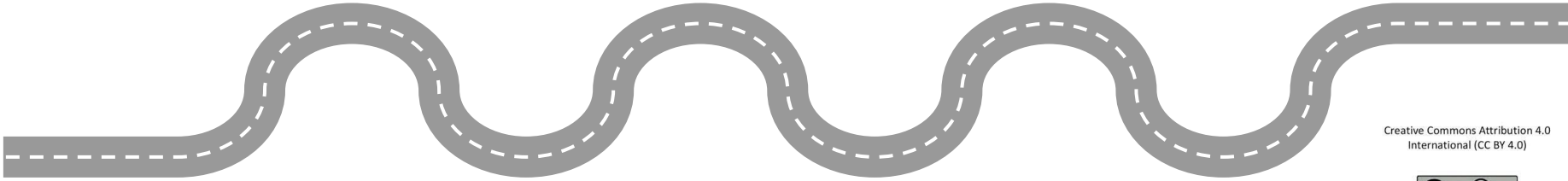
SciLifeLab

Samples Metadata at the SciLifeLab

Dr. Sara El-Gebali

0000-0003-1378-5495

@yalahowy



17-October-2022

Three dimensions of SciLifeLab

Research environment

Approx. 190 affiliated research groups

- Environment and climate change
- Farming and forestry
- Evolution and biodiversity
- Gene editing
- Biofuels and biomaterials
- Microbiology and microbiome
- Drugs and biomedicine
- Healthcare and aging

Infrastructure

Service to ~ 1400 Swedish researchers annually (2020)

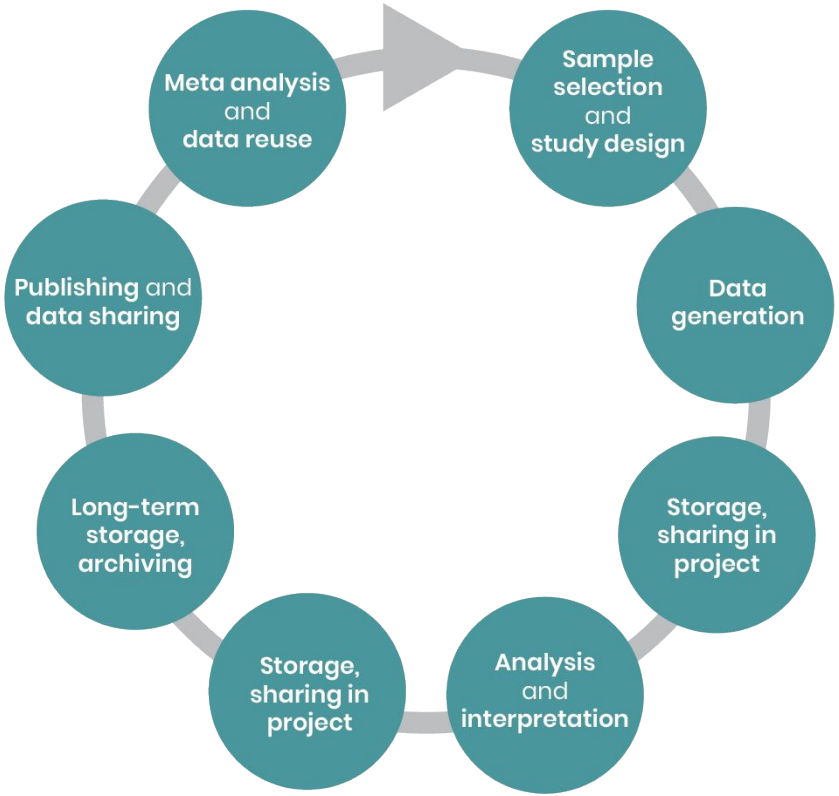
- Bioinformatics
- Cellular and molecular imaging
- Clinical diagnostics
- Single cell biology
- Genomics
- Chemical biology and gene editing
- Drug development
- Proteomics and metabolomics

Data-driven life science

3.1 billion SEK, 12-year-program

Putting Sweden at the forefront of data-driven life science research and fostering the next generation of life scientists

- Four strategic research areas
- Recruiting talent from across the globe
- Academic and industry PhD and postdoc programs
- Sparking collaborations, innovation and interdisciplinary team science
- Building a strong computational and data science base for open, real-time data



The flow of data in a cycle from production to analysis, sharing in projects, publication and re-use to initiate new studies characterizes data-driven life science, which the Data Centre assists with.

Data Centre

Created to **maximize impact** of SciLifeLab generated data

Assists in communication between platforms, users, and research community

Acts as a point of contact for data management questions relating to SciLifeLab generated data

Assists platforms with data tracking and statistics

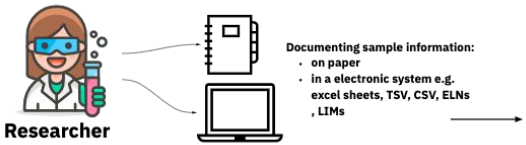
Facilitates providing SciLifeLab generated data with SciLifeLab funded bioinformatics and data management support

Assists with planning the handling of SciLifeLab generated data throughout projects

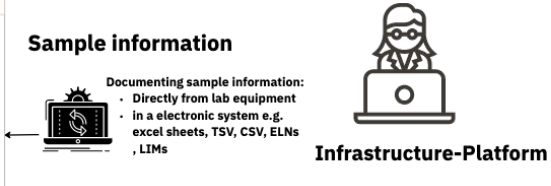
👉 services and resources for data management and IT

👉 promoting FAIR, open science, and good data practises throughout the data lifecycle

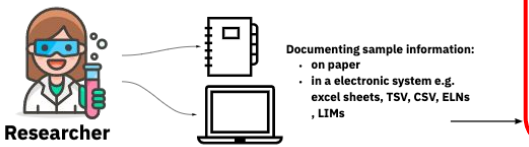
Samples Metadata at SciLifeLab



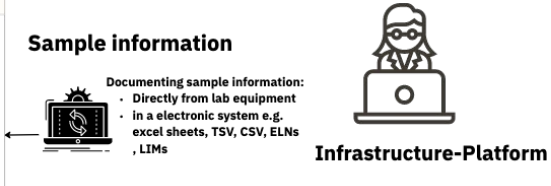
Core descriptors	Recommended descriptors	Additional info
*Minimum set of attributes for basic **discoverability***	*Domain specific*	*Flexible, user specific*
We compared a handful of schemas • IGSN • Biosamples EBI • ISA-Tab (FAIRDOMSEEK) • ENA • iSamples • Bioschemas • By-covid • MAGE-TAB Example of attributes: • Investigation title • Investigation Description • Study title • Study ID • Sample type • Sample name • Sample ID • Sample Description • Collection date (DD/MM/YYYY) • Collection time (HH/MM)	Per use-case, i.e this is what we can work with the researchers/infrastructures to develop, taking into account community standards, e.g.: • Minimum information about a microarray experiment (**MIAME**) • Minimum Information about any (x) Sequence (**MIxS**) • Minimum /nformation for Publication of Quantitative Real-Time PCR (**MIQE**)	Any information users need for their internal processes, flexible; anything else that support FAIRness.



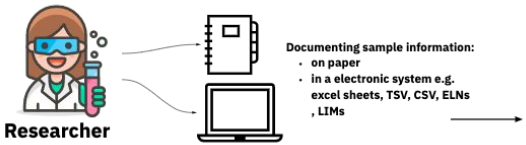
Samples Metadata at SciLifeLab



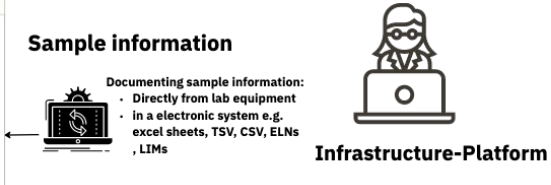
Core descriptors	Recommended descriptors	Additional info
*Minimum set of attributes for basic **discoverability***	*Domain specific*	*Flexible, user specific*
We compared a handful of schemas • IGSN • Biosamples EBI • ISA-Tab (FAIRDOMSEEK) • ENA • iSamples • Bioschemas • By-covid • MAGE-TAB Example of attributes: • Investigation title • Investigation Description • Study title • Study ID • Sample type • Sample name • Sample ID • Sample Description • Collection date (DD/MM/YYYY) • Collection time (HH/MM)	Per use-case, i.e this is what we can work with the researchers/infrastructures to develop, taking into account community standards, e.g.: • Minimum information about a microarray experiment (**MIAME**) • Minimum Information about any (x) Sequence (**MIxS**) • Minimum /nformation for Publication of Quantitative Real-Time PCR (**MIQE**)	Any information users need for their internal processes, flexible; anything else that support FAIRness.



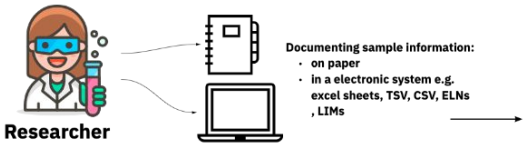
Samples Metadata at SciLifeLab



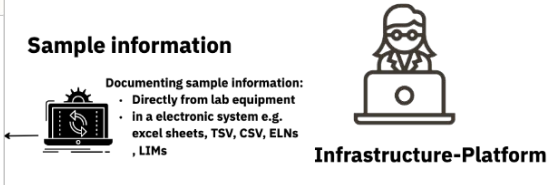
Core descriptors	Recommended descriptors	Additional info
*Minimum set of attributes for basic **discoverability***	*Domain specific*	*Flexible, user specific*
We compared a handful of schemas • IGSN • Biosamples EBI • ISA-Tab (FAIRDOMSEEK) • ENA • iSamples • Bioschemas • By-covid • MAGE-TAB Example of attributes: • Investigation title • Investigation Description • Study title • Study ID • Sample type • Sample name • Sample ID • Sample Description • Collection date (DD/MM/YYYY) • Collection time (HH/MM)	Per use-case, i.e this is what we can work with the researchers/infrastructures to develop, taking into account community standards, e.g.: • Minimum information about a microarray experiment (**MIAME**) • Minimum Information about any (x) Sequence (**MIxS**) • Minimum /nformation for Publication of Quantitative Real-Time PCR (**MIQE**)	Any information users need for their internal processes, flexible; anything else that support FAIRness.



Samples Metadata at SciLifeLab



Core descriptors	Recommended descriptors	Additional info
*Minimum set of attributes for basic **discoverability***	*Domain specific*	*Flexible, user specific*
We compared a handful of schemas • IGSN • Biosamples EBI • ISA-Tab (FAIRDOMSEEK) • ENA • iSamples • Bioschemas • By-covid • MAGE-TAB Example of attributes: • Investigation title • Investigation Description • Study title • Study ID • Sample type • Sample name • Sample ID • Sample Description • Collection date (DD/MM/YYYY) • Collection time (HH/MM)	Per use-case, i.e this is what we can work with the researchers/infrastructures to develop, taking into account community standards, e.g.: • Minimum information about a microarray experiment (**MIAME**) • Minimum Information about any (x) Sequence (**MIxS**) • Minimum /nformation for Publication of Quantitative Real-Time PCR (**MIQE**)	Any information users need for their internal processes, flexible; anything else that support FAIRness.



Community

Persistent identifiers IGSN DataCite FIND, ACCESS, AND REUSE DATA

DataCite IGSN partnership

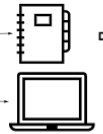
- In October 2021, DataCite and IGSN e.V. announced a formal partnership following the recommendations of the IGSN 2040 Project
  - **DataCite** – Provides IGSN ID registration services and supports ongoing sustainability of IGSN ID infrastructure
  - **IGSN e.V.** – Serves to implement and promote standard methods for identifying, citing, and locating material samples with confidence
- Phase 1: Technical transition of IGSN ID infrastructure and members under DataCite
  - DataCite DOI services for registering IGSN IDs for material samples has been launched!
  - Support and best practice documentation is available: https://support.datacite.org/
- Phase 2: Scaling of community engagement to ensure long-term sustainability
  - Increase discovery, adoption, and use of IGSN IDs
  - Establish Communities of Practice for support of standardized methods to identify, cite, & locate samples
    - Promote use of PIDs for material samples
    - Articulate use cases within (sub)disciplines – mapping out the workflow
    - Describe material samples – minimum useful metadata description and extensions

Contact Rorie Edmunds (rorie.edmunds@datacite.org) for more information on IGSN IDs and CoPs

SEEK at SciLifeLab

Researcher

Sample information



Documenting sample information:
- **on paper**
- **in a electronic system e.g. excel sheets, TSV, CSV, ELNs , LIMs**

Core descriptors	Recommended descriptors	Additional info
*Minimum set of attributes for basic **discoverability***	*Domain specific*	*Flexible, user specific*
We compared a handful of schemas • IGSN • Biosamples EBI • ISA-Tab (FAIRDOMSEEK) • ENA • iSamples • Bioschemas • By-covid • MAGE-TAB Example of attributes: • Investigation title • Investigation Description • Study title • Study ID • Sample type • Sample name • Sample ID • Sample Description • Collection date (DD/MM/YYYY) • Collection time (HH/MM)	Per use-case, i.e this is what we can work with the researchers/infrastructures to develop, taking into account community standards, e.g.: • Minimum information about a microarray experiment (**MIAME**) • Minimum Information about any (x) Sequence (**MIxS**) • Minimum /nformation for Publication of Quantitative Real-Time PCR (**MIQE**)	Any information users need for their internal processes, flexible; anything else that support FAIRness.

Sample information

Documenting sample information:
- **Directly from lab equipment**
- **in a electronic system e.g. excel sheets, TSV, CSV, ELNs , LIMs**

Infrastructure-Platform

FAIRDOM-SEEK

Sample information

Persistent identifiers (IGSN, ORCID, ROR)

Data files, directories, workflows

Standards and code

RoCrate

SciLifeLab- Sample information Database

Persistent landing page for each sample

Sample ID	related data/Metadata	Results/Data sets	XYZ

- Central institutional repository for sample metadata storage
- Enhanced Findability and Accessibility for samples
- Linked to research outputs
- Increased complex multi-omics data integration enabling higher power analyses

EMBL-EBI Repositories

ENA, Biosamples, more

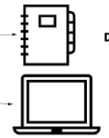
SEEK at SciLifeLab

SEEK

Order portal integration

Sample information

Researcher



Documenting sample information:
- **on paper**
- **in a electronic system e.g. excel sheets, TSV, CSV, ELNs , LIMs**

Core descriptors	Recommended descriptors	Additional info
*Minimum set of attributes for basic **discoverability***	*Domain specific*	*Flexible, user specific*
We compared a handful of schemas • IGSN • Biosamples EBI • ISA-Tab (FAIRDOMSEEK) • ENA • iSamples • Bioschemas • By-covid • MAGE-TAB Example of attributes: • Investigation title • Investigation Description • Study title • Study ID • Sample type • Sample name • Sample ID • Sample Description • Collection date (DD/MM/YYYY) • Collection time (HH/MM)	Per use-case, i.e this is what we can work with the researchers/infrastructures to develop, taking into account community standards, e.g.: • Minimum information about a microarray experiment (**MIAME**) • Minimum Information about any (x) Sequence (**MIxS**) • Minimum *I*nformation for Publication of *Q*uantitative Real-Time PCR (**MIQE**)	Any information users need for their internal processes, flexible; anything else that support FAIRness.

Sample information

Documenting sample information:
- **Directly from lab equipment**
- **in a electronic system e.g. excel sheets, TSV, CSV, ELNs , LIMs**

Infrastructure Units

Data delivery system integration

Workflows for infrastructures

Storage

DataStewardship Wizard

Figshare

Galaxy **Protocols.io**

Communication between users and infrastructures **PIDs** **ELNs**

Compute environments

Analysis and Visualisation tools **Stackn**

Brokering **Interface UX/UI**

Search functions **+ other integrations codeveloped with Datahub**

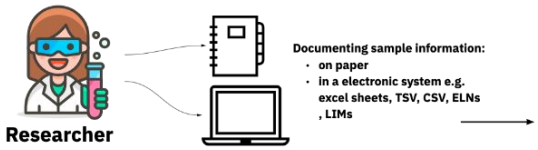
SEEK at SciLifeLab

SEEK

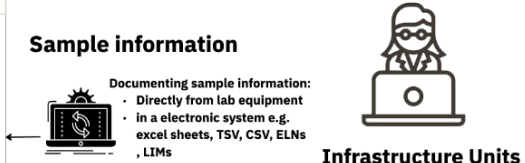
Core descriptors	Recommended descriptors	Additional info
*Minimum set of attributes for basic **discoverability***	*Domain specific*	*Flexible, user specific*
We compared a handful of schemas • IGSN • Biosamples EBI • ISA-Tab (FAIRDOMSEEK) • ENA • iSamples • Bioschemas • By-covid • MAGE-TAB Example of attributes: • Investigation title • Investigation Description • Study title • Study ID • Sample type • Sample name • Sample ID • Sample Description • Collection date (DD/MM/YYYY) • Collection time (HH/MM)	Per use-case, i.e this is what we can work with the researchers/infrastructures to develop, taking into account community standards, e.g.: • Minimum information about a microarray experiment (**MIAME**) • Minimum Information about any (x) Sequence (**MIxS**) • Minimum *I*nformation for Publication of *Q*uantitative Real-Time PCR (**MIQE**)	Any information users need for their internal processes, flexible; anything else that support FAIRness.

Order portal integration

Sample information



Sample information



Data delivery system integration

Workflows for infrastructures

Storage

DataStewardship Wizard

Figshare

Galaxy **Protocols.io**

Communication between users and infrastructures **PIDs** **ELNs**

Analysis and Visualisation tools **Stackn**

Compute environments

Brokering **Interface UX/UI**

Search functions **+ other integrations codeveloped with Datahub**

3x Developers
1x UX/UI
1x Community coordinator
1x DataSteward

Stay in touch

Email: datacentre@scilifelab.se

Twitter: @scilifelab_DC

Linkedin: scilifelab-data-centre

Funders

Knut and Alice Wallenberg Foundation

Swedish Research Council

SciLifeLab