, and develop new analytical methods
  - Reduce duplication and waste
- Innovation and utilization outside research
  - Public authorities, companies, and private persons outside research can make use of the data
- Citation
  - Citation of data will be a merit for the researcher that produced it

Data loss is real and significant, while data growth is staggering



Nature news, 19 December 2013



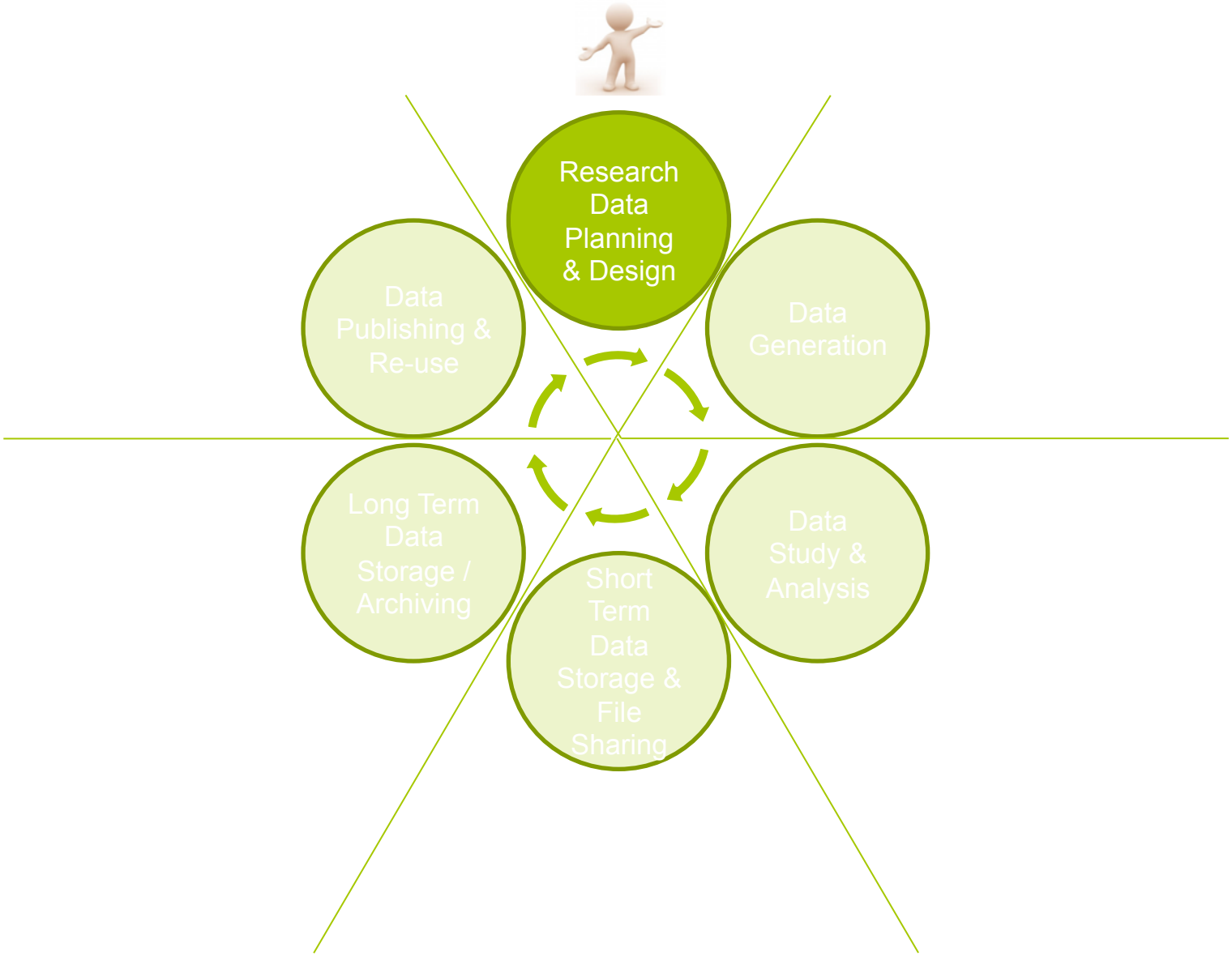
- DNA sequence data is doubling every 6-8 months and looks to continue for this decade
- Projected to surpass astronomy data in the coming decade

'Oops, that link was the laptop of my PhD student'

Slide stolen from Barend Mons

The Research Data Life Cycle

Planning & Design

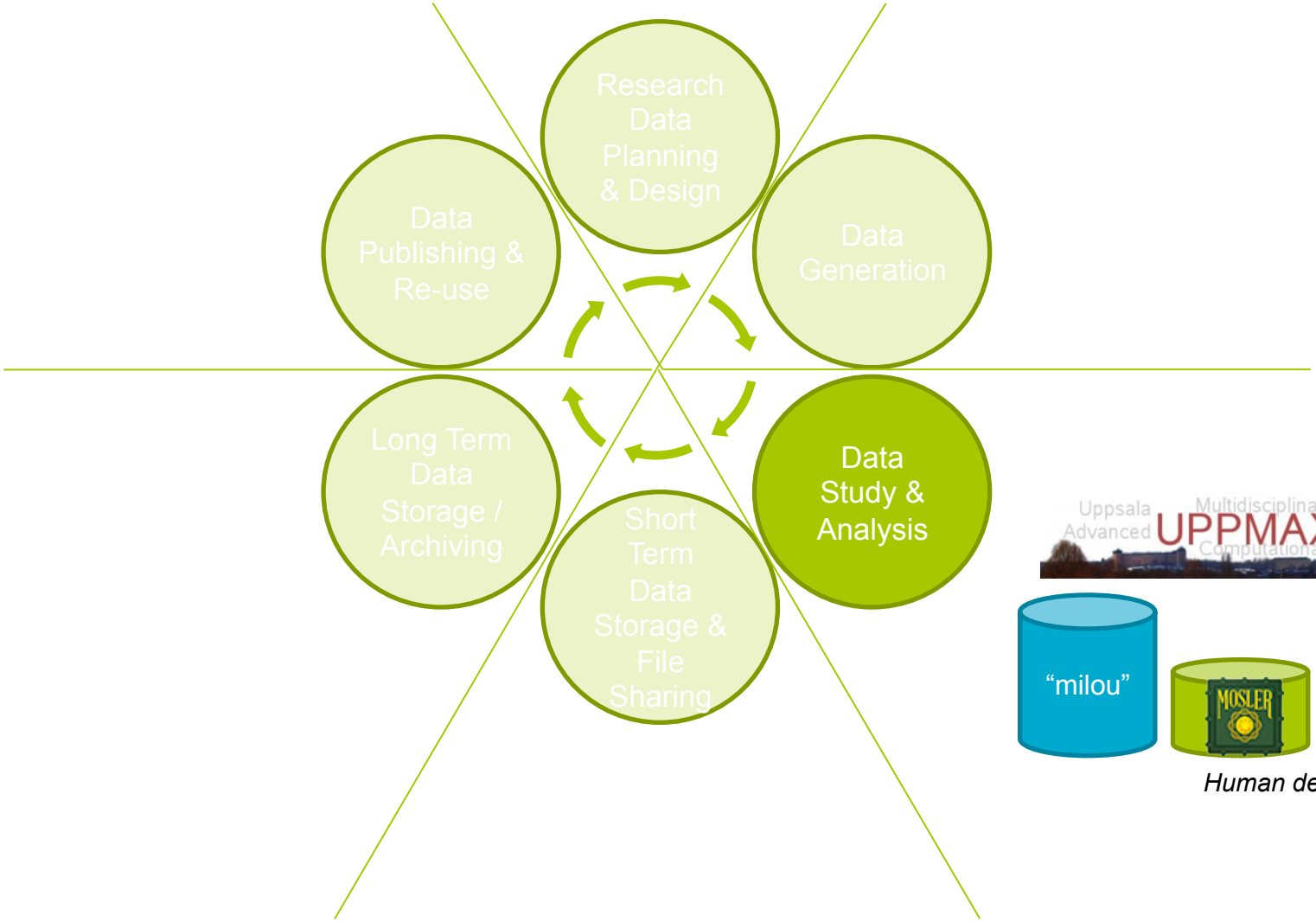


Planning & Design

- Data Management planning
  - Data types
    - Sizes, were to store, etc
  - **Metadata**
    - Study, Samples, Experiments, etc
    - Use standards!
      - *But not straight-forward...* >600 life science data standards
      - Ontologies & contolled vocabularies
      - http://biosharing.org

- *Data Management Plans*
  - Will become a standard part of the research funding application process
  - What will be collected?, Size?, Organized?, Documented?, Stored and preserved?, Disseminated?, Policies?, Budget?

Study & Analysis



Uppsala Multidisciplinary Center for Advanced Computational Science UPPMAX

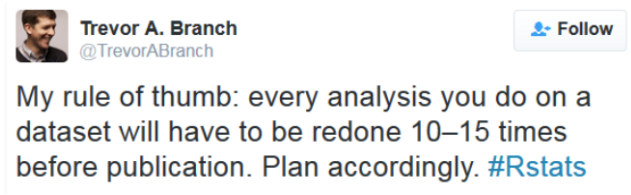
"milou"

"bianca"

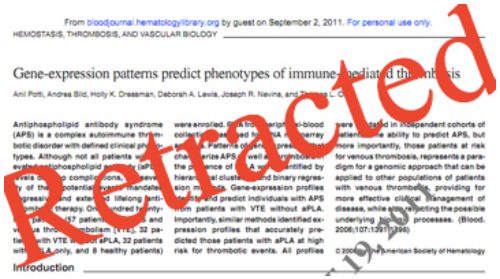
Human derived data

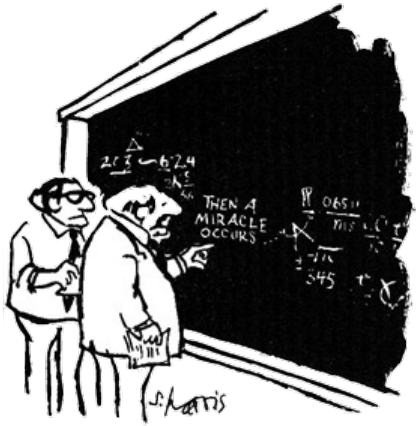
Study & Analysis

- Guiding principle
  - *“Someone unfamiliar with your project should be able to look at your computer files and understand in detail what you did and why.”*
- Research reality
  - *"Everything you do, you will have to do over and over again”*
    – Murphy’s law



- Structuring data for analysis
  - Poor organizational choices lead to significantly slower research progress.
  - It is critical to make results reproducible.





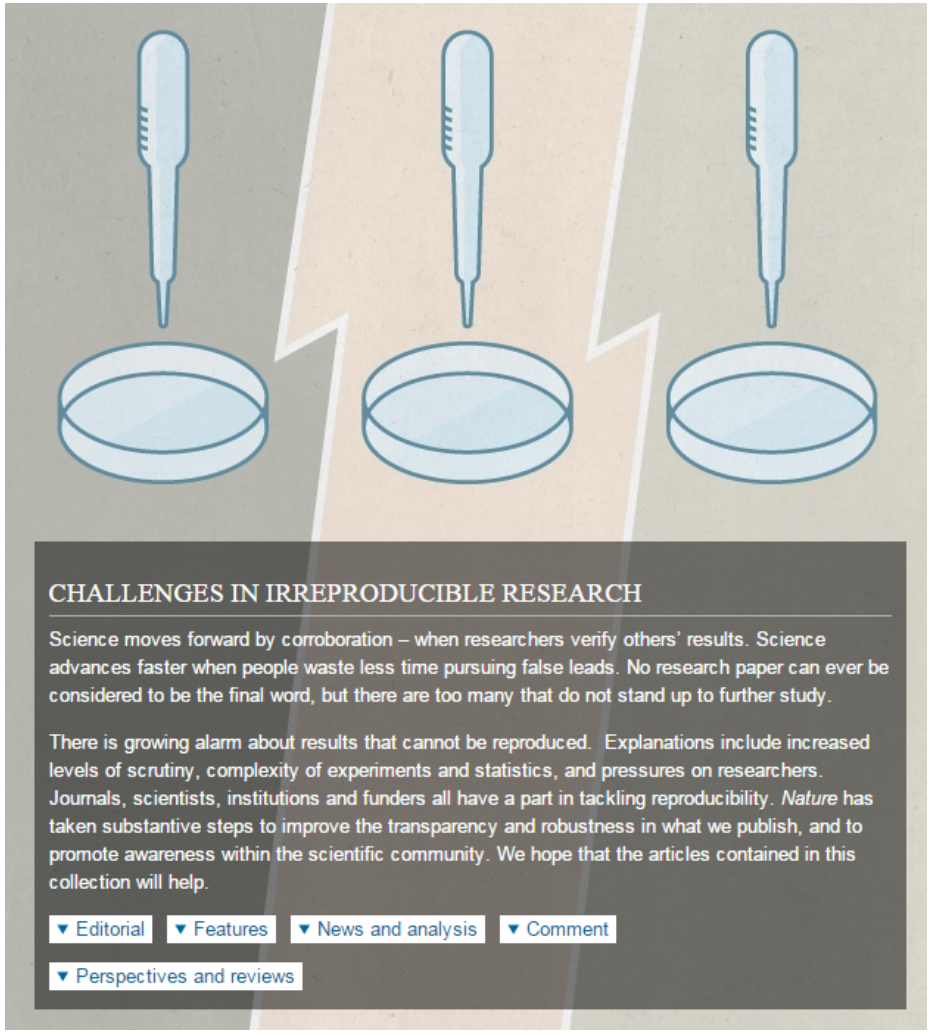
“I think you should be more explicit here in step two.”

A reproducibility crisis

Nature special issue

http://www.nature.com/news/reproducibility-1.17552

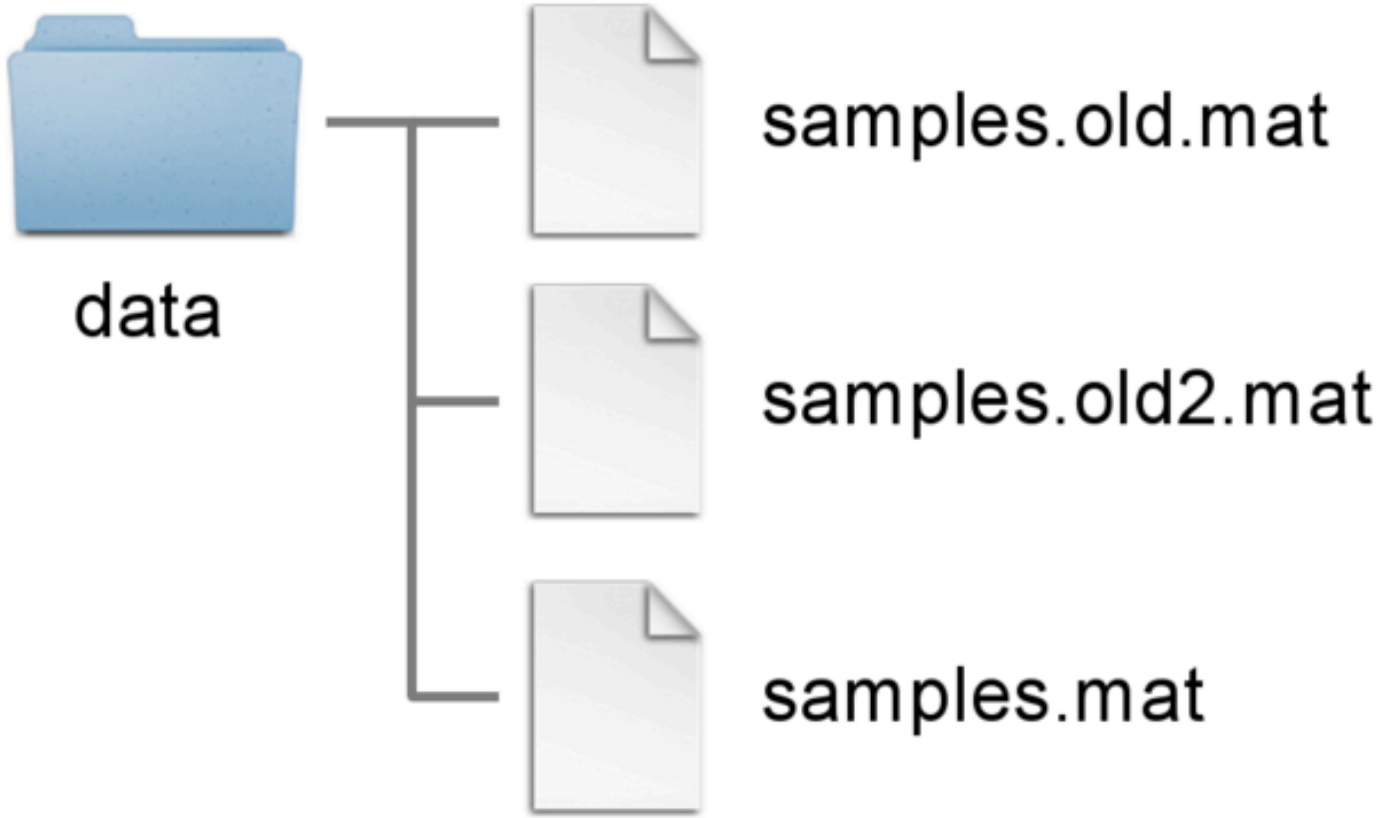
Several studies have shown alarming numbers of published papers that don't stand up to scrutiny



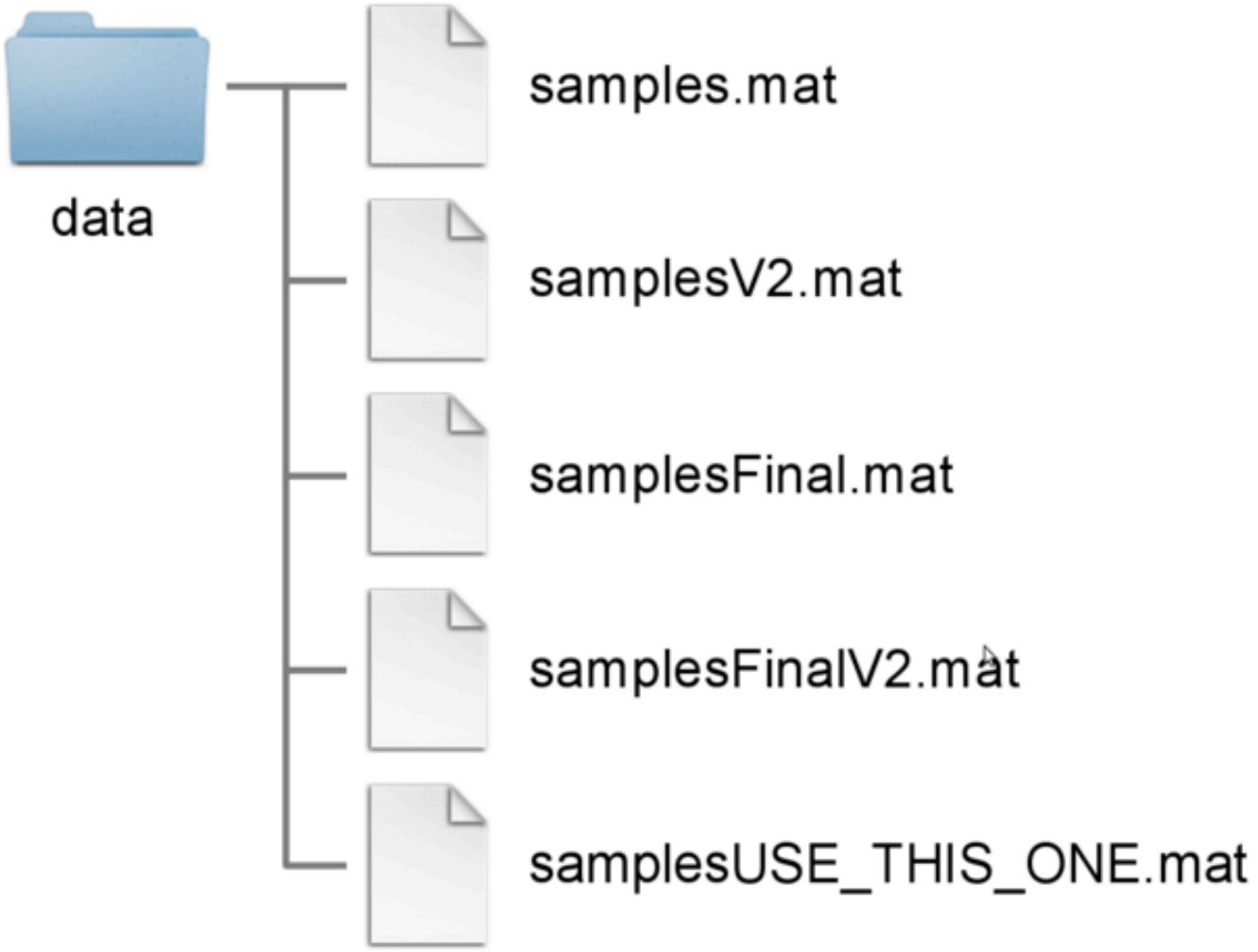
Organization - So far, so good…

Now what?

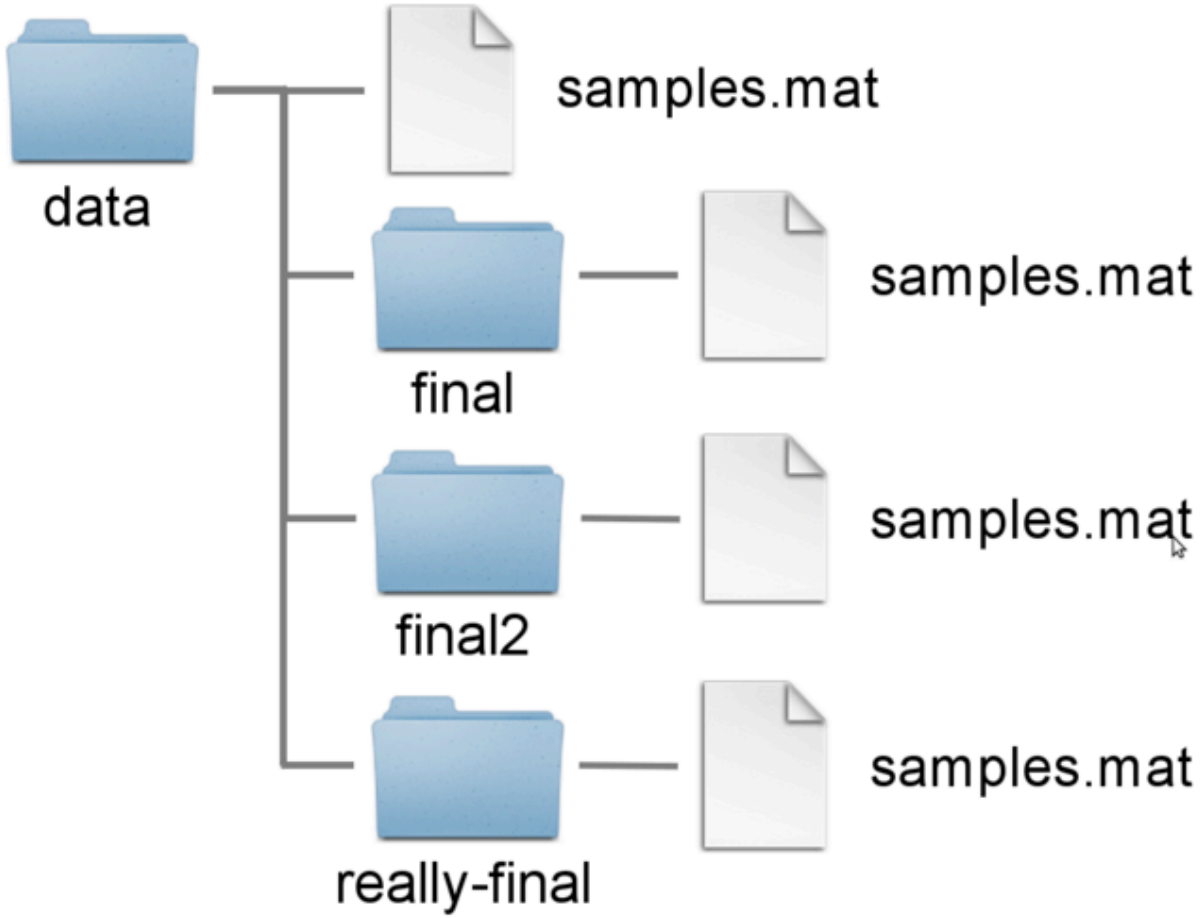
I guess this is alright



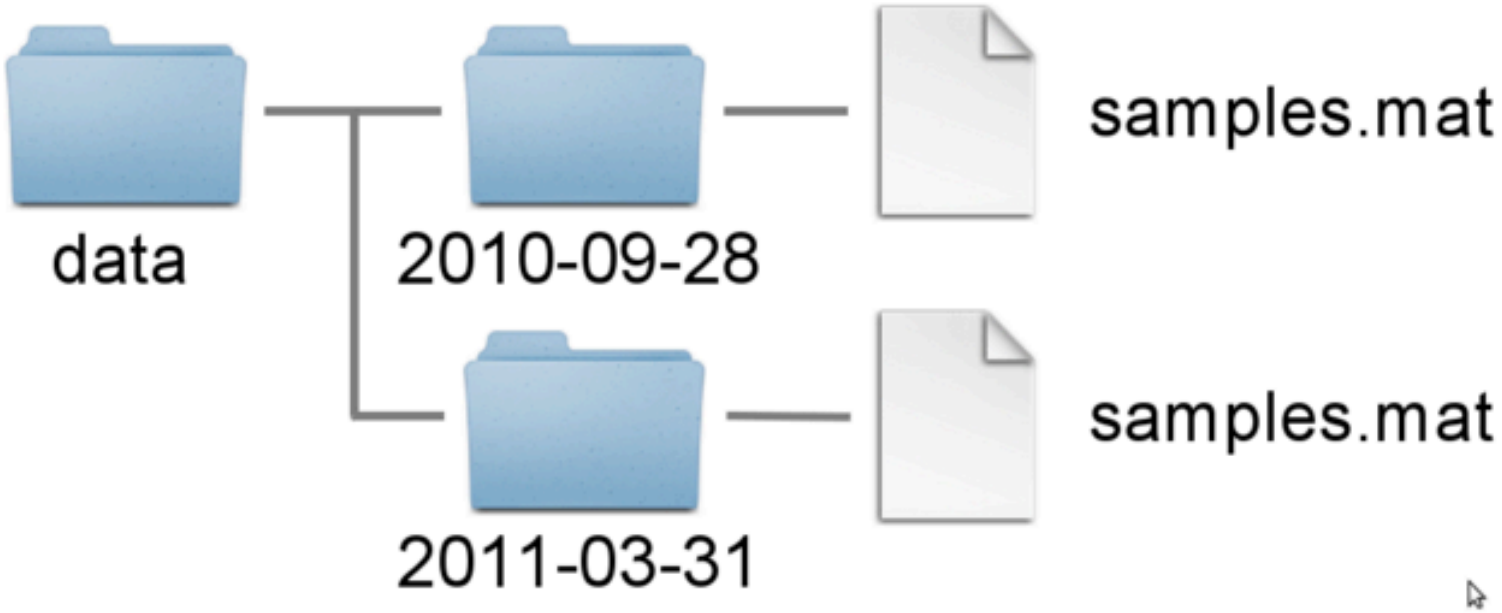
Which one is the most recent?



Another (bad) common approach



A possible solution

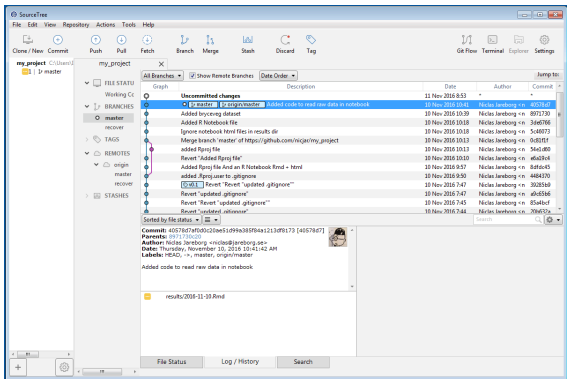


Suggested best practices

- There is a **folder for the raw data**, which do not get altered, or intermixed with data that is the result of manual or programmatic manipulation. I.e., derived data is kept separate from raw data, and **raw data are not duplicated**.
- **Code is kept separate from data**.
- Use a **version control system** (at least for code) – e.g. **git**
- There is a **scratch directory for experimentation**. Everything in the scratch directory can be deleted at any time without negative impact.
- There should be a **README in every directory**, describing the purpose of the directory and its contents.
- Use **non-proprietary formats** – *.csv* rather than *.xlsx*
- Etc…

Version control

- What is it?
  - A system that keeps records of your changes
  - Allows for collaborative development
  - Allows you to know who made what changes and when
  - Allows you to revert any changes and go back to a previous state
- Several systems available
  - Git, RCS, CVS, SVN, Perforce, Mercurial, Bazaar
  - Git
    - Command line & GUIs
    - Remote repository hosting
      - GitHub, Bitbucket, etc



Suggested best practices

- There is a **folder for the raw data**, which do not get altered, or intermixed with data that is the result of manual or programmatic manipulation. I.e., derived data is kept separate from raw data, and **raw data are not duplicated**.
- **Code is kept separate from data**.
- Use a **version control system** (at least for code) – e.g. **git**
- There is a **scratch directory for experimentation**. Everything in the scratch directory can be deleted at any time without negative impact.
- There should be a **README in every directory**, describing the purpose of the directory and its contents.
- Use **non-proprietary formats** – *.csv* rather than *.xlsx*
- Etc…

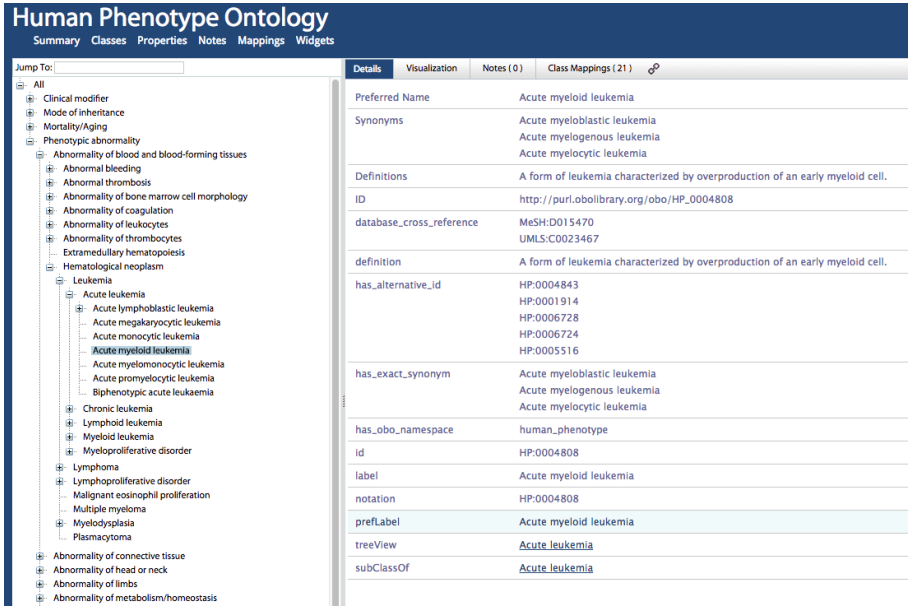
Non-proprietary formats

- A text-based format is more future-safe, than a proprietary binary format by a commercial vendor
- Markdown is a nice way of getting nice output from text.
  - Simple & readable formating
  - Can be converted to lots of different outputs
    - HTML, pdf, MS Word, slides etc

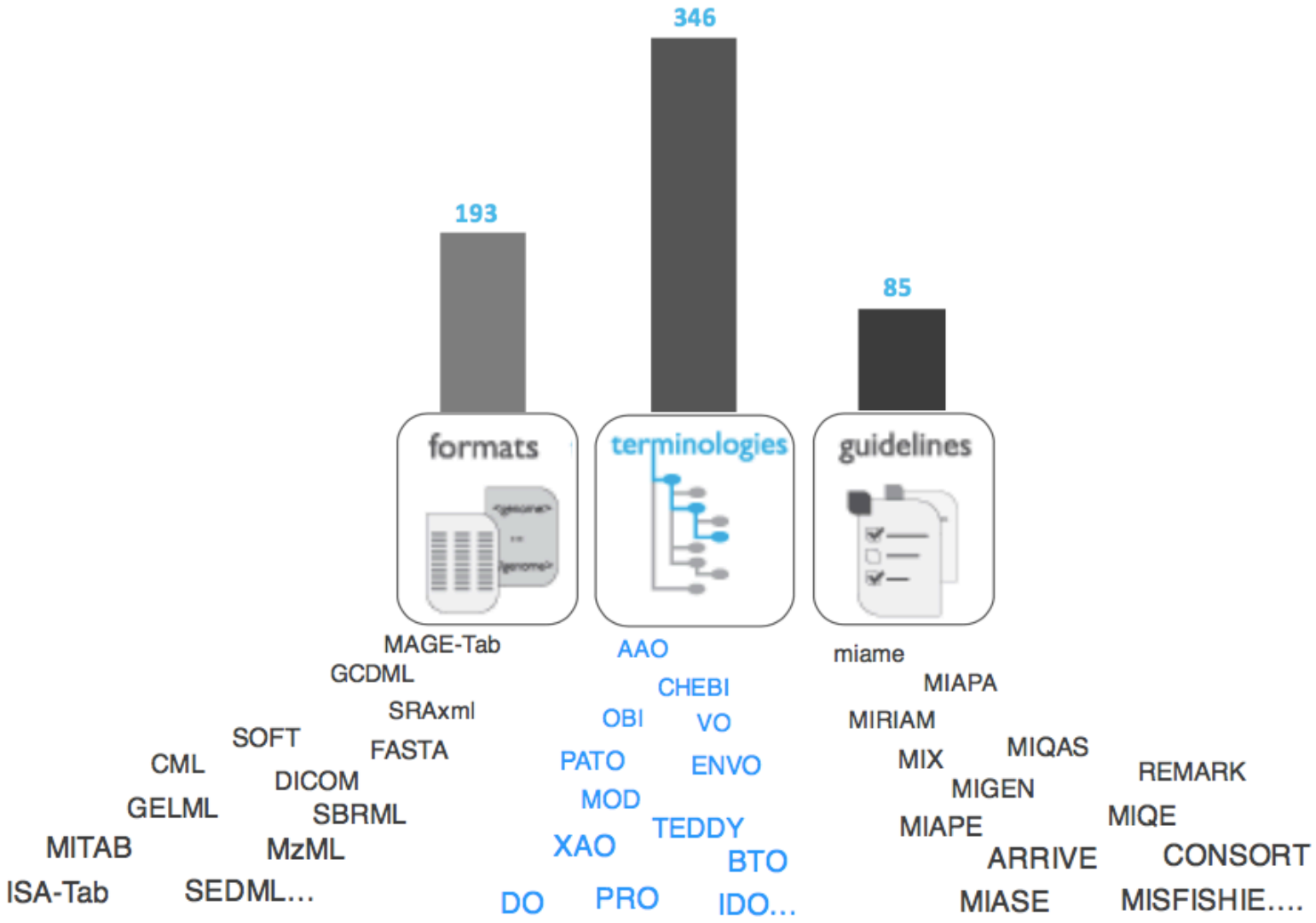
- *Never, never, never use **Excel** for scientific analysis!*
  - Script analysis – bash, python, R, …

Still missing something

- Need context → document **metadata**
  - How was the data generated?
  - From what was the data generated?
  - What where the experimental conditions?
  - Etc
- Use standards
  - Controlled vocabularies / Ontologies
  - *Not straight-forward…*

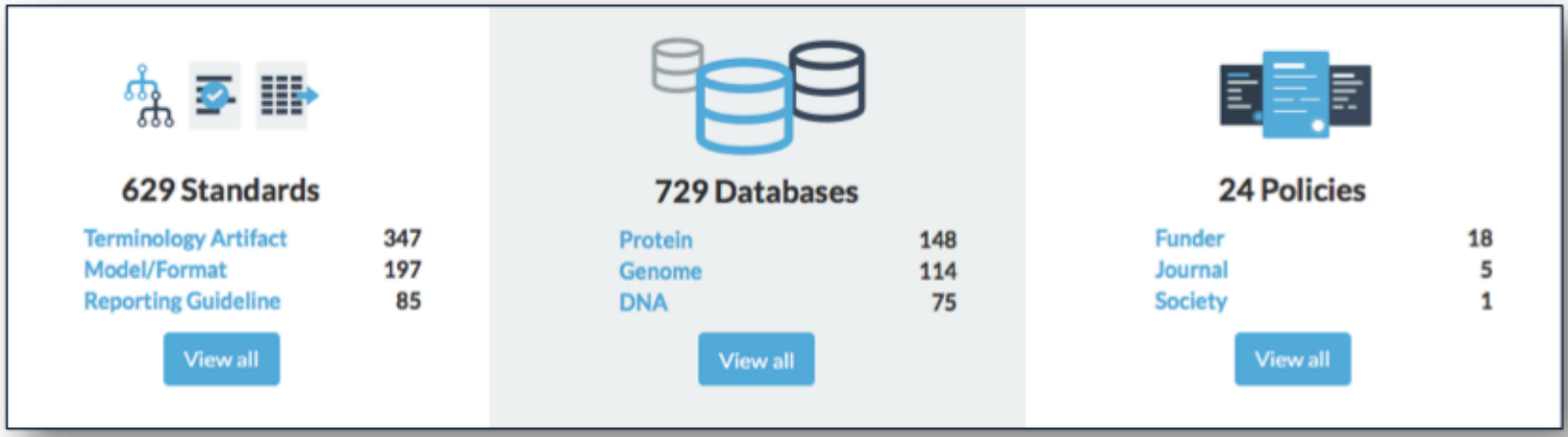


In the life sciences there are >600 *content standards*



biosharing.org
Information Resources

1,379 records and growing

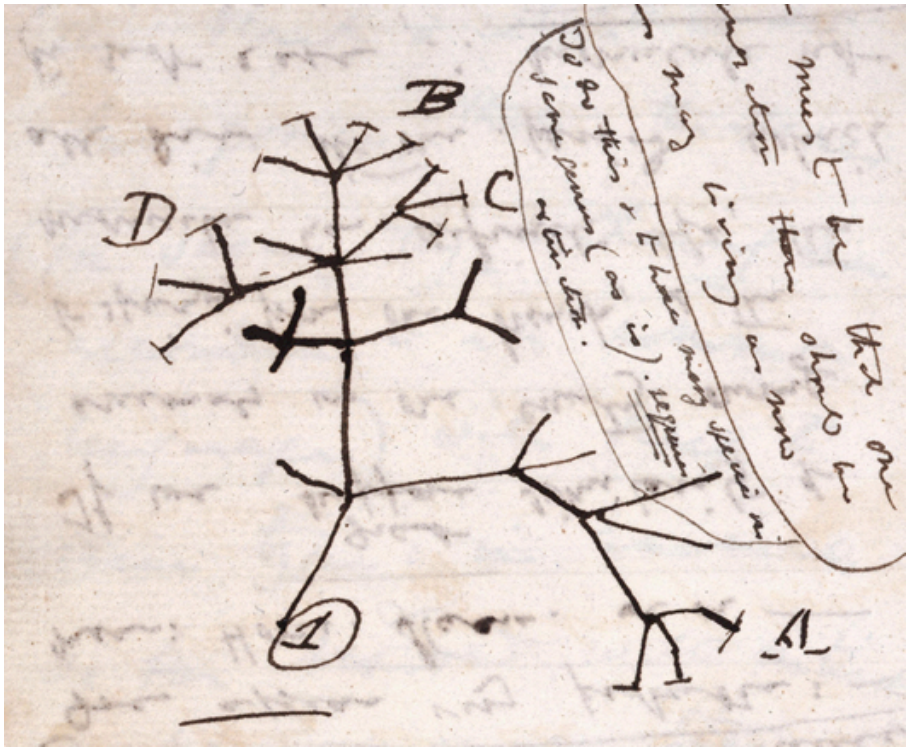


Mapping the landscape of 'standards' in the life sciences

A **web-based**, **curated** and **searchable registry** ensuring that standards and databases are *registered*, *informative* and *discoverable*; monitoring development and **evolution** of standards, their **use** in databases and adoption of both in data policies

Lab notebooks

- Why?
  - You have to understand what you have done
  - **Others should be able to reproduce what you have done**
    - Dated entries
    - Point to commands run and results generated

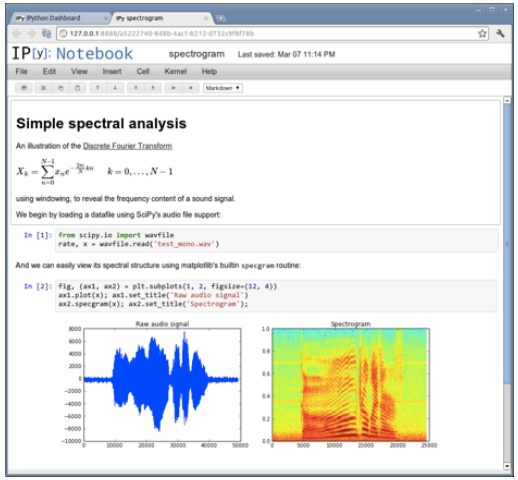


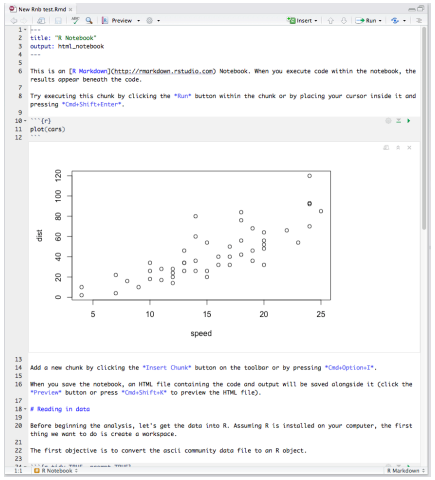
Lab notes – useful practices

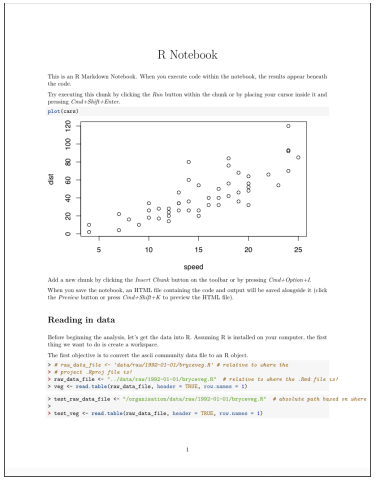
- Put in *results* directory
- Dated entries
- Entries relatively verbose
- Link to data and code (including versions)
- Embedded images or tables showing results of analysis done
- Observations, Conclusions, and ideas for future work
- Also document analysis that doesn't work, so that it can be understood why you choose a particular way of doing the analysis in the end

Where to keep notes

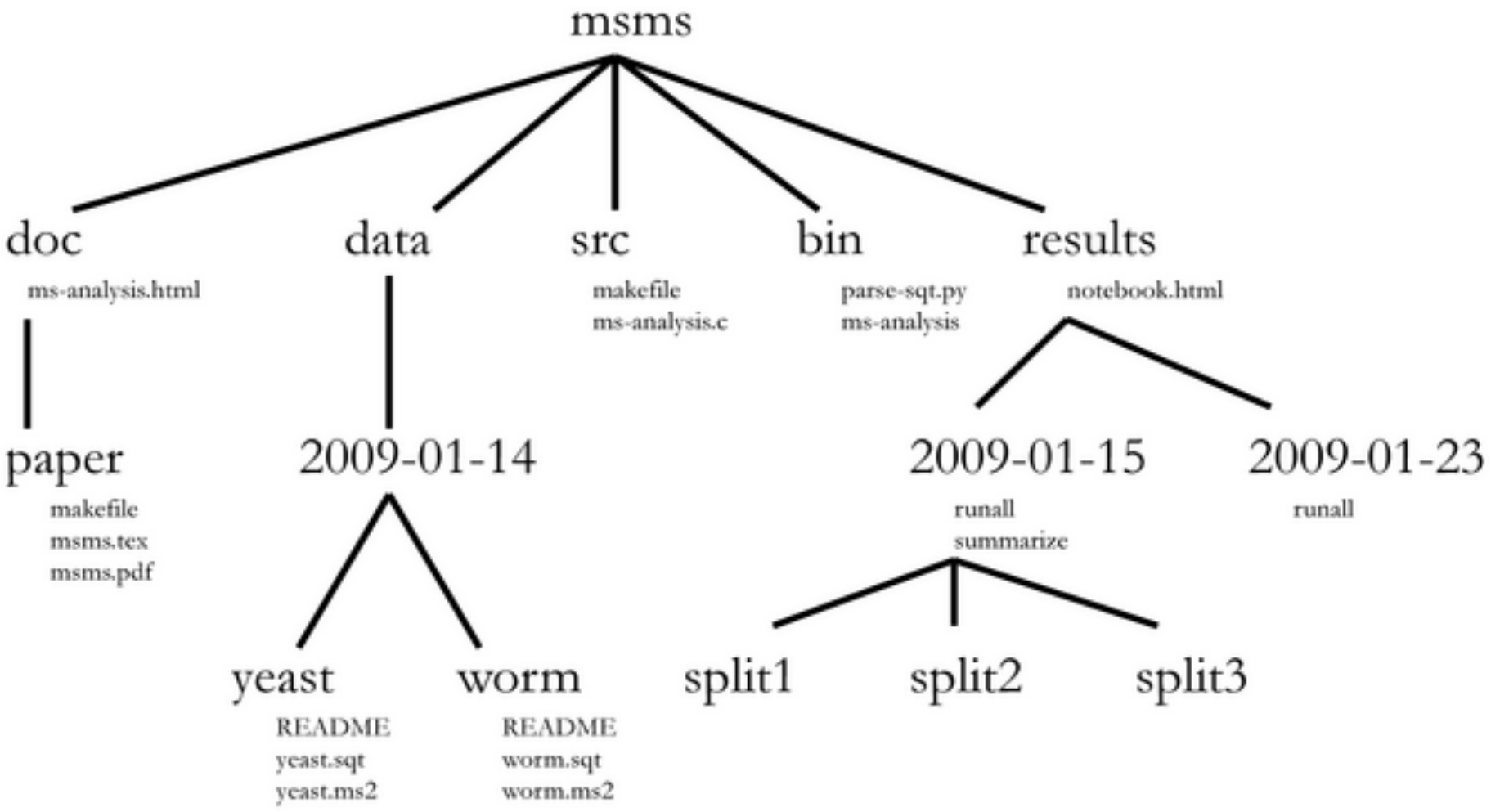
- Paper Notebook
- Word processor program / Text files
- Electronic Lab Notebooks
- 'Interactive' Electronic Notebooks
  - e.g. jupyther, R Notebooks in RStudio
  - Plain text - work well with version control
  - Embed and execute code
  - Convert to other output formats
    - html, pdf, word







Directory structure for a sample project



Noble WS (2009) A Quick Guide to Organizing Computational Biology Projects. PLoS Comput Biol 5(7): e1000424. doi:10.1371/journal.pcbi.1000424
http://journals.plos.org/ploscompbiol/article?id=info:doi/10.1371/journal.pcbi.1000424

Or like this…

```
├── bin  <---------------# Binary files and executables (jar files & proj-wide scripts etc)
├── conf <---------------# Project-wide configuraiotn
├── doc  <---------------# Any documents, such as manuscripts being written
├── experiments  <-------#  The main experiments folder
│   ├── 2000-01-01-exa <-# An example Experiment
│   │   ├── audit  <-----# Audit logs from workflow runs (higher level than normal logs)
│   │   ├── bin   <------# Experiment-specific executables and scripts
│   │   ├── conf  <------# Experiment-specific config
│   │   ├── data  <------# Any data generated by workflows
│   │   ├── doc   <------# Experiment-specific documents
│   │   ├── log   <------# Log files from workflow runs (lower level than audit logs)
│   │   ├── raw   <------# Raw-data to be used in the experiment (not to be changed)
│   │   ├── results  <---# Results from workflow runs
│   │   ├── run   <------# All files rel. to running experiment: Workflows, run confs/scripts...
│   │   └── tmp   <------# Any temporary files not supposed to be saved
├── raw  <---------------# Project-wide raw data
├── results  <-----------# Project-wide results
└── src  <---------------# Project-wide source code (that needs to be compiled)
```

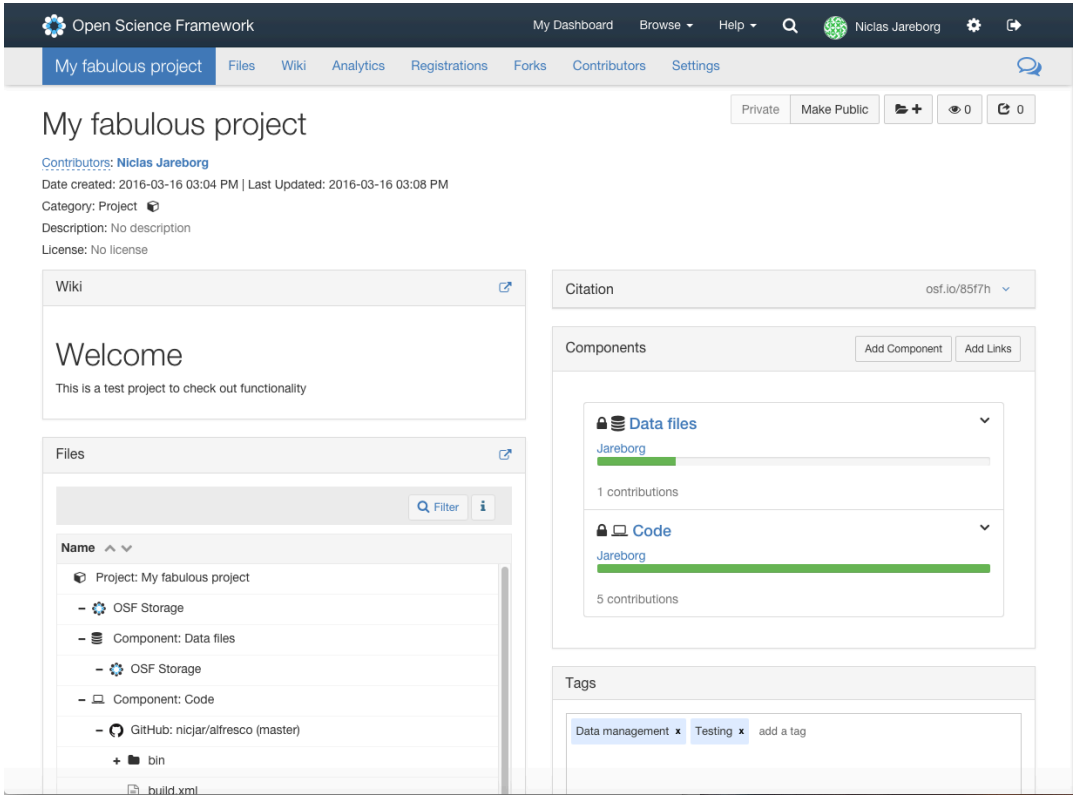
From Samuel Lampa's blog: http://bionics.it/posts/organizing-compbio-projects

Project organization

- There's no perfect set-up
  - Pick one! e.g.
    - https://github.com/chendaniely/computational-project-cookie-cutter
    - https://github.com/Reproducible-Science-Curriculum/rr-init
    - https://github.com/nylander/ptemplate
    - …
- Communicate structure to collaborators
- Document as you go
- Done well it might reduce post-project explaining

Project collaboration tools

- Open Science Framework – http://osf.io
  - Organize research project documentation and outputs
  - Control access for collaboration
  - 3rd party integrations
    - Google Drive
    - Dropbox
    - GitHub
    - External links
    - Etc
  - Persistent identifiers



NBIS
NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

Personal data

Personal data - Legislation

- Personal Data Act (*Personuppgiftslagen (PUL)*)
- Act concerning the Ethical Review of Research Involving Humans (*Lag om etikprövning av forskning som avser människor*)

Personal Data Act

- All kinds of information that is directly or indirectly referable to a natural person who is alive constitute personal data

- Sensitive data
  - It is **prohibited** to process personal data that discloses *ethnic origin*, *political opinions*, *religious or philosophical convictions, membership of trade unions*, as well as personal data relating to ***health*** or *sexual life*.
  - Sensitive personal data can be handled for **research purposes** if person has given **explicit consent**

- The Data Inspection Board (*Datainspektionen*) is the supervisory authority under the Personal Data Act

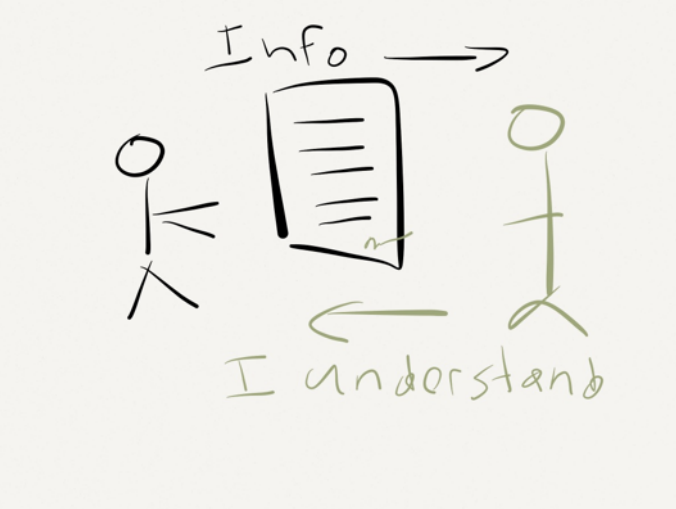
Personal Data Act – cont.

- The (legal) person that decides why and how personal data should be processed is called the **controller of personal data** (*personuppgiftsansvarig*)
  - e.g. the employing university
- The controller of personal data can delegate processing of personal data to a **personal data assistant** (*personuppgiftsbiträde*)
  - e.g. UPPMAX/Uppsala university
- A **personal data representative** (*personuppgiftsombud*) is a natural person who, on the assignment of the controller, shall ensure that personal data is processed in a lawful and proper manner
- Obligation to report handling of personal data to the Data Inspection Board
  - Or, notify the Board of the named representative

Act concerning the Ethical Review

- Research that concerns studies of biological material that has been taken from a living person and that can be traced back to that person may only be conducted if it has been approved subsequent to an ethical vetting

- Informed consent
  - The subject must be informed about the purpose or the research and the consequences and risks that the research might entail
  - The subject must consent



Genetic information

- The genetic information of an individual is personal data
  - **Sensitive** personal data (as it relates to health)
  - Even if *anonymized* / *pseudonymized*
  - In principle, **no** difference between WGS, Exome, Transcriptome or GWAS data

- Theoretically possible to identify the individual person from which the sequence was derived from the sequence itself
  - The more associated metadata there is, the easier this gets
  - Gymrek et al. “Identifying Personal Genomes by Surname Inference”. Science 339, 321 (2013); DOI:10.1126/science.1229566

- *“The controller is liable to implement technical and organizational measures to protect the personal data. The measures shall attain a suitable level of security.”*

Mosler

- e-Infrastructure for working with sensitive data for academic research
  - Owned by NBIS / Operated and hosted by UPPMAX
- Inspired by Norwegian solution (TSD)
- Designed to look like UPPMAX clusters
  - UPPMAX modules
  - UPPMAX can assist with installing custom tools

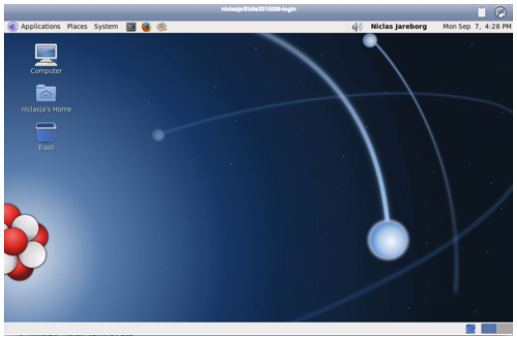
- Implementation project completed Nov 2015
- “Pilot-size system”
  - 24 nodes, 270 TB

- Provide users with a compute environment for sensitive data, with a *suitable level of security*

https://nbis.se/infrastructure/mosler.html

Mosler

- High-performance computing in a virtualized environment (OpenStack)
  - Each project environment is isolated from all other projects
    - Separated private networks and file systems
    - No internet access
    - No root access
- Only accessible over remote Linux desktop (ThinLinc) via a web dashboard
- 2-factor authentication for login
- Restricted data transfer in/out
  - Via a file gateway
  - Project members can transfer IN / only PI allowed to transfer out
  - Not possible to copy/paste out



- *Future*
  - SNIC Sens – "**bianca**"
    - Swedish Research Council funded
    - Being implemented at UPPMAX
    - In Pilot testing stage
    - Open early 2017

Tryggve – collaboration for sensitive biomedical data

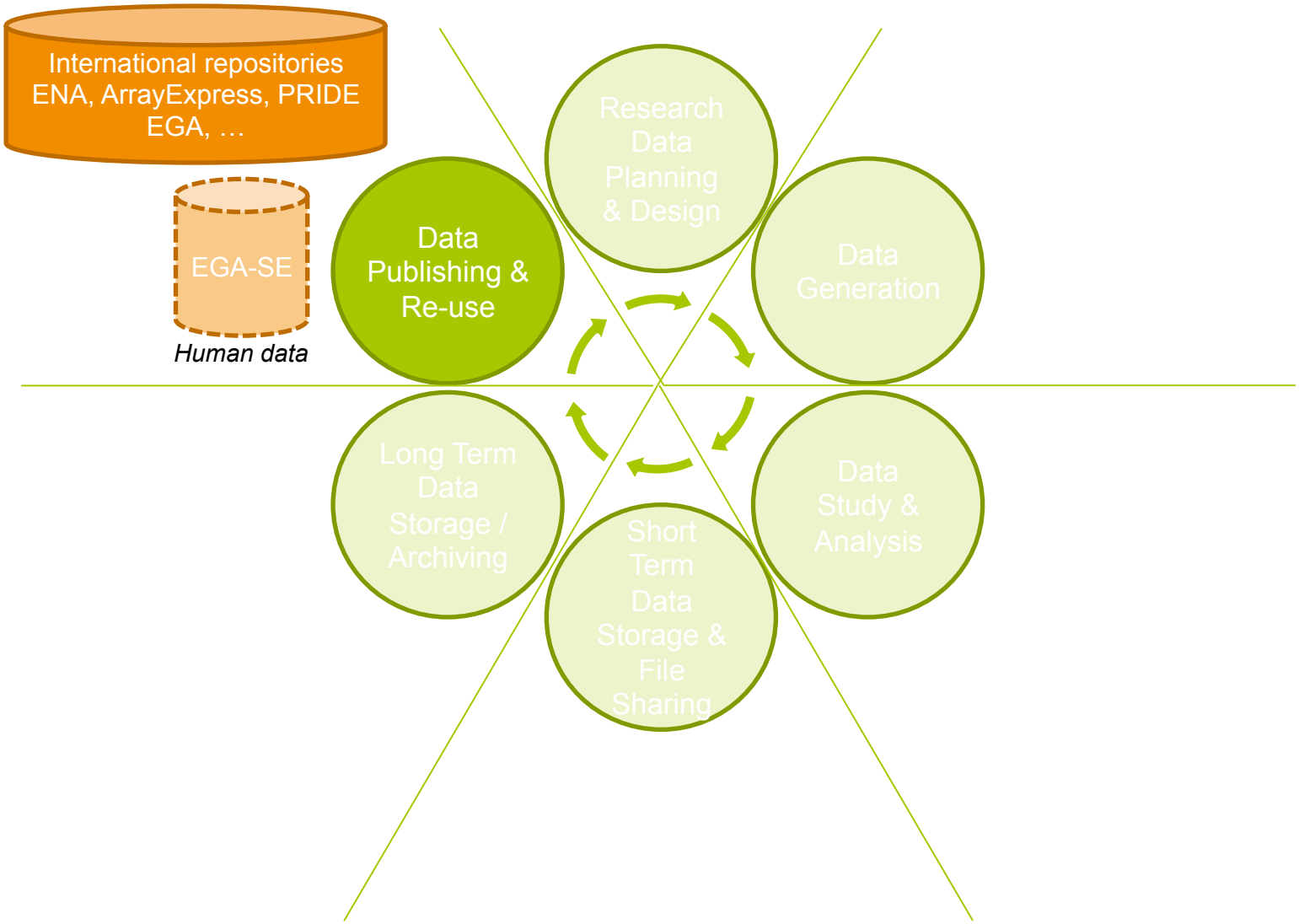
- Project aims to strengthen Nordic biomedical research by facilitating use of **sensitive data in cross-border projects**
- Collaborators and funders are NeIC and ELIXIR Nodes in Denmark, Finland, Norway and Sweden
- Project will build on strong existing capacities and resources in Nordic countries

Tryggve works on 6 themes

1. Technical development
   - Building blocks: Secure systems in Den, Fin, Nor & Swe
2. Interoperability of systems
   - Data transfer service – *sFTP beamer*
   - Portable software installations – *docker containers*
   - Shared computing resources – *Mosler-ePouta*
   - Investigate common authentication and authorization mechanisms
3. Process development
   - Knowledge-sharing (e.g. IT security, administrative processes, harmonizing user agreements)
   - Code of Conduct
4. Legal framework
   - Assessing relevant legislation
   - Analyzing legal requirements in use cases
5. **Use cases**
   - **Implement and support concrete use cases to facilitate cross-border research, and to connect project to actual user demands.**
6. Communication and outreach

https://wiki.neic.no/wiki/Tryggve_Getting_Started

Data Publishing & Re-use

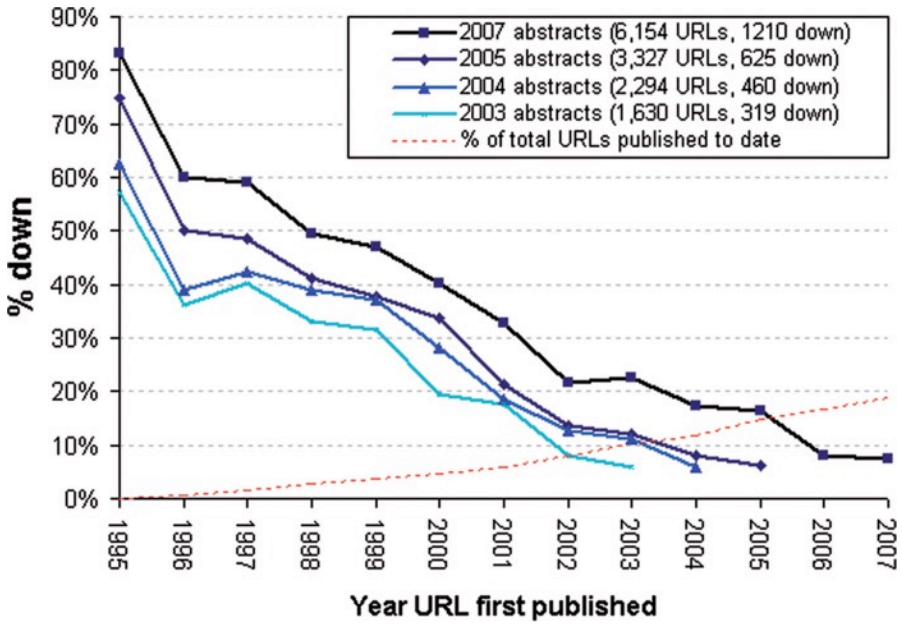


Data persistency issues

URL decay in MEDLINE—a 4-year follow-up study

Jonathan D. Wren*
+ Author Affiliations
*To whom correspondence should be addressed.

Received January 22, 2008.
Revision received March 11, 2008.
Accepted April 6, 2008.



Jonathan D. Wren Bioinformatics 2008;24:1381-1385

- Link rot – more 404 errors generated over time
- Reference rot* – link rot plus content drift i.e. webpages evolving and no longer reflecting original content cited

* Term coined by Hiberlink http://hiberlink.org

Data Publishing & Re-use

- *Research Data Publishing is a cornerstone of Open Access*

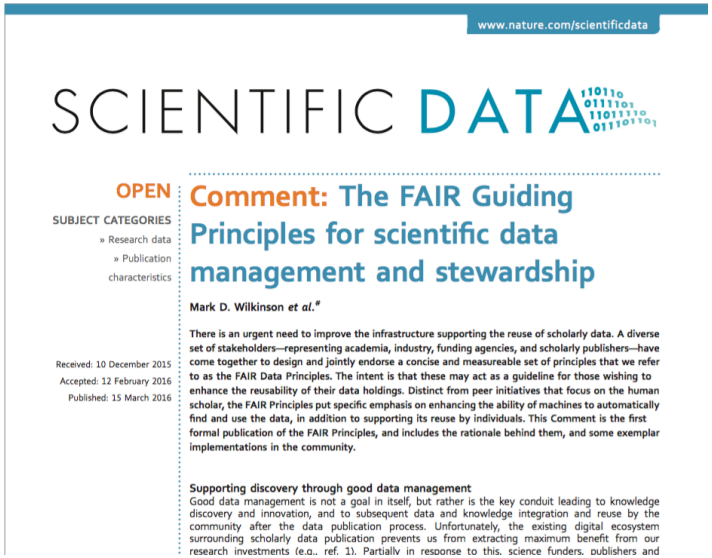
- Long-term storage
  - Data should not disappear
- Persistent identifiers
  - Possibility to refer to a dataset over long periods of time
  - Unique
  - e.g. DOIs (Digital Object Identifiers)

- Discoverability
  - Expose dataset metadata through search functionalities

FAIR

- To be useful for others data should be
  - ***FAIR*** - Findable, Accessible, Interoperable, and Reusable
    … for both Machines and Humans

Wilkinson, Mark et al. *"The FAIR Guiding Principles for scientific data management and stewardship"*. Scientific Data 3, Article number: 160018 (2016)
http://dx.doi.org/10.1038/sdata.2016.18



www.nature.com/scientificdata

SCIENTIFIC DATA

OPEN

SUBJECT CATEGORIES
» Research data
» Publication characteristics

Comment: The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson *et al.*#

There is an urgent need to improve the infrastructure supporting the reuse of scholarly data. A diverse set of stakeholders—representing academia, industry, funding agencies, and scholarly publishers—have come together to design and jointly endorse a concise and measureable set of principles that we refer to as the FAIR Data Principles. The intent is that these may act as a guideline for those wishing to enhance the reusability of their data holdings. Distinct from peer initiatives that focus on the human scholar, the FAIR Principles put specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals. This Comment is the first formal publication of the FAIR Principles, and includes the rationale behind them, and some exemplar implementations in the community.

Received: 10 December 2015
Accepted: 12 February 2016
Published: 15 March 2016

Supporting discovery through good data management

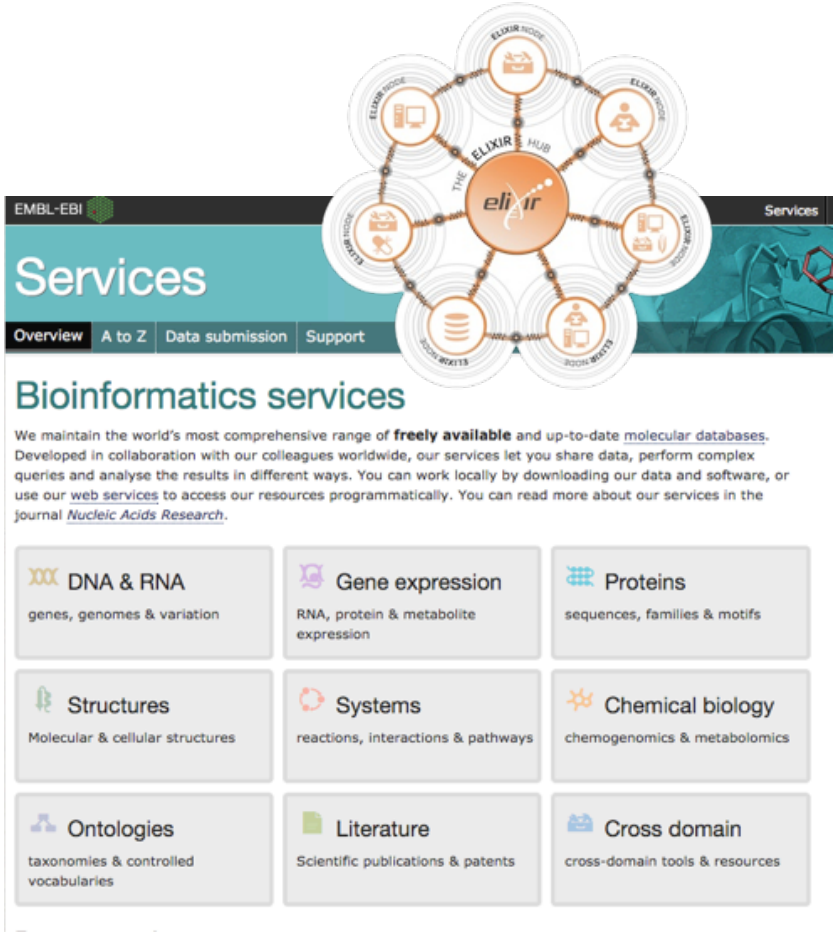
Good data management is not a goal in itself, but rather is the key conduit leading to knowledge discovery and innovation, and to subsequent data and knowledge integration and reuse by the community after the data publication process. Unfortunately, the existing digital ecosystem surrounding scholarly data publication prevents us from extracting maximum benefit from our research investments (e.g., ref. 1). Partially in response to this, science funders, publishers and

DOI: 10.1038/sdata.2016.18

International public repositories

- Best way to make data findable and re-usable
- Domain-specific metadata standards
- *Not always straight-forward!*

- **EBI** databases
  - ENA, Array Express, PRIDE etc



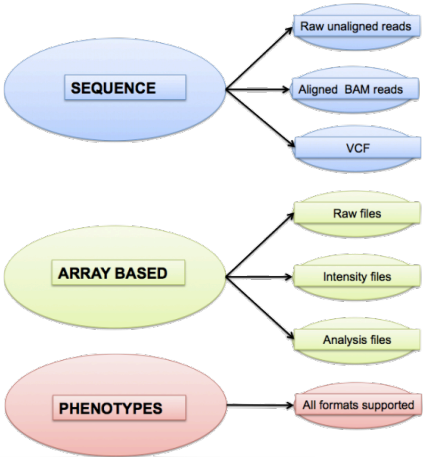
Surprisingly few submit to international repositories

- NIH funded research
  - Only 12% of articles from NIH funded research mention data deposited in international repositories
  - Estimated 200000+ “invisible” data sets / year

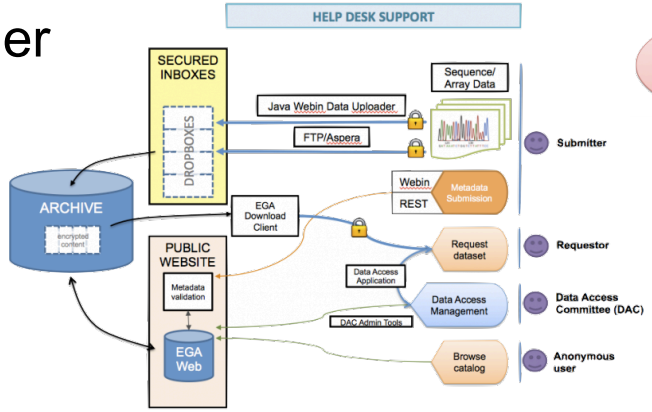
Read et al. “Sizing the Problem of Improving Discovery and Access to NIH-Funded Data: A Preliminary Study” (2015)
PLoS ONE 10(7): e0132735. doi: 10.1371/journal.pone.0132735

What about sensitive data?

- **EGA** – European Genome-phenome Archive
  - Repository that promotes the distribution and sharing of genetic and phenotypic data consented for specific approved uses but not fully open, public distribution.
  - All types of sequence and genotype experiments, including case-control, population, and family studies.

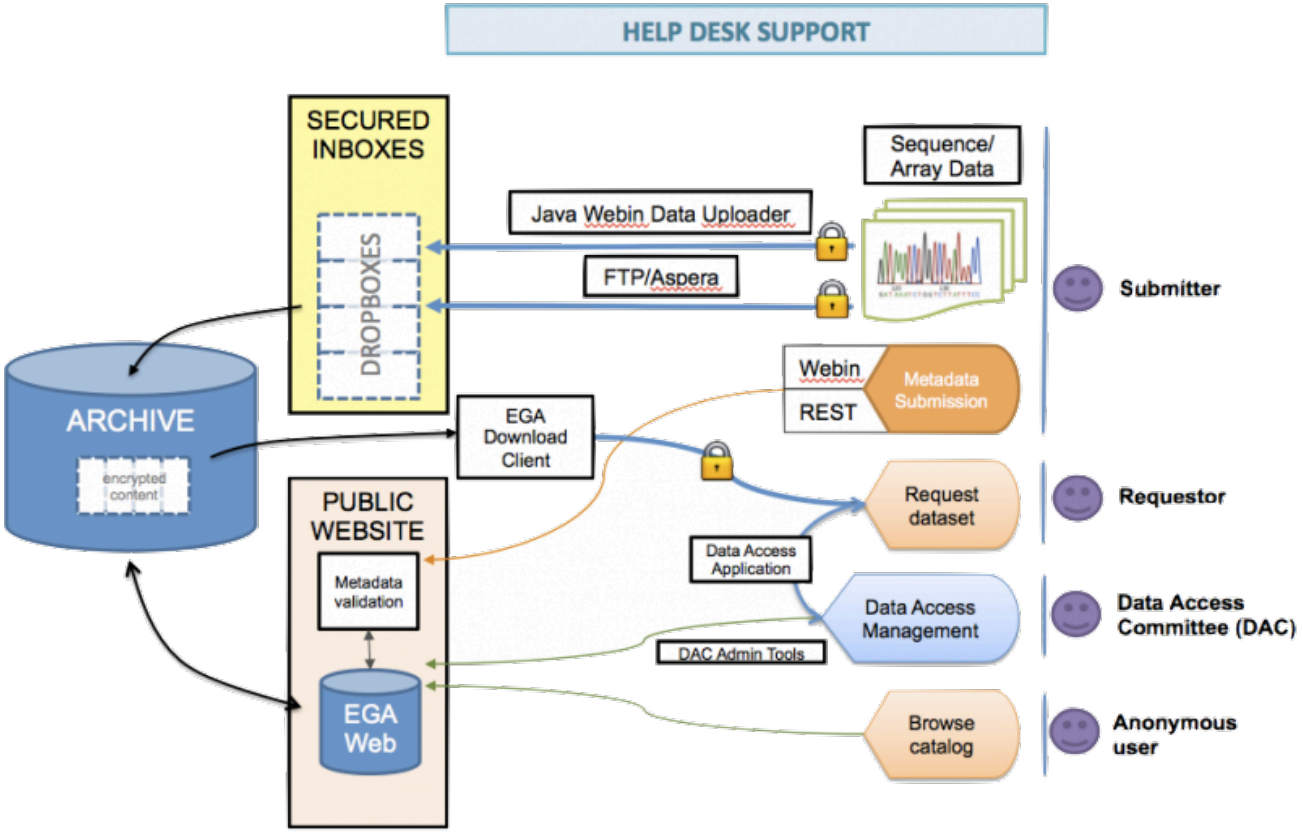


- Data Access Agreement
  - Defined by the data owner
- Data Access Committee – DAC
  - Decided by the data owner



What is different from ENA?

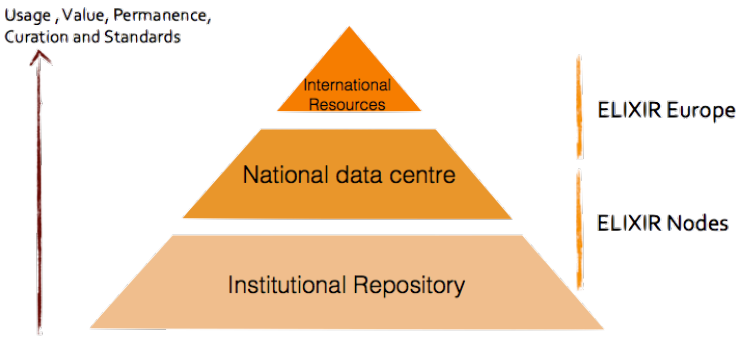
- Data Access Agreement
  - Defined by the data owner
- Data Access Committee – DAC
  - Decided by the data owner



Local EGA-SE

- Federated EGA
  - Metadata stored centrally
  - Data stored nationally/regionally/locally

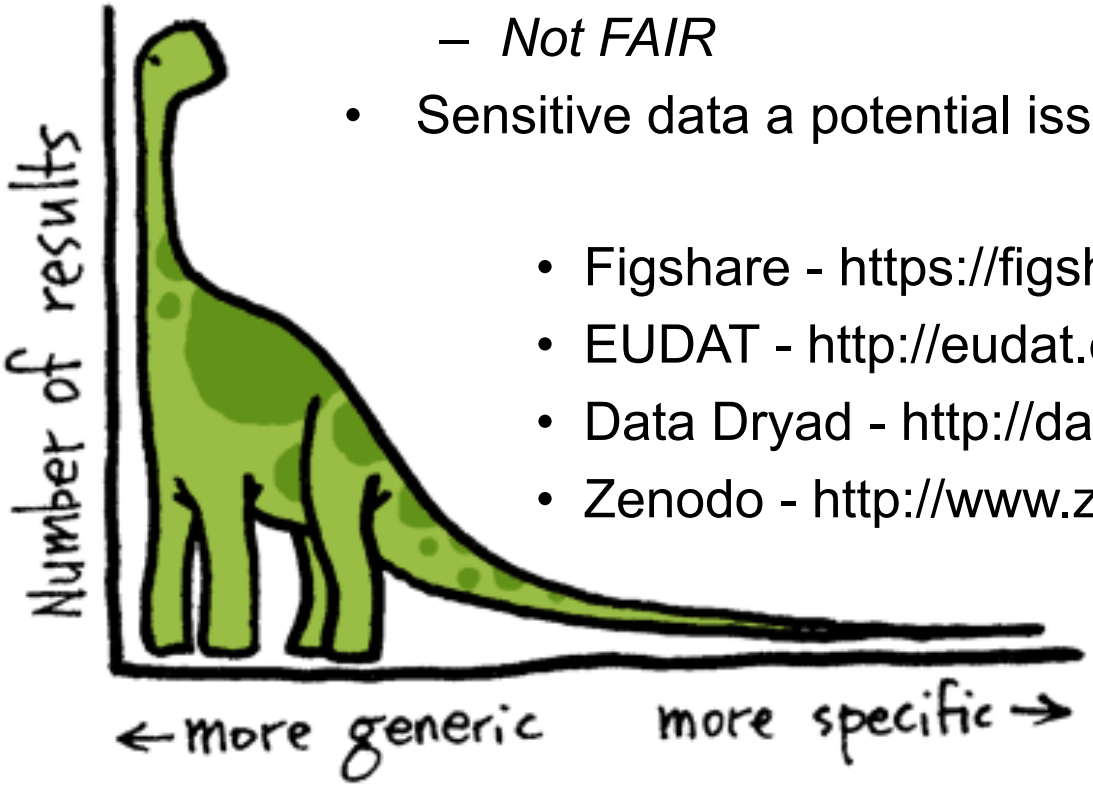
- ELIXIR-Excelerate WP9 (& WP10) activity



- Establish easy-to-use submission route for human sequence data produced by NGI

“Long-tail data” repositories

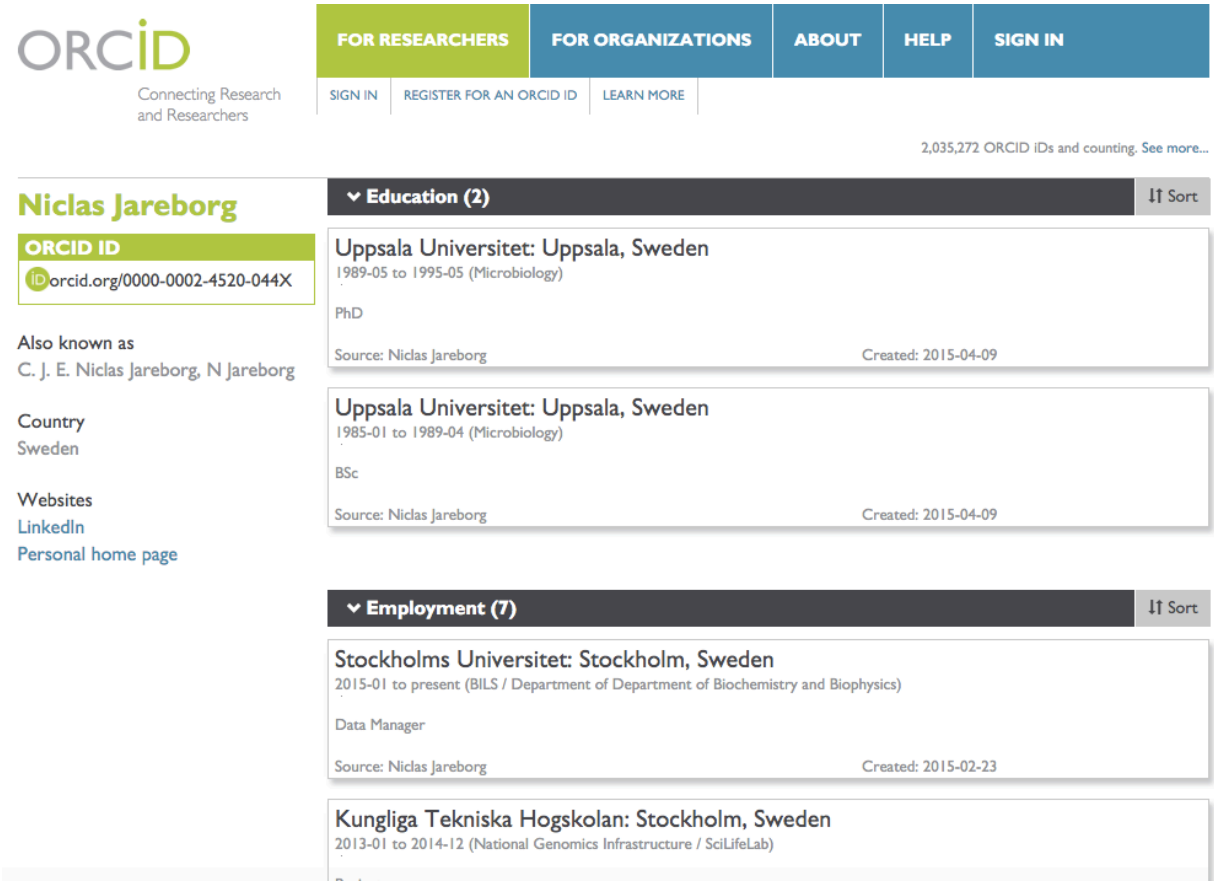
- Research data that doesn’t fit in structured data repositories
- Data publication – persistent identifiers
- Metadata submission – not tailored to Life Science
  - *Affects discoverability*
  - *Not FAIR*
- Sensitive data a potential issue



- Figshare - https://figshare.com/
- EUDAT - http://eudat.eu/
- Data Dryad - http://datadryad.org/
- Zenodo - http://www.zenodo.org/

Who are you?

- ORCID is an open, non-profit, community-driven effort to create and maintain a registry of unique researcher identifiers and a transparent method of linking research activities and outputs to these identifiers.
- http://orcid.org



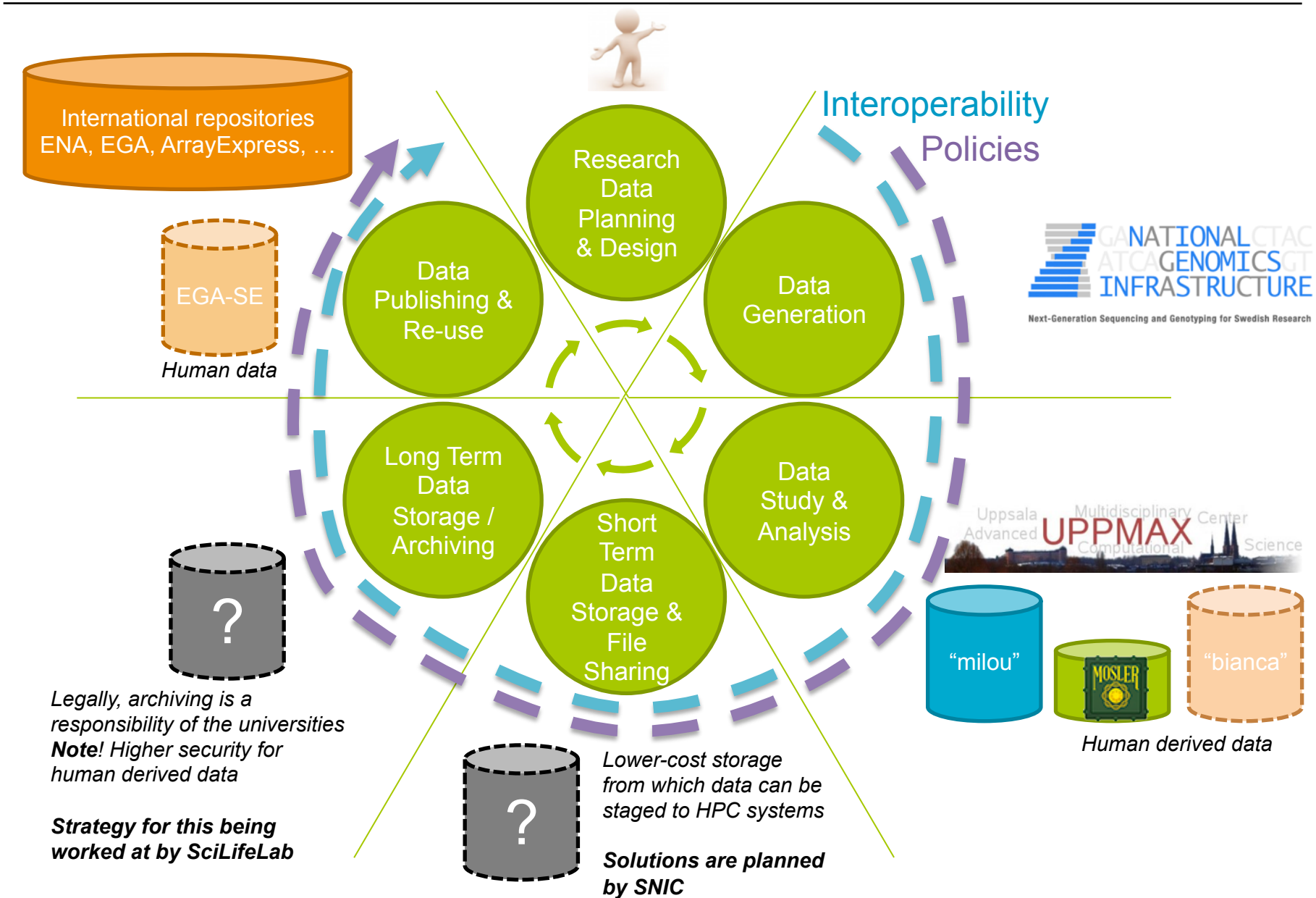
NBIS Data Management support

- Project planning
  - Metadata
  - File formats
  - Licensing
  - *Data Management Plans*

- Data analysis

- Data publication and submission
  - Automate submissions to public repositories
  - Metadata
  - Licensing

Gaps in the NGS Data Life Cycle



Source Acknowledgements

- Research Data Management, EUDAT - http://hdl.handle.net/11304/79db27e2-c12a-11e5-9bb4-2b0aad496318
- Barend Mons – FAIR Data
- Antti Pursula – Tryggve https://wiki.neic.no/wiki/Tryggve
- Noble WS (2009) A Quick Guide to Organizing Computational Biology Projects. PLoS Comput Biol 5(7): e1000424. doi:10.1371/journal.pcbi.1000424
- Samuel Lampa - http://bionics.it/posts/organizing-compbio-projects
- Reproducible Science Curriculum – https://github.com/Reproducible-Science-Curriculum/rr-init