
Project Data Management

Niclas Jareborg, NBIS
niclas.jareborg@bils.se

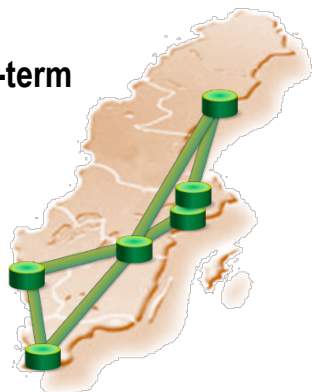
Data Management Seminar, 2016-03-18





Four Facilities

- **Short- and medium-term support**
Wide competence in bioinformatics distributed at all large university cities
- **Long-term support**
Large collaborative projects selected by scientific ranking
- **Compute and storage**
Providing computational and storage resources for bioinformatics, especially next-generation sequencing
- **Systems biology**
Network analyses and Integrative bioinformatics



Support and Infrastructure

- Total 80+ FTEs
- "Drop in" every week at all six sites – check calendar at <http://nbis.se> (<http://bils.se>)
- Study design consultation (free)
- Short- & Medium-term support (user fee 800 SEK per hour; first 8 hours per PI and year for free; application rounds every 2 weeks)
- Long-term support (≤ 500 h; free: scientific evaluation)
- Compute and storage (free; SNIC/UPPMAX)

Multiple areas of expertise

- Genomics/NGS/Metagenomics
- Genome annotation and assembly
- MS-Proteomics
- Systems biology and Metabolomics
- Protein bioinformatics
- Biostatistics
- Systems development
- Data publication
- Data management plans

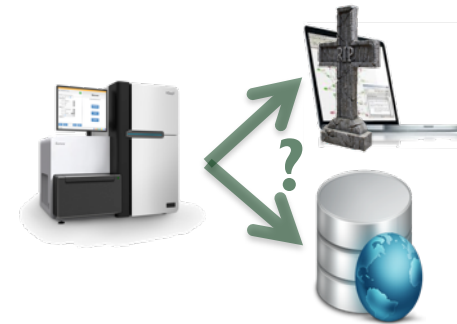
Coordination of the Swedish node in the European infrastructure for biological data **ELIXIR**



Funding from:

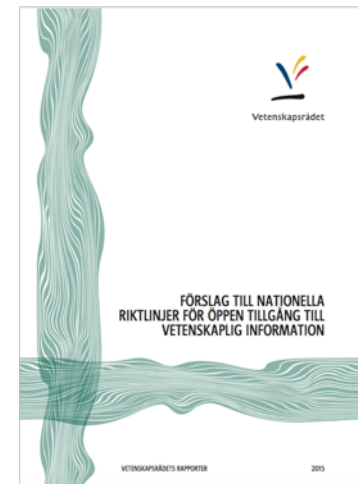
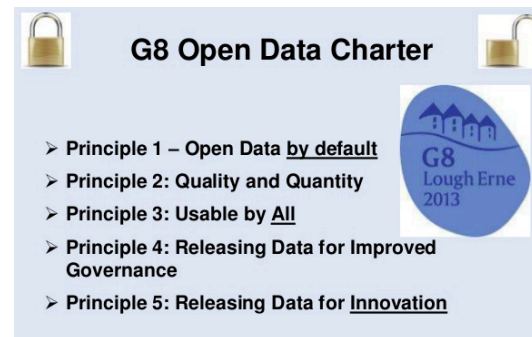


- To make your research easier!
- To stop yourself drowning in irrelevant stuff
- In case you need the data later
- To avoid accusations of fraud or bad science
- To share your data for others to use and learn from
- To get credit for producing it
- Because funders or your organisation require it



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

- *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 last year

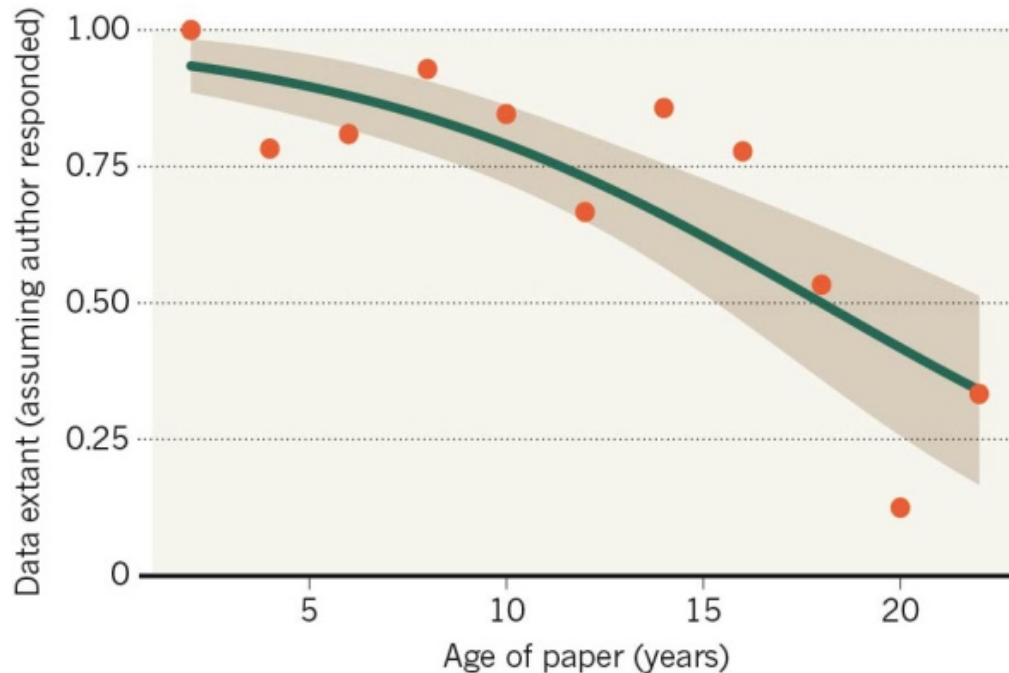


- Democracy and transparency
 - Publicly funded research data should be accessible to all
 - Published results and conclusions should be possible to check by others
- Research
 - Enables others to combine data, address new questions, and develop new analytical methods
 - Reduce duplication and waste
- 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



MISSING DATA

As research articles age, the odds of their raw data being extant drop dramatically.



Nature news, 19 December 2013



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

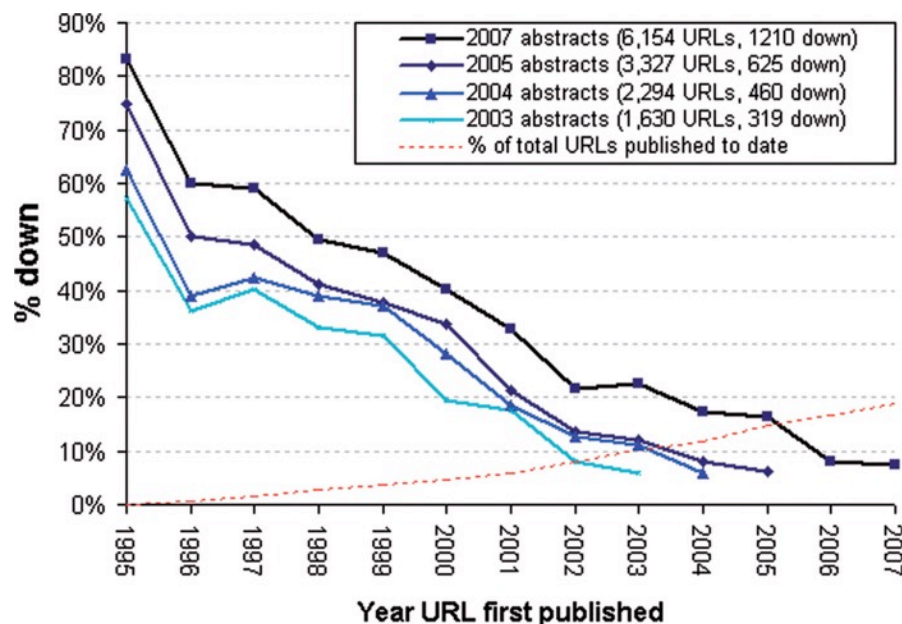
URL decay in MEDLINE—a 4-year follow-up study

Jonathan D. Wren*

+ Author Affiliations

*To whom correspondence should be addressed.

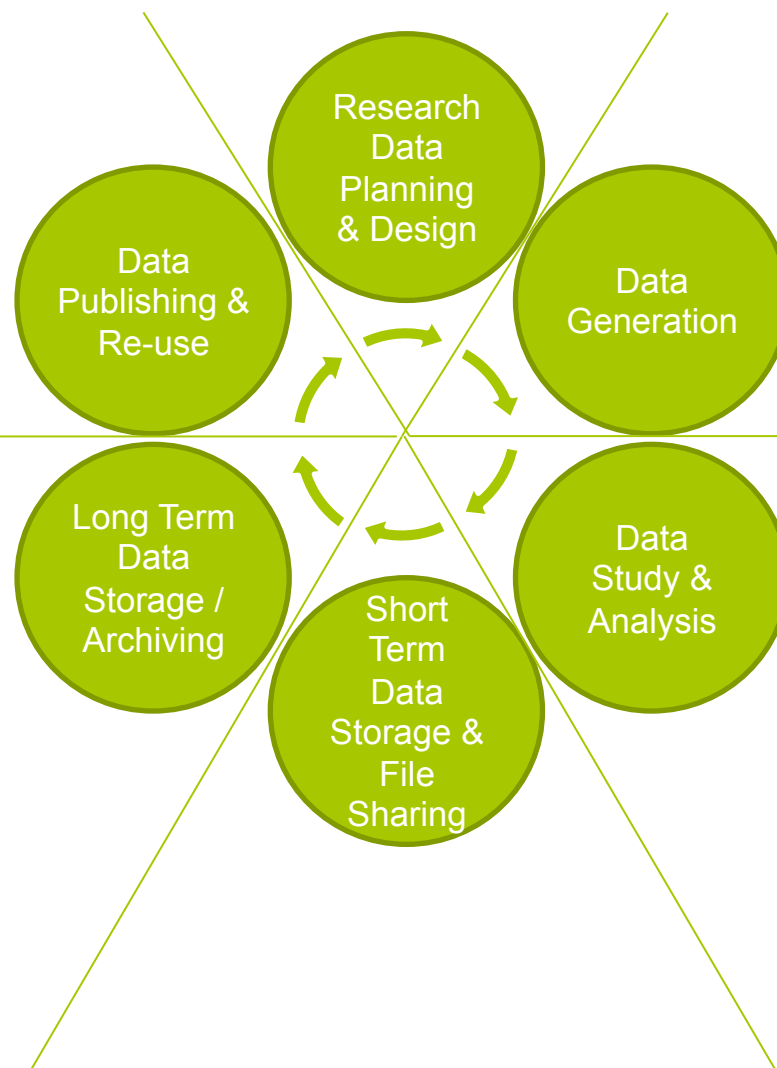
Received January 22, 2008.
Revision received March 11, 2008.
Accepted April 6, 2008.

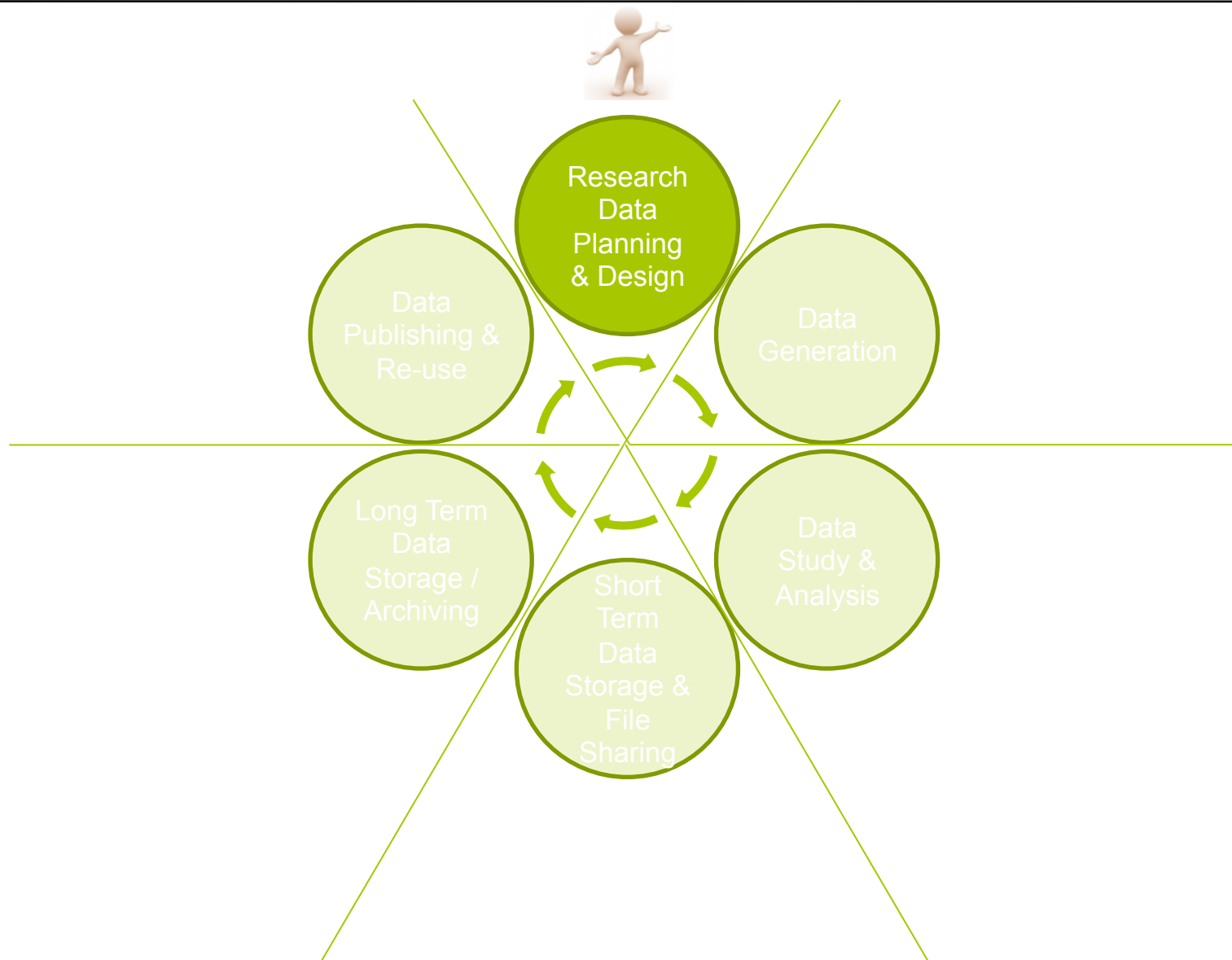


Jonathan D. Wren Bioinformatics 2008;24:1381-1385

- 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

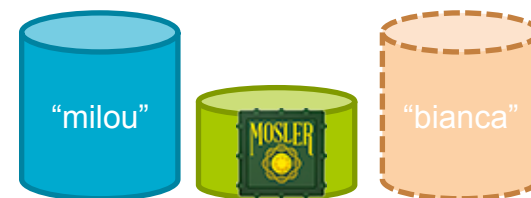
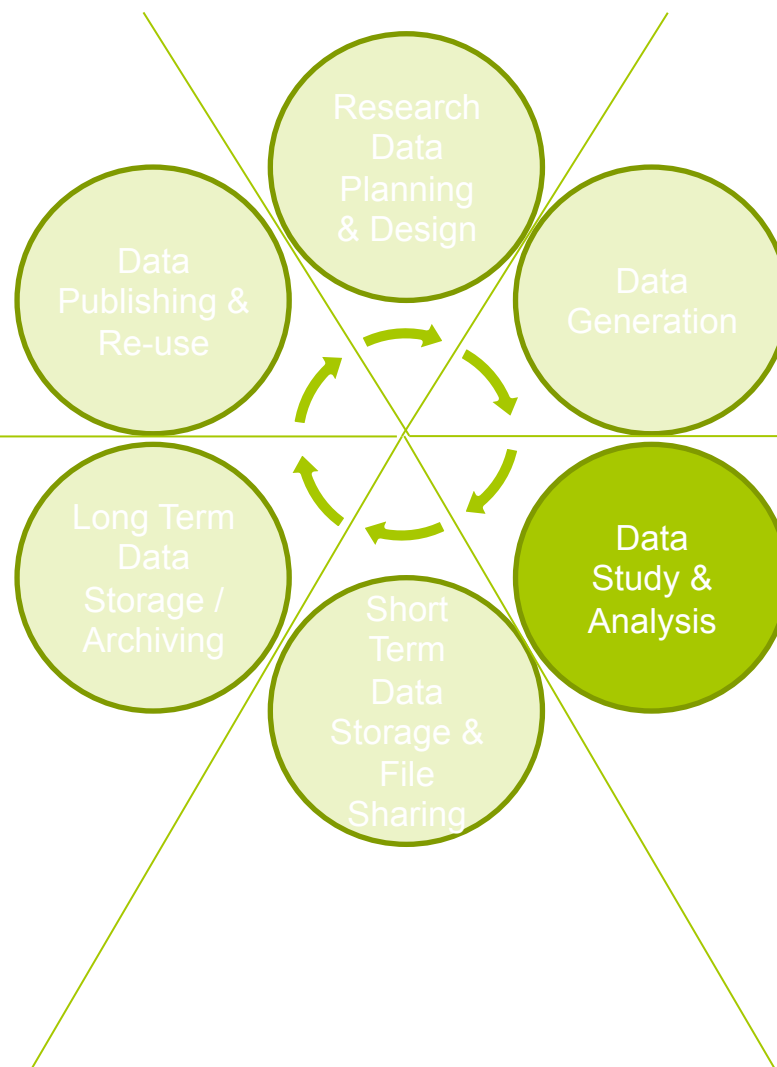
* Term coined by Hiberlink <http://hiberlink.org>





- Data 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?, Organized?, Documented?, Stored and preserved?, Disseminated?, Policies?, Budget?





Human derived data

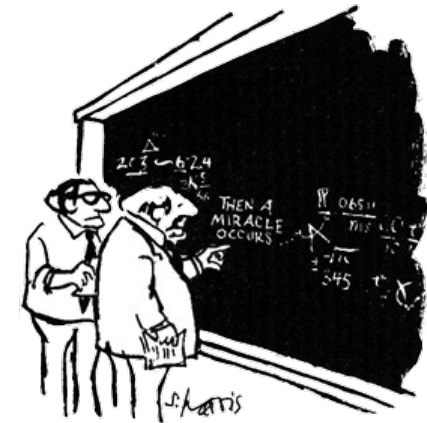
- Principles

- “Someone unfamiliar with your project should be able to look at your computer files and understand in detail what you did and why.”
- “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.

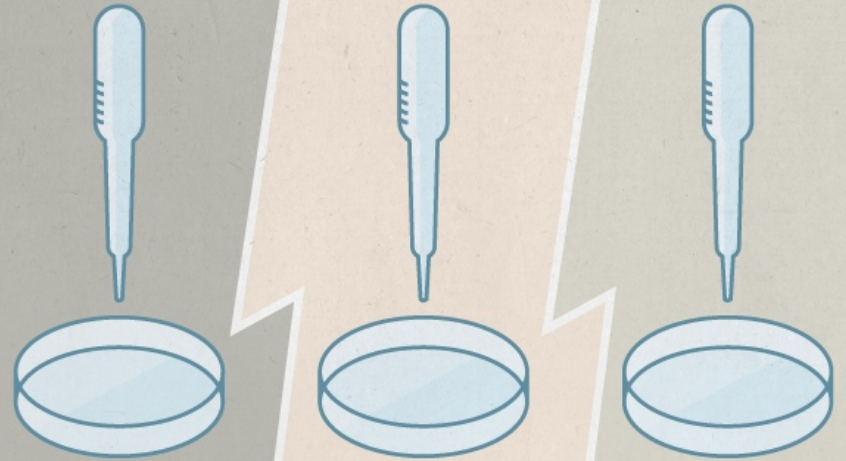


“I think you should be more explicit here in step two.”

Nature special issue

[http://www.nature.com/
news/reproducibility-1.17552](http://www.nature.com/news/reproducibility-1.17552)

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



CHALLENGES IN IRREPRODUCIBLE RESEARCH

Science moves forward by corroboration – when researchers verify others' results. Science advances faster when people waste less time pursuing false leads. No research paper can ever be considered to be the final word, but there are too many that do not stand up to further study.

There is growing alarm about results that cannot be reproduced. Explanations include increased levels of scrutiny, complexity of experiments and statistics, and pressures on researchers. Journals, scientists, institutions and funders all have a part in tackling reproducibility. *Nature* has taken substantive steps to improve the transparency and robustness in what we publish, and to promote awareness within the scientific community. We hope that the articles contained in this collection will help.

▼ Editorial ▼ Features ▼ News and analysis ▼ Comment

▼ Perspectives and reviews

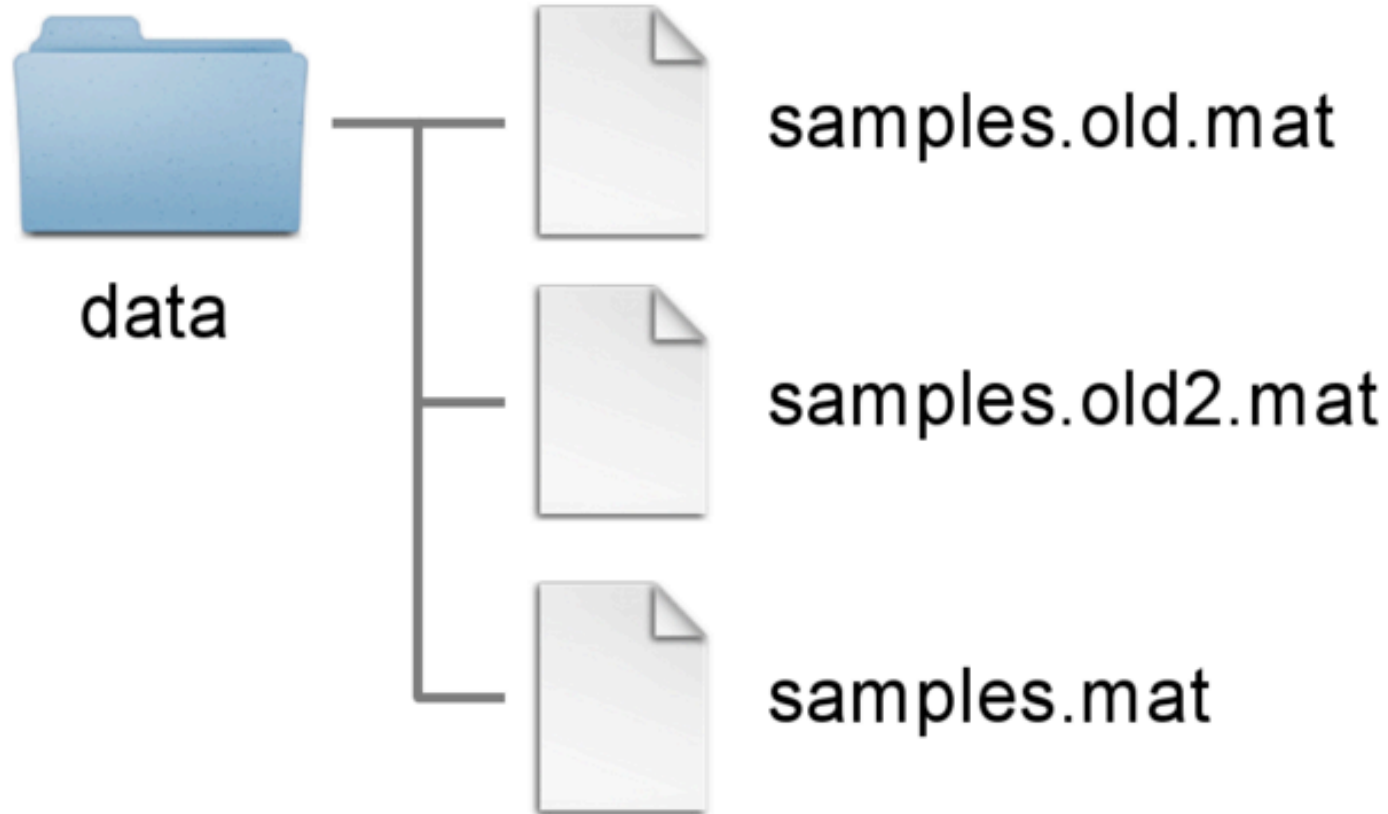


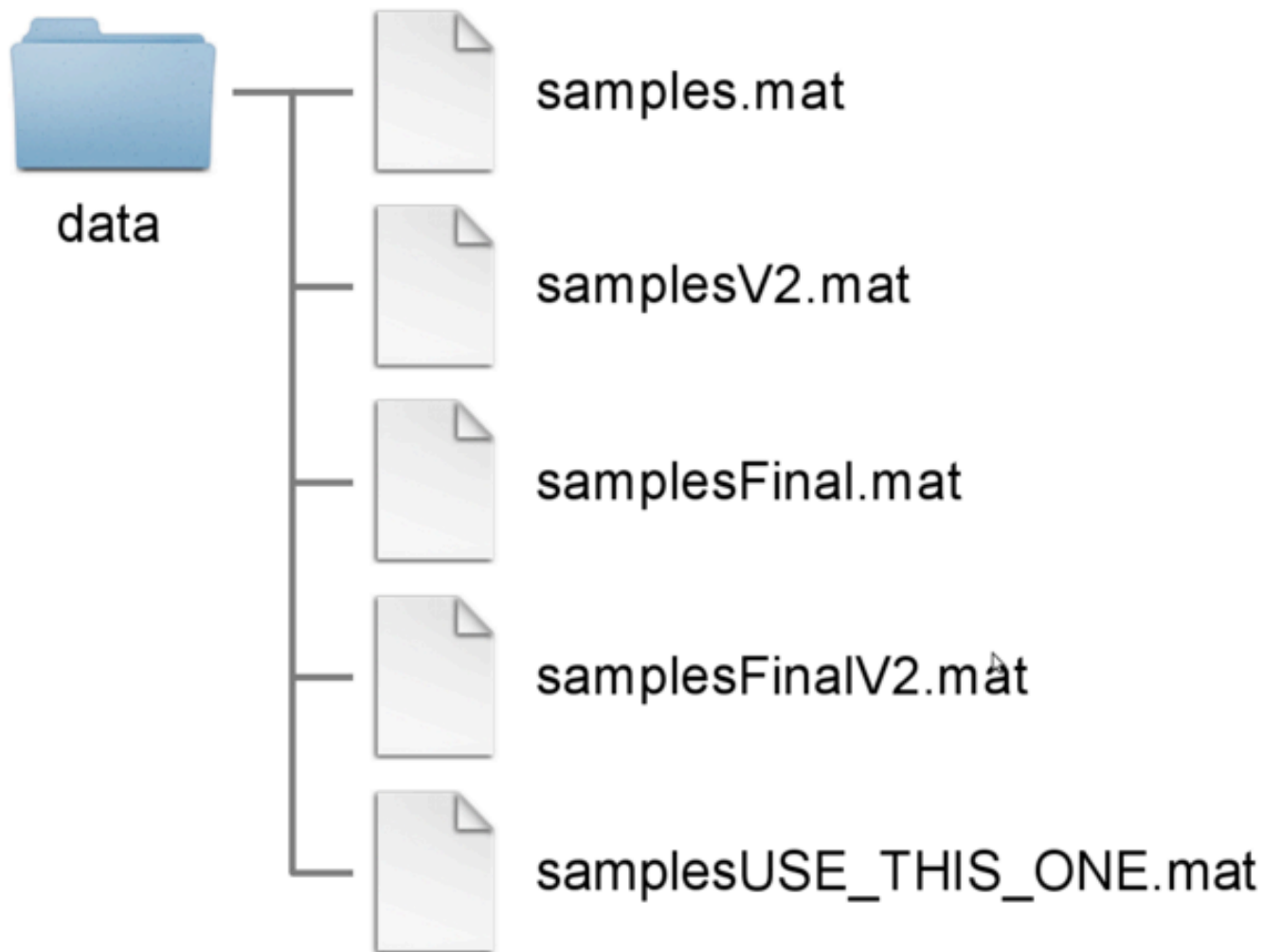
data

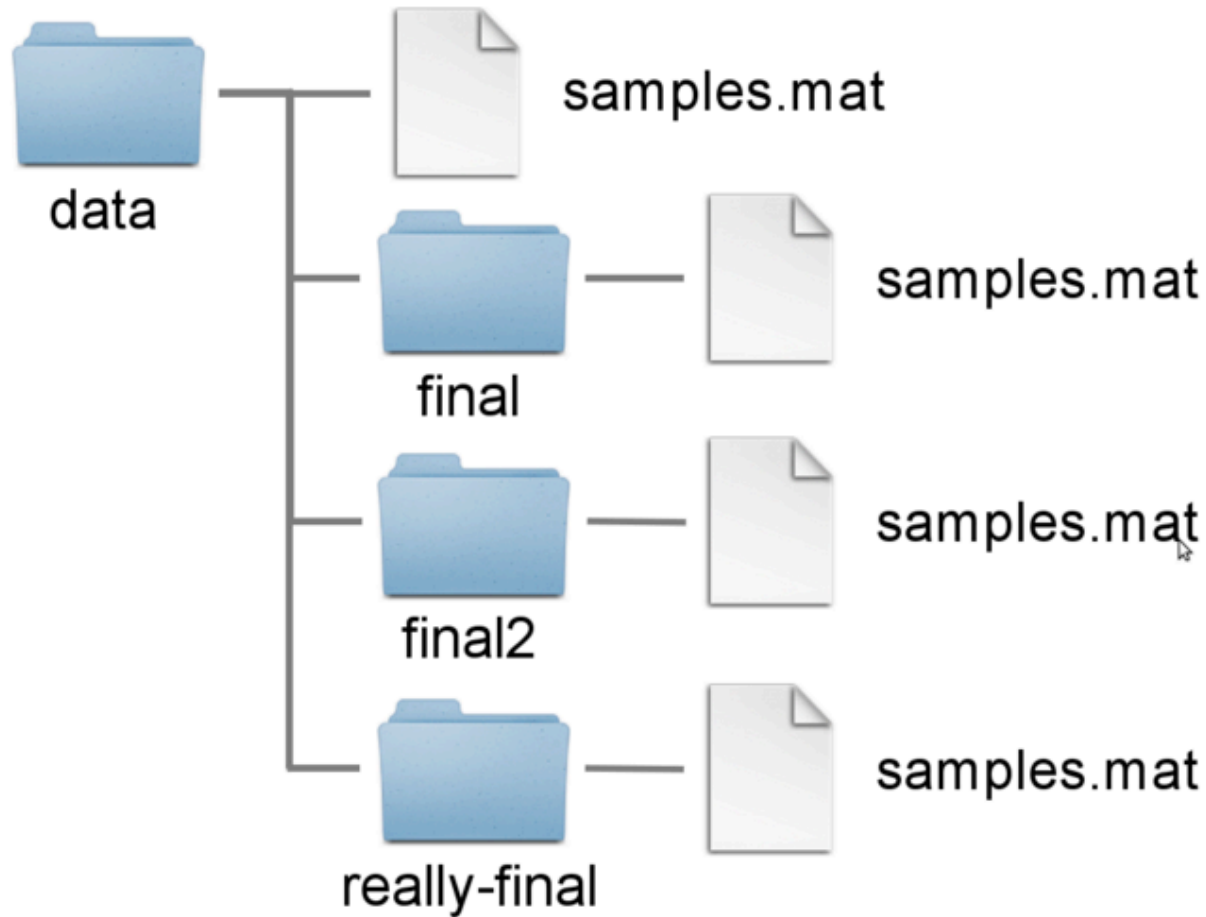


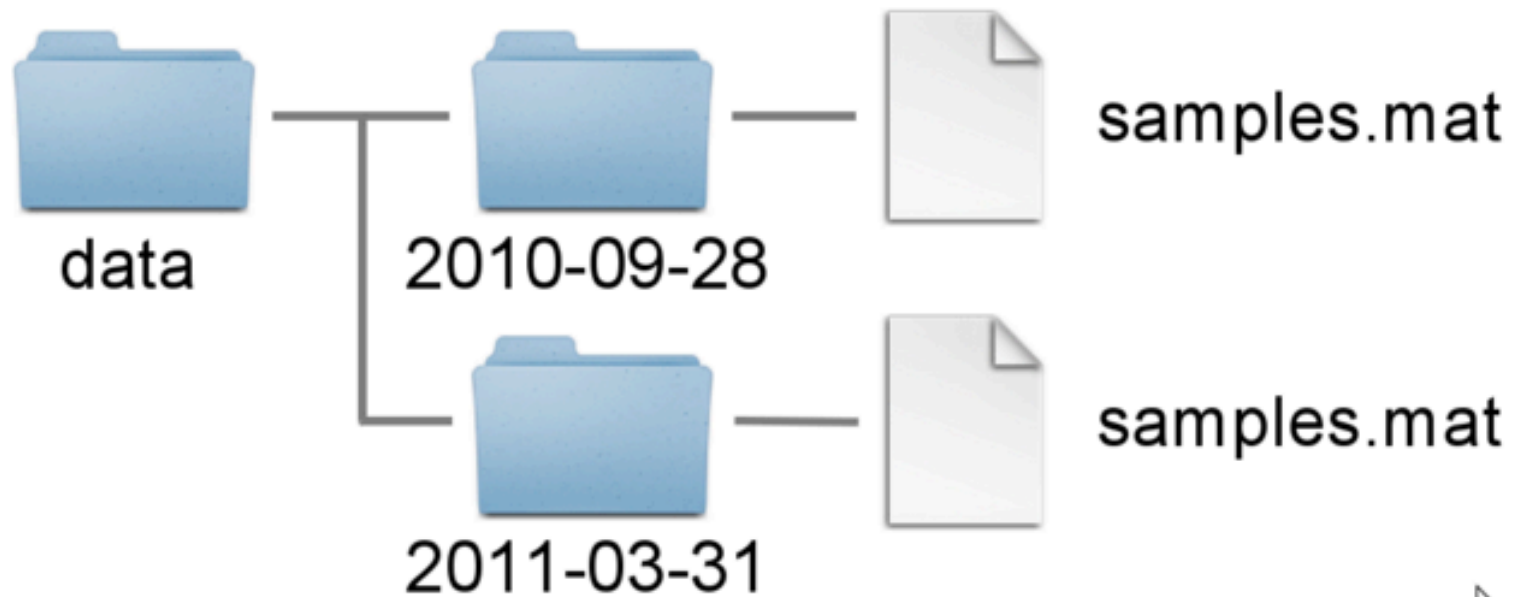
samples.mat



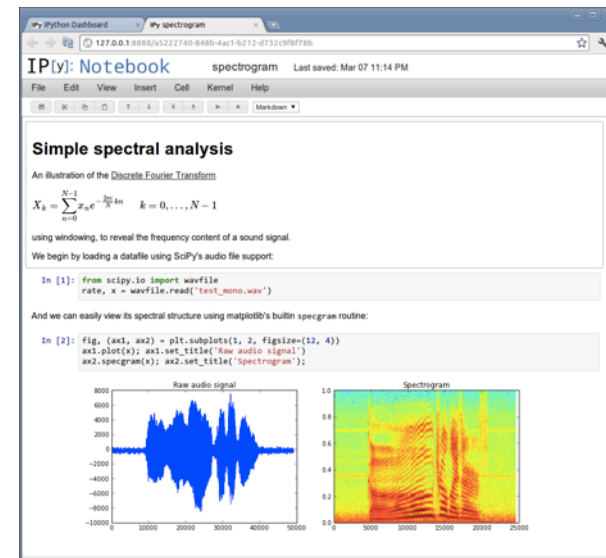
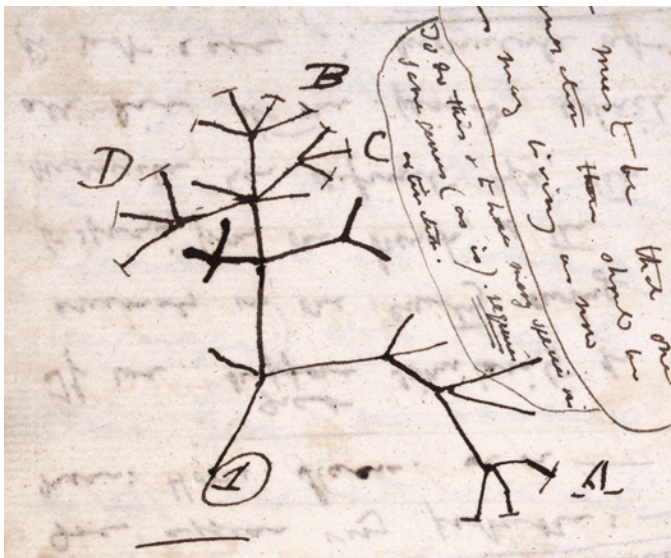




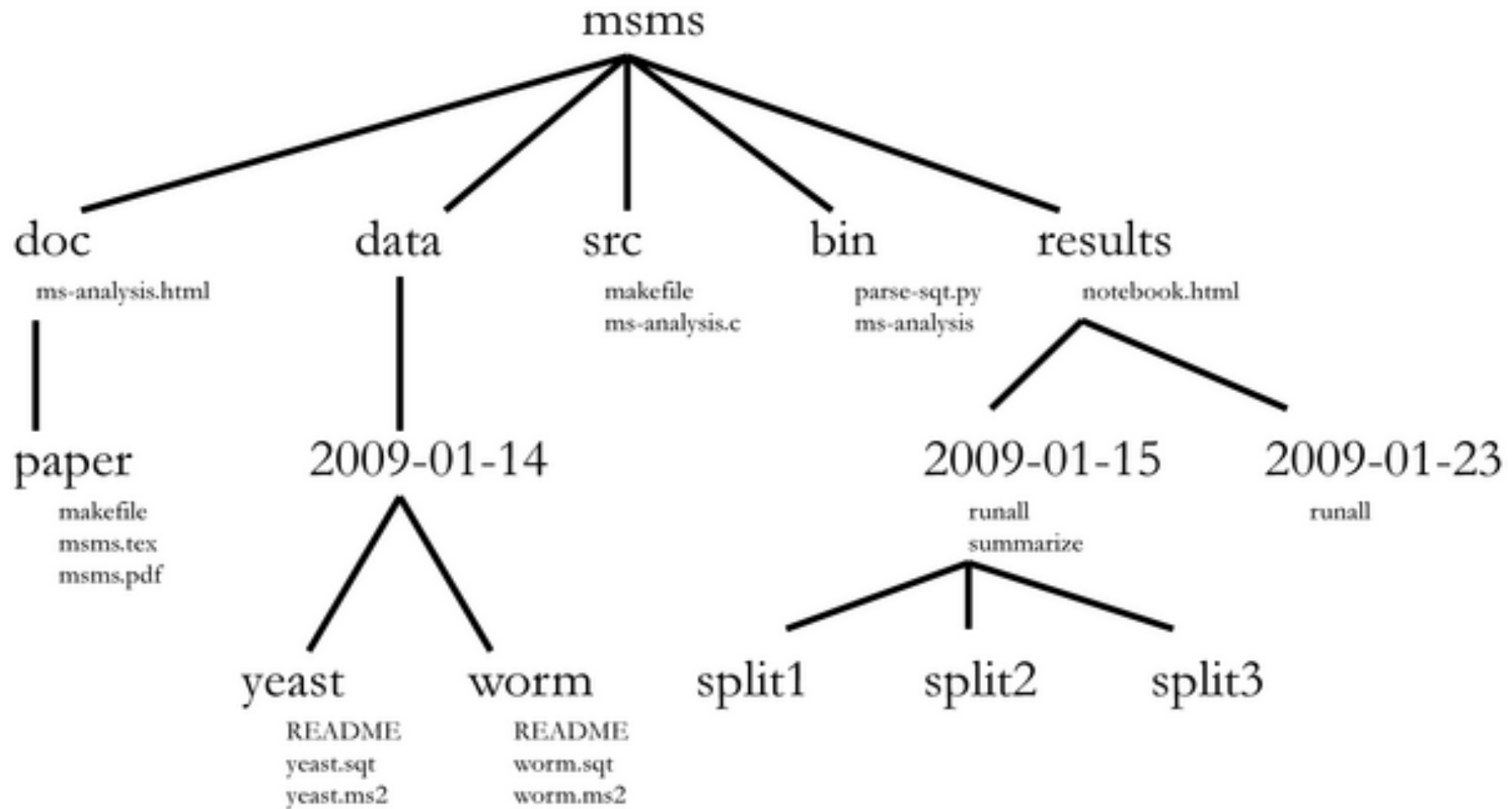




- Need context → document **metadata**
 - How was the data generated?
 - From what was the data generated?
 - What were the experimental conditions?
 - Etc
- Need to describe what is done → **Lab notebook**
 - Dated entries
 - Point to commands run and results generated



-
- 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 **non-proprietary formats** – .csv rather than .x/sx
 - 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.
 - **Manuscript production** output is kept **separate** from everything else.
 - Etc...



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>

```

— 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)

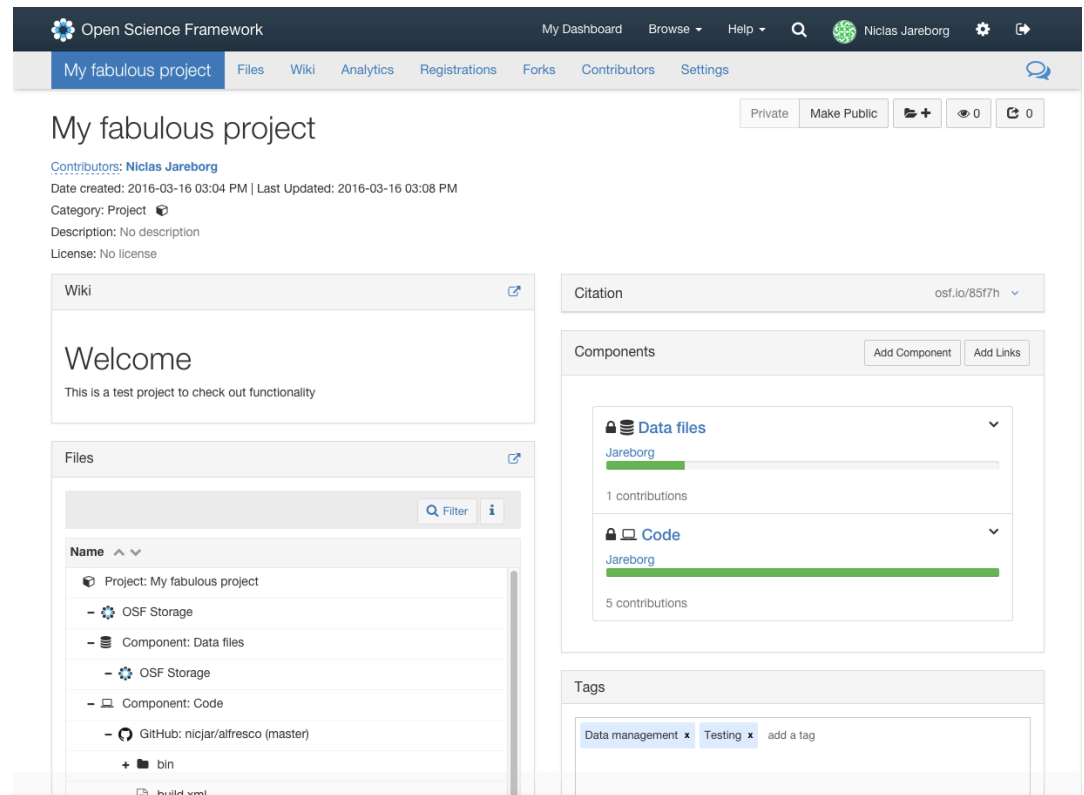
```

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

- 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



- 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



Personal data



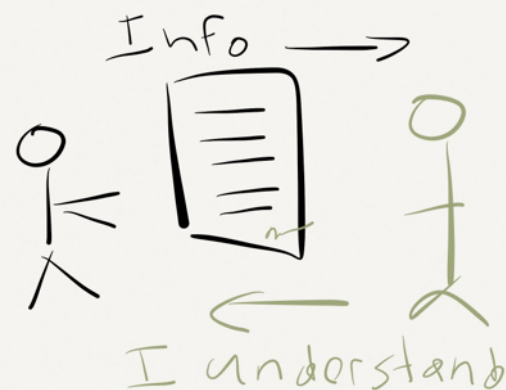
- Personal Data Act (*Personuppgiftslagen (PUL)*)
- Act concerning the Ethical Review of Research Involving Humans (*Lag om etikprövning av forskning som avser människor*)



- 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

- 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

- 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



- 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.”*

- e-Infrastructure for working with sensitive data for academic research
- Inspired by Norwegian solution (TSD)
- Designed to look like UPPMAX clusters
- Project
 - Funded and coordinated by BILS
 - Project lead: Jonas Hagberg (Niclas J)
 - Completed 2015-11-05
- System
 - Owned by BILS
 - Operated and hosted by UPPMAX



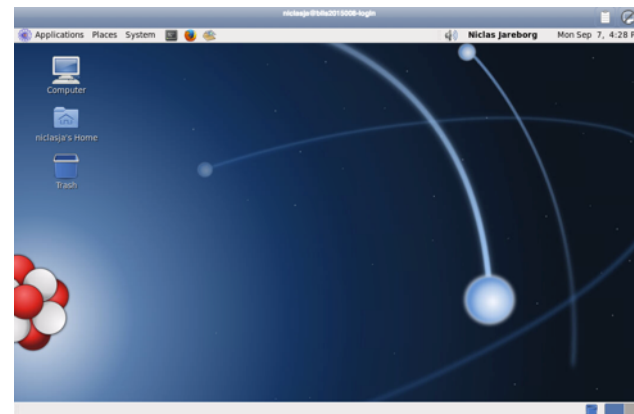
<https://bils.se/resources/mosler.html>

https://wiki.bils.se/wiki/Mosler_user_documentation

- 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
- “Pilot system”
 - 24 nodes, 270 TB
- User administration through SNIC SUPR
- 2-factor authentication for login
 - F2F identity check for users accounts
 - 2-factor (token) administrators
 - Katarina Truvé (Göteborg)
 - Malin Larsson (Linköping)
 - Dag Ahrén (Lund)
 - Mikael Borg (Stockholm)
 - Niclas Jareborg (Stockholm)
 - Johan Viklund (Uppsala)
 - Thomas Källman (Uppsala)
 - Jeanette Tångrot (Umeå)



- Only accessible over remote Linux desktop (ThinLinc) via a web dashboard
 - *from specified IP ranges*
- 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
- All user activity logged
- Software
 - UPPMAX modules
 - UPPMAX can assist with installing custom tools
- *Future*
 - SNIC Sens – “bianca”
 - Plan to have in operation in June



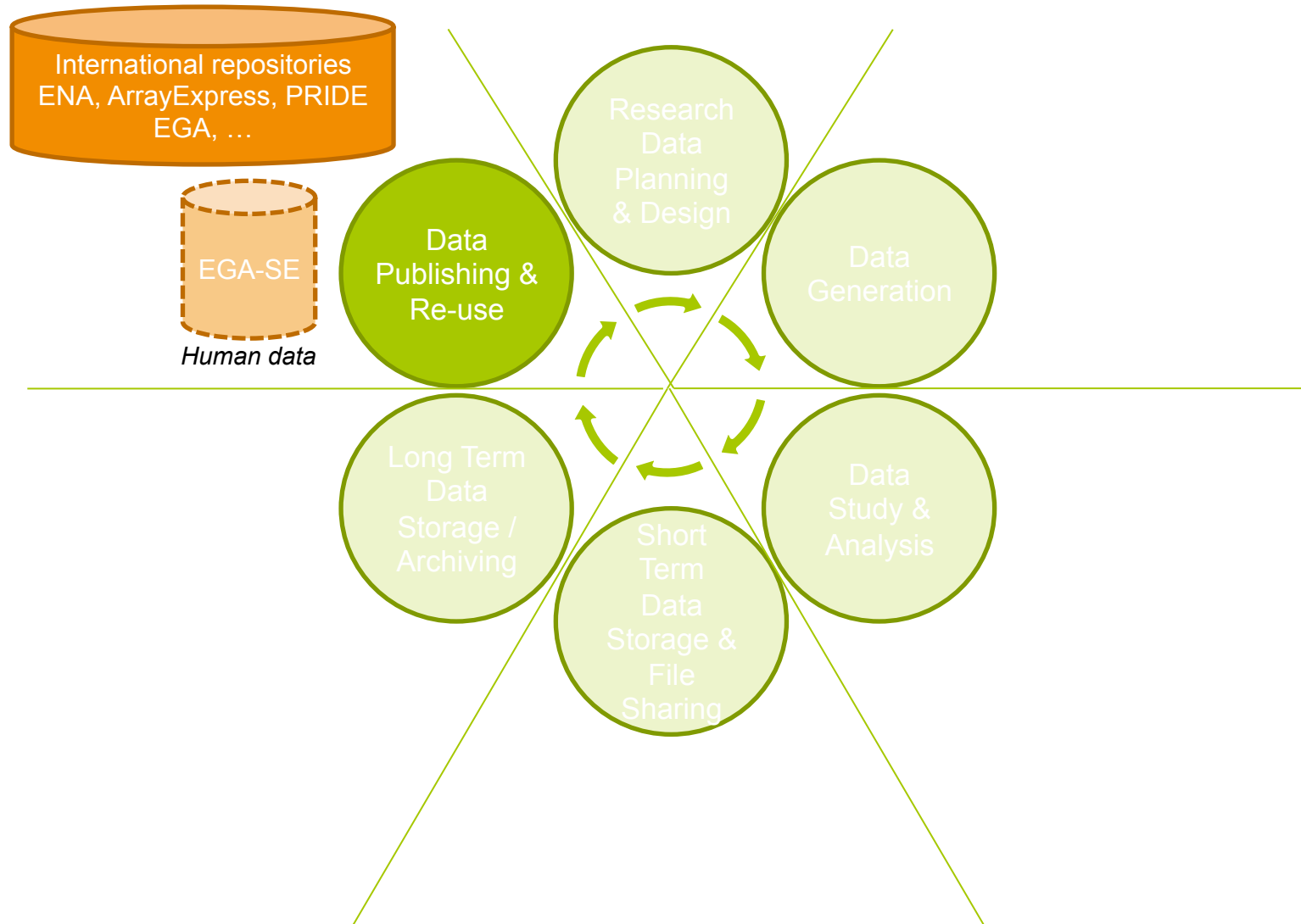
Tryggve - collaboration for sensitive biomedical data

- 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
 - Analysing 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




- Long-term storage
- Persistent identifiers
- Discoverability
 - Metadata
- To be useful for others data should be
 - *FAIR* - Findable, Accessible, Interoperable, and Reusable

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.^a

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 measurable 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.

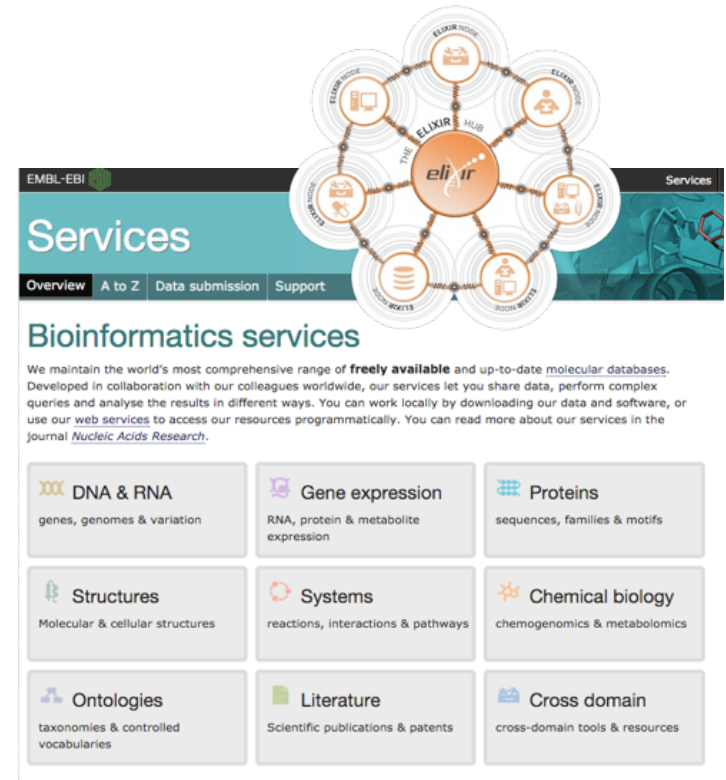
Supporting discovery through good data management

Received: 10 December 2015

Accepted: 12 February 2016

Published: 15 March 2016

- Best way to make data findable and re-usable
- EBI databases
 - ENA, Array Express, PRIDE etc
- Domain-specific metadata standards
- *Not always straight-forward for users*



The image shows a screenshot of the EMBL-EBI Services page. At the top right, there is a circular diagram labeled 'THE ELIXIR HUB' with 'elixir' in the center. It is surrounded by icons representing various services: ELIXIR NODE, ELIXIR DATA, ELIXIR PEOPLE, ELIXIR TOOLS, ELIXIR RESEARCH, ELIXIR INFRASTRUCTURE, ELIXIR COMMUNITY, and ELIXIR SERVICES. The main page content includes a navigation bar with 'Overview', 'A to Z', 'Data submission', and 'Support'. Below this is the 'Services' section, which states: 'We maintain the world's most comprehensive range of **freely available** and up-to-date molecular databases. Developed in collaboration with our colleagues worldwide, our services let you share data, perform complex queries and analyse the results in different ways. You can work locally by downloading our data and software, or use our web services to access our resources programmatically. You can read more about our services in the journal Nucleic Acids Research.' Below this text is a grid of nine service categories, each with an icon and a brief description:

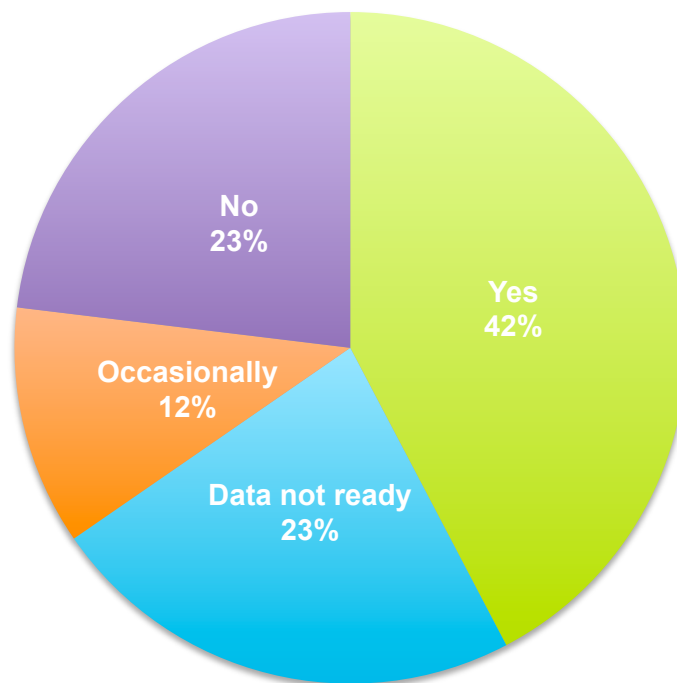
DNA & RNA genes, genomes & variation	Gene expression RNA, protein & metabolite expression	Proteins sequences, families & motifs
Structures Molecular & cellular structures	Systems reactions, interactions & pathways	Chemical biology chemogenomics & metabolomics
Ontologies taxonomies & controlled vocabularies	Literature Scientific publications & patents	Cross domain cross-domain tools & resources

- 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

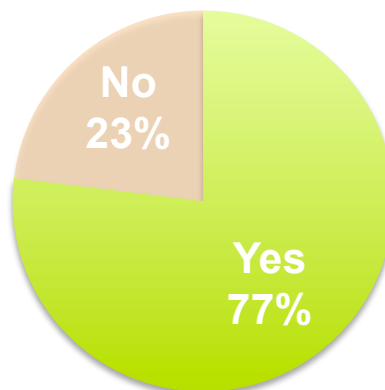
- NGI user mailing list
 - 22 April – 15 May 2015
 - 52 responses
- *Have you submitted your sequence data to public repositories?*
 - 1/4 have not submitted their data to a public repository



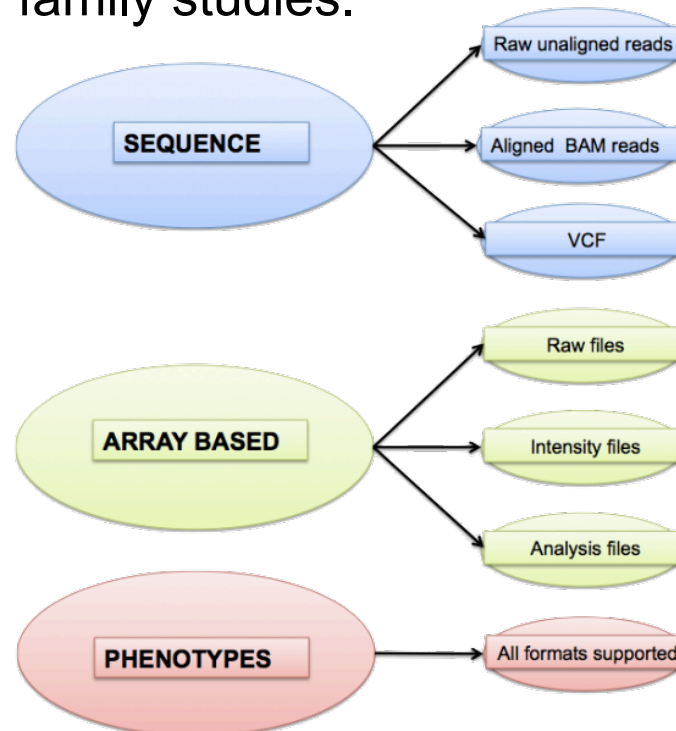
- Reasons for not submitting data*



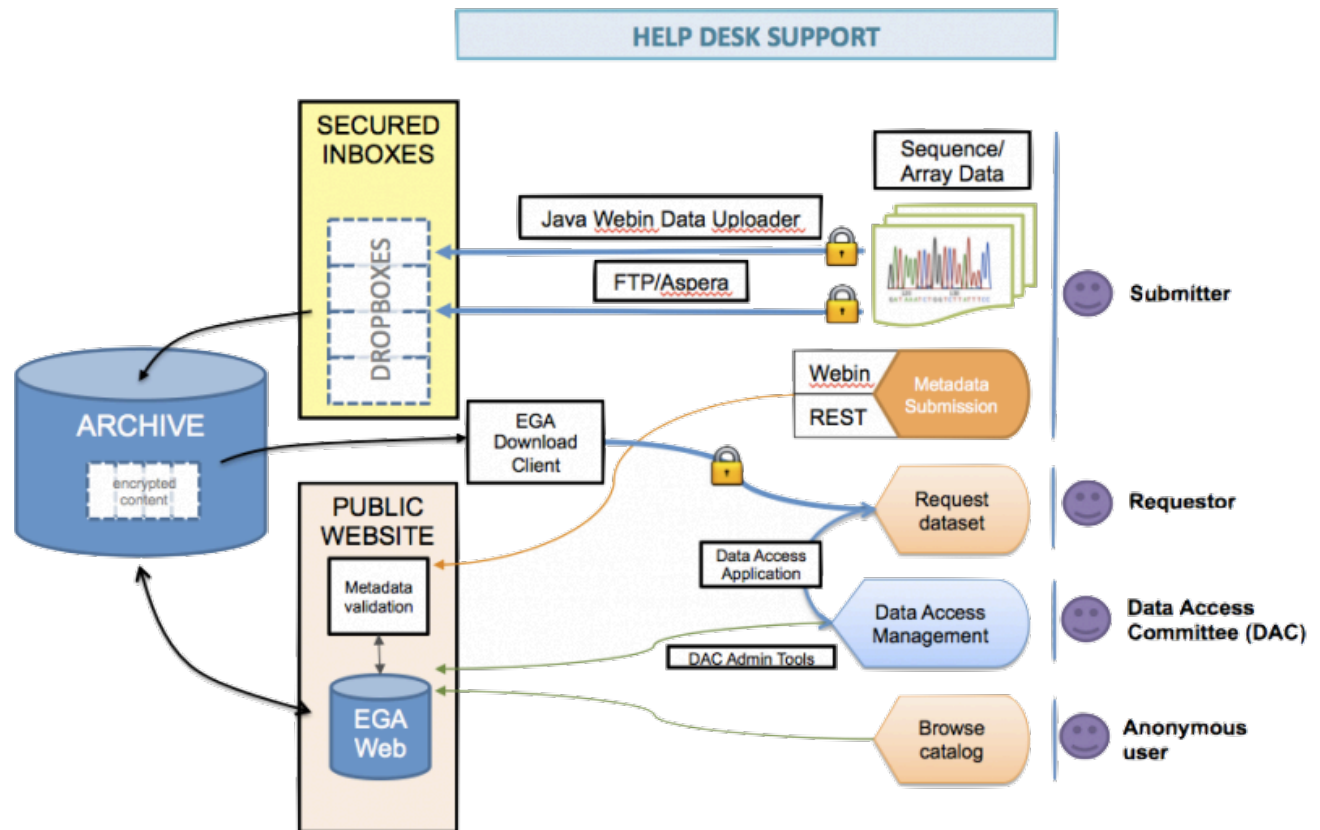
- Want support to submit to public repositories?*



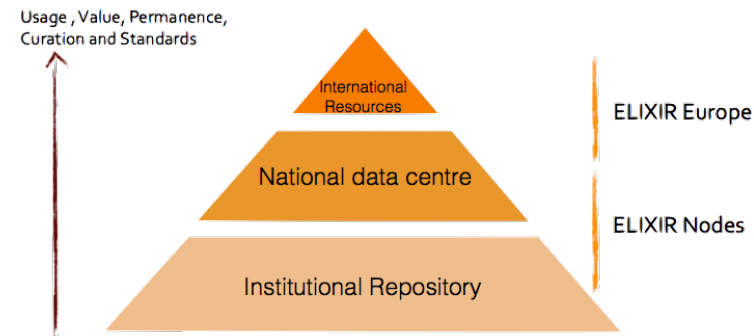
- EGA – European Genome-phenome Archive
 - Repository that promotes the distribution and sharing of genetic and phenotypic data consented for specific approved uses but not fully open, public distribution.
 - All types of sequence and genotype experiments, including case-control, population, and family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



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



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



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

-
- Research data that doesn't fit in structured data repositories
 - Data publication – persistent identifiers
 - Metadata submission – not tailored to Life Science
 - *Affects discoverability*
 - Figshare - <https://figshare.com/>
 - EUDAT - <http://eudat.eu/>
 - Data Dryad - <http://datadryad.org/>
 - Zenodo - <http://www.zenodo.org/>

- 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>

The screenshot shows the ORCID iD profile for Niclas Jareborg. The header includes the ORCID logo and navigation links: FOR RESEARCHERS, FOR ORGANIZATIONS, ABOUT, HELP, and SIGN IN. Below the header, it states '2,035,272 ORCID iDs and counting. See more...'. The profile section for Niclas Jareborg displays his ORCID ID (0000-0002-4520-044X) and lists his education and employment history.

ORCID
Connecting Research and Researchers

Niclas Jareborg

ORCID ID
ID: orcid.org/0000-0002-4520-044X

Also known as
C. J. E. Niclas Jareborg, N Jareborg

Country
Sweden

Websites
[LinkedIn](#)
[Personal home page](#)

Education (2)

Uppsala Universitet: Uppsala, Sweden
1989-05 to 1995-05 (Microbiology)
PhD
Source: Niclas Jareborg Created: 2015-04-09

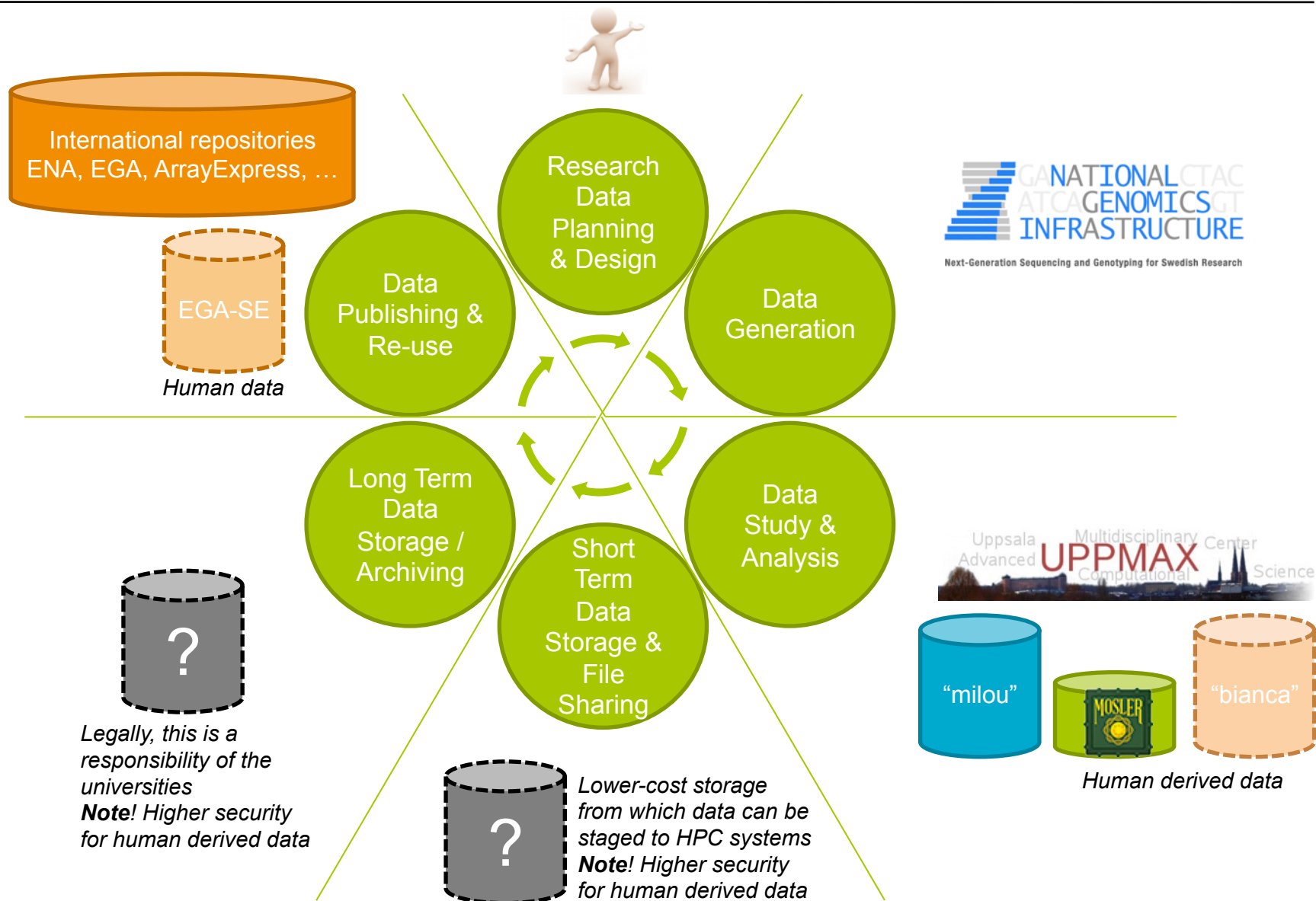
Uppsala Universitet: Uppsala, Sweden
1985-01 to 1989-04 (Microbiology)
BSc
Source: Niclas Jareborg Created: 2015-04-09

Employment (7)

Stockholms Universitet: Stockholm, Sweden
2015-01 to present (BILS / Department of Department of Biochemistry and Biophysics)
Data Manager
Source: Niclas Jareborg Created: 2015-02-23

Kungliga Tekniska Hogskolan: Stockholm, Sweden
2013-01 to 2014-12 (National Genomics Infrastructure / SciLifeLab)

- Project planning
 - Metadata
 - File formats
 - Licensing
 - *Data Management Plans*
- Data analysis
- Data publication and submission
 - Automate submissions to public repositories
 - Metadata
 - Licensing



- 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](https://doi.org/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>