FAIR raziskovalni podatki v biologiji

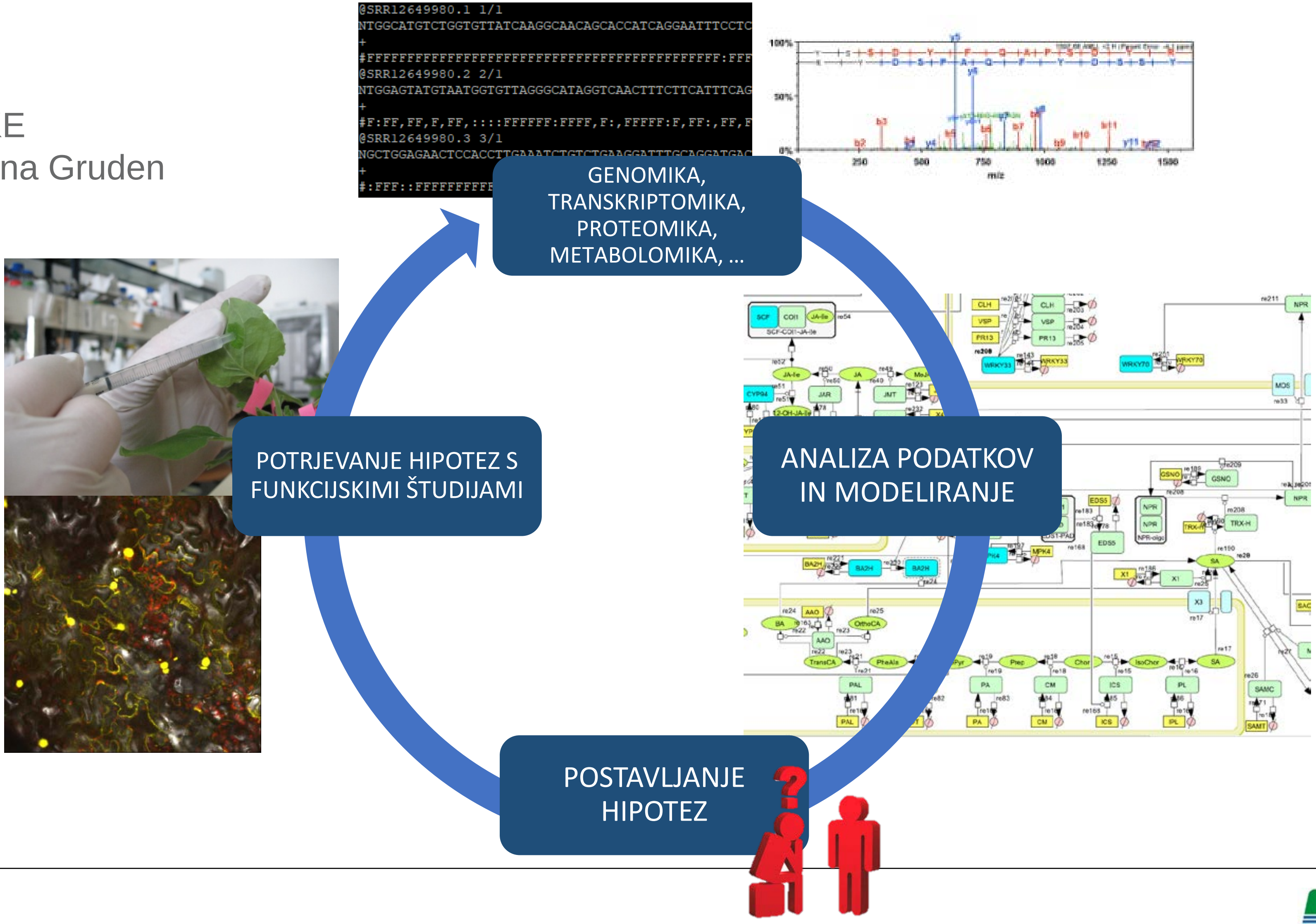
NACIONALNI INŠTITUT ZA BIOLOGIJO
NATIONAL INSTITUTE OF BIOLOGY

dr. Marko Petek
marko.petek@nib.si

12.05.2022

Raziskave na Oddelku za biotehnologijo in sistemsko biologijo

Delovni sklop: OMIKE
Vodja: prof. dr. Kristina Gruden



```
@SRR12649980.1 1/1
NTGGCATGTCTGGTGTTATCAAGGCAACAGCACCATCAGGAATTTCCTC
+
#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFF
@SRR12649980.2 2/1
NTGGAGTATGTAATGGTGTTAGGGCATAGGTCAACTTTCTTCATTTCAG
+
#F:FF,FF,F,FF,::::FFFFFF:FFFF,F:,FFFFF:F,FF:,FF,F
@SRR12649980.3 3/1
NGCTGGAGAACTCCACCTTGAAATCTGTCTGAAGGATTTGCAGGATGAC
+
#:FFF::FFFFFFFFF
```

GENOMIKA, TRANSKRIPTOMIKA, PROTEOMIKA, METABOLOMIKA, ...

ANALIZA PODATKOV IN MODELIRANJE

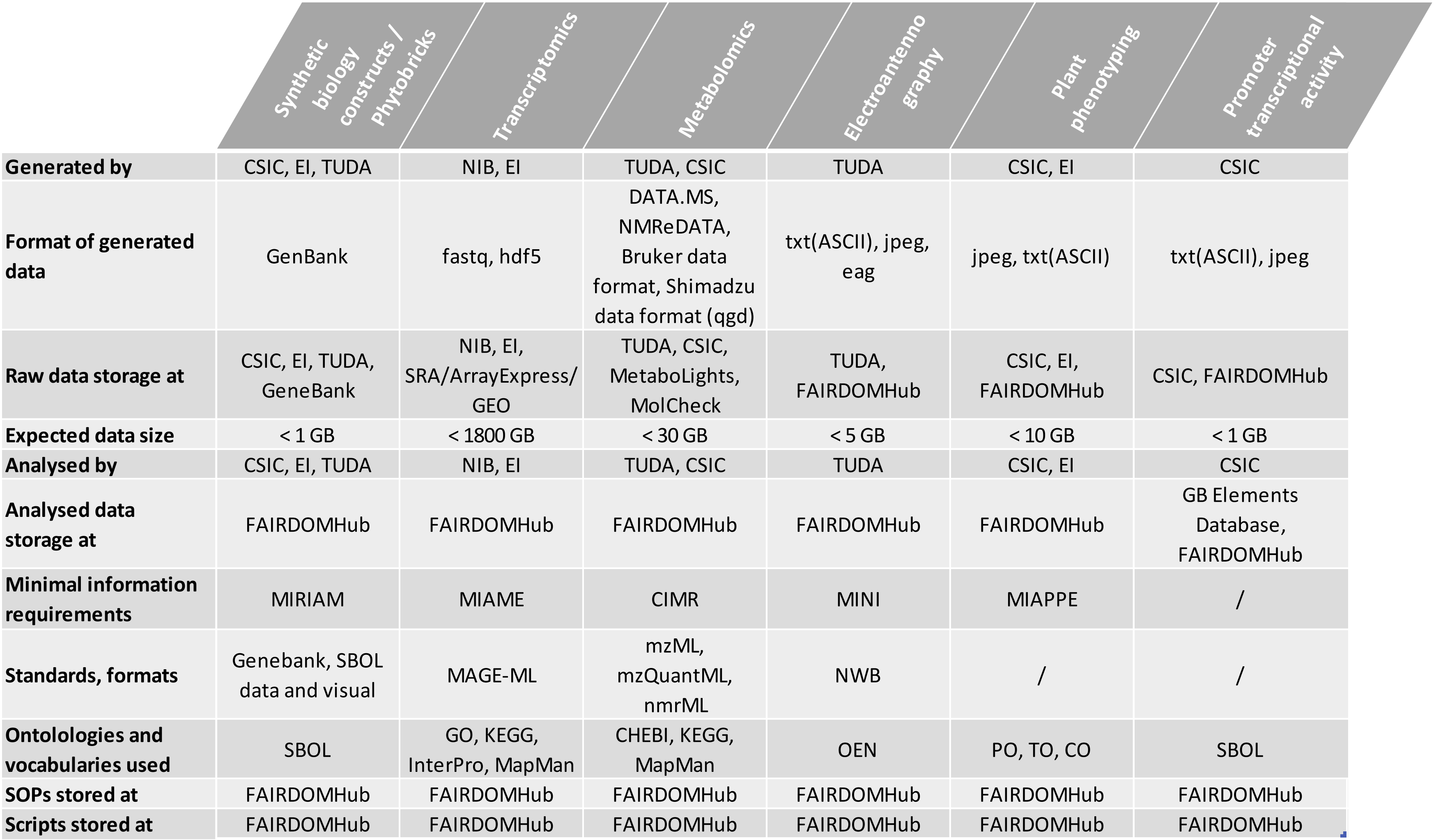
POSTAVLJANJE HIPOTEZ

POTRJEVANJE HIPOTEZ S FUNKCIJSKIMI ŠTUDIJAMI

Kakšne raziskovane podatke generiramo?

- Surovi podatki
- Metapodatki
- Vmesni in končni rezultati obdelave podatkov in modeliranja
- Publikacije (članki, poročila, predstavitve, posterji etc)

Načrt ravnanja s projektnimi podatki (“data management plan”)



	Synthetic biology constructs / Phytobricks	Transcriptomics	Metabolomics	Electroantenno graphy	Plant phenotyping	Promoter transcriptional activity
Generated by	CSIC, EI, TUDA	NIB, EI	TUDA, CSIC	TUDA	CSIC, EI	CSIC
Format of generated data	GenBank	fastq, hdf5	DATA.MS, NMReDATA, Bruker data format, Shimadzu data format (qgd)	txt(ASCII), jpeg, eag	jpeg, txt(ASCII)	txt(ASCII), jpeg
Raw data storage at	CSIC, EI, TUDA, GeneBank	NIB, EI, SRA/ArrayExpress/ GEO	TUDA, CSIC, MetaboLights, MolCheck	TUDA, FAIRDOMHub	CSIC, EI, FAIRDOMHub	CSIC, FAIRDOMHub
Expected data size	< 1 GB	< 1800 GB	< 30 GB	< 5 GB	< 10 GB	< 1 GB
Analysed by	CSIC, EI, TUDA	NIB, EI	TUDA, CSIC	TUDA	CSIC, EI	CSIC
Analysed data storage at	FAIRDOMHub	FAIRDOMHub	FAIRDOMHub	FAIRDOMHub	FAIRDOMHub	GB Elements Database, FAIRDOMHub
Minimal information requirements	MIRIAM	MIAME	CIMR	MINI	MIAPPE	/
Standards, formats	Genebank, SBOL data and visual	MAGE-ML	mzML, mzQuantML, nmrML	NWB	/	/
Ontolologies and vocabularies used	SBOL	GO, KEGG, InterPro, MapMan	CHEBI, KEGG, MapMan	OEN	PO, TO, CO	SBOL
SOPs stored at	FAIRDOMHub	FAIRDOMHub	FAIRDOMHub	FAIRDOMHub	FAIRDOMHub	FAIRDOMHub
Scripts stored at	FAIRDOMHub	FAIRDOMHub	FAIRDOMHub	FAIRDOMHub	FAIRDOMHub	FAIRDOMHub

Javni podatkovni repozitoriji

SPECIALIZIRANI REPOZITORIJI ZA SUROVE BIOLOŠKE PODATKE

GEO
Gene Expression Omnibus

ENA
European Nucleotide Archive

SRA

MetaboLights

PubChem

e!DAL
Plant Genomics & Phenomics Research Data Repository

PRIDE

EMPIAR
Electron Microscopy Public Image Archive

SPLOŠNI REPOZITORIJI ZA PODATKE

zenodo

figshare

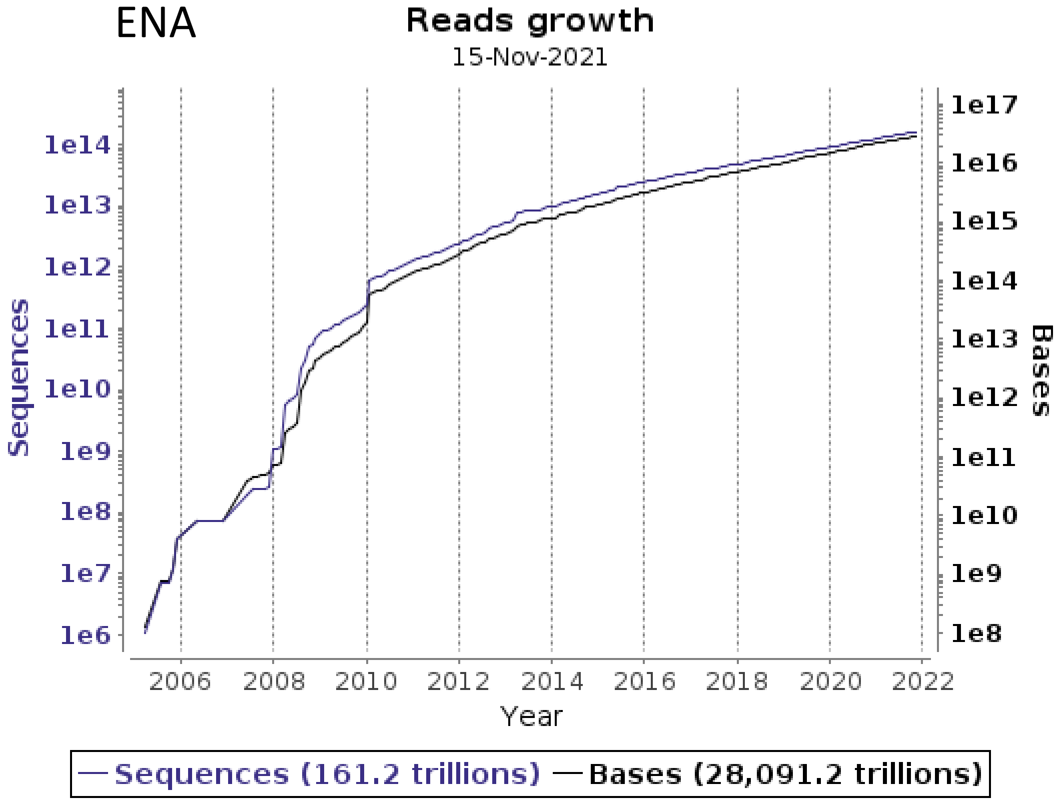
FAIRDOM HUB

GitHub

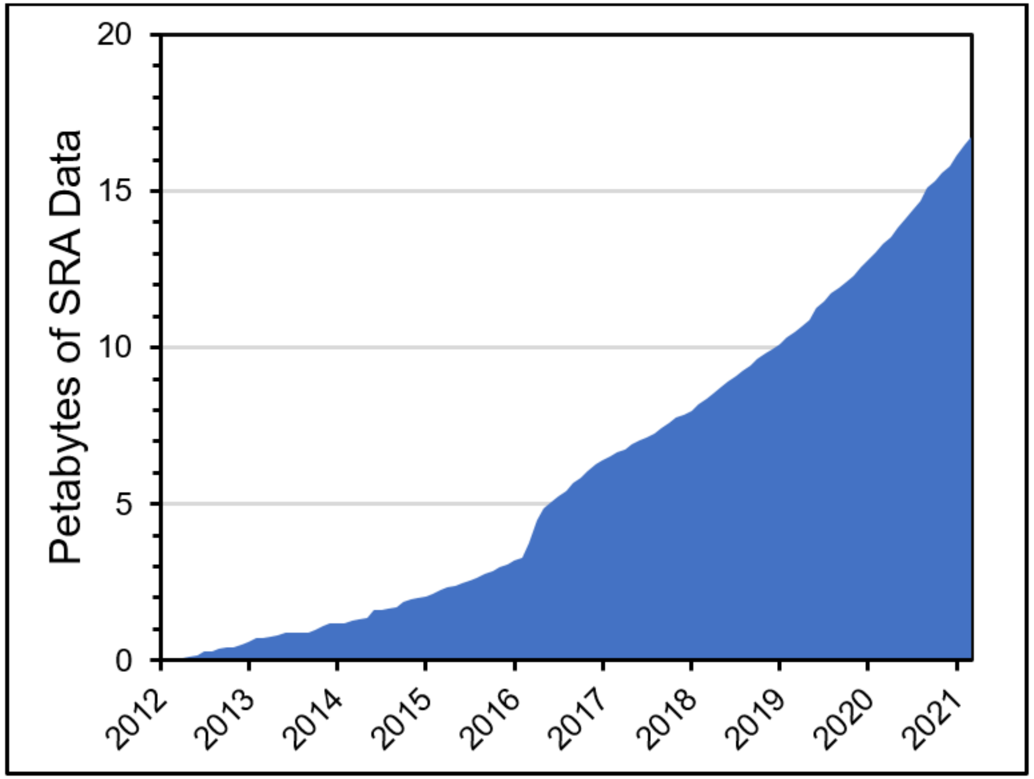
Javni podatkovni repozitoriji

Rast količine podatkov visokozmogljivostnega sekvenciranja DNA (in RNA):

ENA

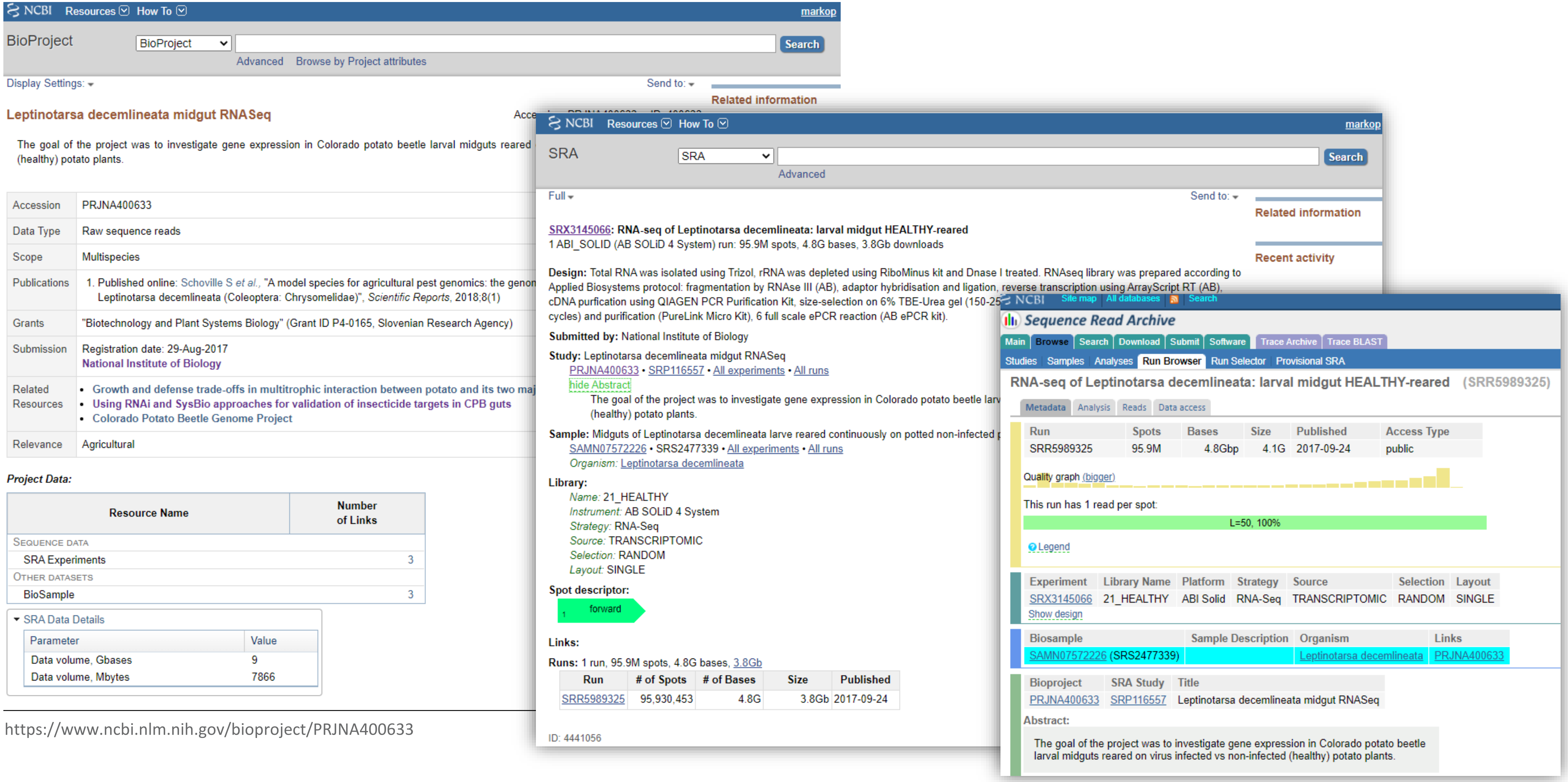


slika: EBI ENA
(https://www.ebi.ac.uk/ena/browser/about/statistics)



slika: NCBI Insights
(https://ncbiinsights.ncbi.nlm.nih.gov/2021/08/09/espsss-workshop/#more-6180)

Deponiranje surovih podatkov v javne repozitorije



Leptinotarsa decemlineata midgut RNASeq

Acce

The goal of the project was to investigate gene expression in Colorado potato beetle larval midguts reared
(healthy) potato plants.

Accession	PRJNA400633
Data Type	Raw sequence reads
Scope	Multispecies
Publications	1. Published online: Schoville S *et al.*, "A model species for agricultural pest genomics: the genon Leptinotarsa decemlineata (Coleoptera: Chrysomelidae)", *Scientific Reports*, 2018;8(1)
Grants	"Biotechnology and Plant Systems Biology" (Grant ID P4-0165, Slovenian Research Agency)
Submission	Registration date: 29-Aug-2017 National Institute of Biology
Related Resources	• Growth and defense trade-offs in multitrophic interaction between potato and its two maj • Using RNAi and SysBio approaches for validation of insecticide targets in CPB guts • Colorado Potato Beetle Genome Project
Relevance	Agricultural

Project Data:

Resource Name	Number of Links
SEQUENCE DATA	
SRA Experiments	3
OTHER DATASETS	
BioSample	3

▾ SRA Data Details

Parameter	Value
Data volume, Gbases	9
Data volume, Mbytes	7866

SRX3145066: **RNA-seq of Leptinotarsa decemlineata: larval midgut HEALTHY-reared**
1 ABI_SOLID (AB SOLiD 4 System) run: 95.9M spots, 4.8G bases, 3.8Gb downloads

Design: Total RNA was isolated using Trizol, rRNA was depleted using RiboMinus kit and Dnase I treated. RNAseq library was prepared according to Applied Biosystems protocol: fragmentation by RNAse III (AB), adaptor hybridisation and ligation, reverse transcription using ArrayScript RT (AB), cDNA purfication using QIAGEN PCR Purification Kit, size-selection on 6% TBE-Urea gel (150-25
cycles) and purification (PureLink Micro Kit), 6 full scale ePCR reaction (AB ePCR kit).

Submitted by: National Institute of Biology

Study: Leptinotarsa decemlineata midgut RNASeq
PRJNA400633 • SRP116557 • All experiments • All runs
hide Abstract
The goal of the project was to investigate gene expression in Colorado potato beetle larv
(healthy) potato plants.

Sample: Midguts of Leptinotarsa decemlineata larve reared continuously on potted non-infected p
SAMN07572226 • SRS2477339 • All experiments • All runs
Organism: Leptinotarsa decemlineata

Library:
Name: 21_HEALTHY
Instrument: AB SOLiD 4 System
Strategy: RNA-Seq
Source: TRANSCRIPTOMIC
Selection: RANDOM
Layout: SINGLE

Spot descriptor:

Links:

Runs: 1 run, 95.9M spots, 4.8G bases, 3.8Gb

Run	# of Spots	# of Bases	Size	Published
SRR5989325	95,930,453	4.8G	3.8Gb	2017-09-24

ID: 4441056

RNA-seq of Leptinotarsa decemlineata: larval midgut HEALTHY-reared (SRR5989325)

Run	Spots	Bases	Size	Published	Access Type
SRR5989325	95.9M	4.8Gbp	4.1G	2017-09-24	public

Quality graph (bigger)

This run has 1 read per spot:

L=50, 100%

Legend

Experiment	Library Name	Platform	Strategy	Source	Selection	Layout
SRX3145066	21_HEALTHY	ABI Solid	RNA-Seq	TRANSCRIPTOMIC	RANDOM	SINGLE

Show design

Biosample	Sample Description	Organism	Links
SAMN07572226 (SRS2477339)		Leptinotarsa decemlineata	PRJNA400633

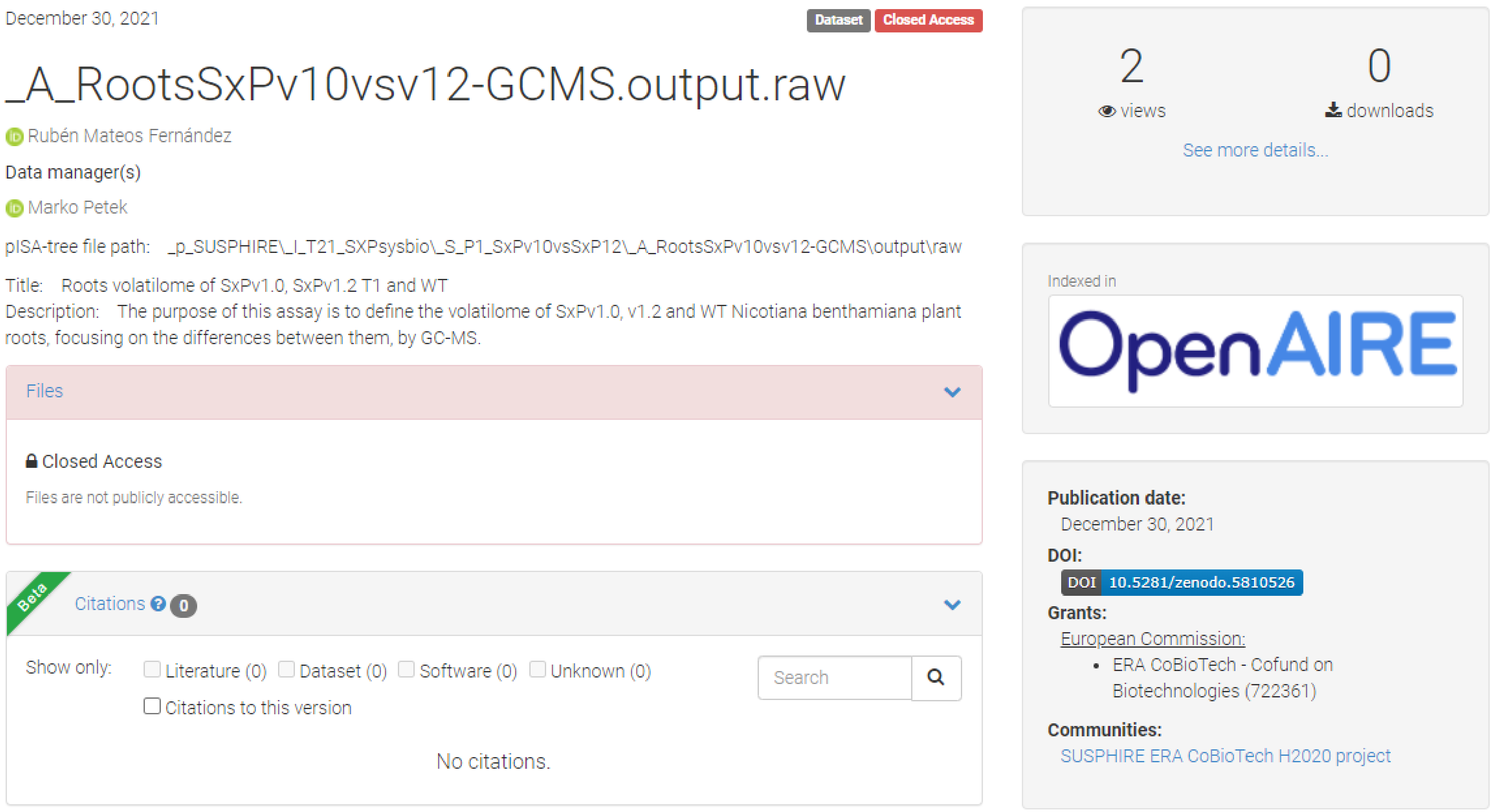
Bioproject	SRA Study	Title
PRJNA400633	SRP116557	Leptinotarsa decemlineata midgut RNASeq

Abstract:

The goal of the project was to investigate gene expression in Colorado potato beetle larval midguts reared on virus infected vs non-infected (healthy) potato plants.

https://www.ncbi.nlm.nih.gov/bioproject/PRJNA400633

Deponiranje surovih podatkov v javne repozitorije



December 30, 2021

Dataset Closed Access

_A_RootsSxPv10vsv12-GCMS.output.raw

Rubén Mateos Fernández

Data manager(s)

Marko Petek

pISA-tree file path: _p_SUSPHIRE_I_T21_SXPsysbio_S_P1_SxPv10vsSxP12_A_RootsSxPv10vsv12-GCMS\output\raw

Title: Roots volatilome of SxPv1.0, SxPv1.2 T1 and WT
Description: The purpose of this assay is to define the volatilome of SxPv1.0, v1.2 and WT Nicotiana benthamiana plant roots, focusing on the differences between them, by GC-MS.

Files

Closed Access

Files are not publicly accessible.

Beta

Citations ? 0

Show only: □ Literature (0) □ Dataset (0) □ Software (0) □ Unknown (0)

□ Citations to this version

Search

No citations.

2 views

0 downloads

See more details...

Indexed in

OpenAIRE

Publication date:
December 30, 2021

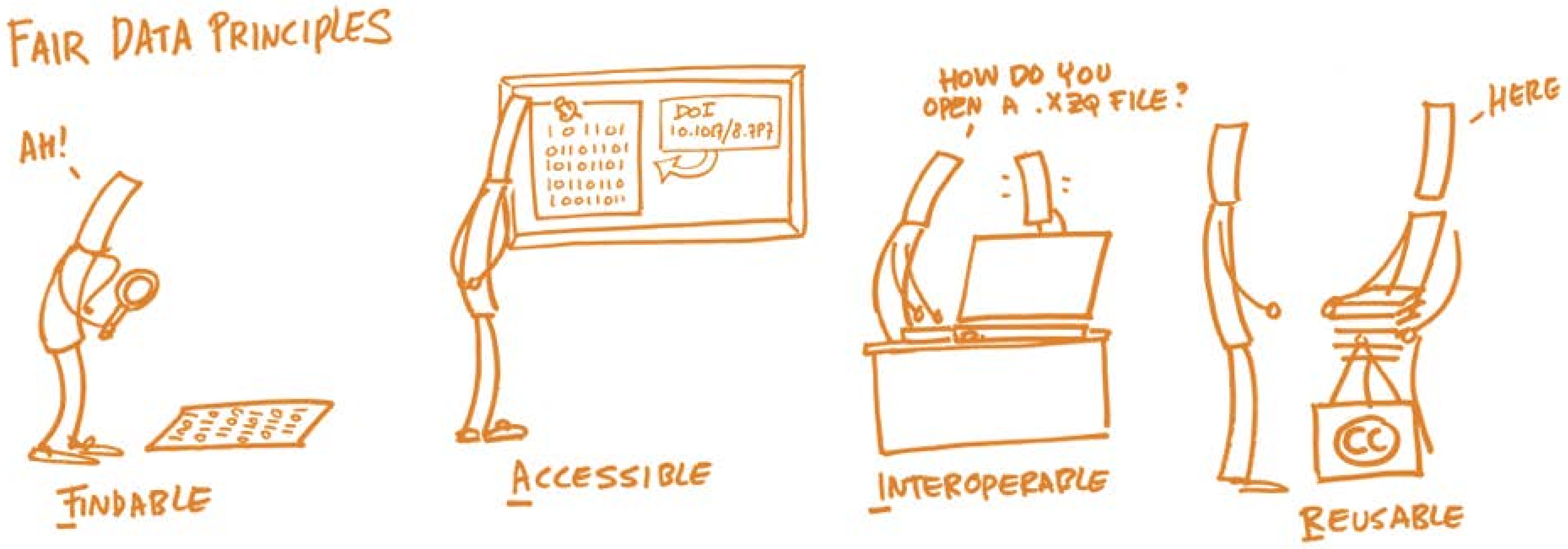
DOI:
DOI 10.5281/zenodo.5810526

Grants:
European Commission:
- ERA CoBioTech - Cofund on Biotechnologies (722361)

Communities:
SUSPHIRE ERA CoBioTech H2020 project

https://doi.org/10.5281/zenodo.5810526

Kaj pomenijo FAIR podatki za nas?



- enoznačni in perzistentni ID-ji
- strojno berljivi metapodatki

- takojšen dostop
- ali jasna pravila za dostop

- datotečni formati
- ontologije
- genski identifikatorji

- metapodatki z ustreznimi attributi za reanalizo
- licence
- standardi različnih ved

slika: The Open Science Training Handbook
https://book.fosteropenscience.eu/
(CC0 1.0 Universal)

Kako ravnamo s podatki preden jih deponiramo v javnih repozitorijih?

Kako (na NIB) ravnamo s podatki preden jih deponiramo v javnih repozitorijih?

- Surovi podatki
- Metapodatki
- Vmesni in končni rezultati obdelave podatkov in modeliranja
- Publikacije (članki, poročila, predstavitve, posterji etc)

stork

mrežni strežnik z varnostno kopijo in omejenim dostopom

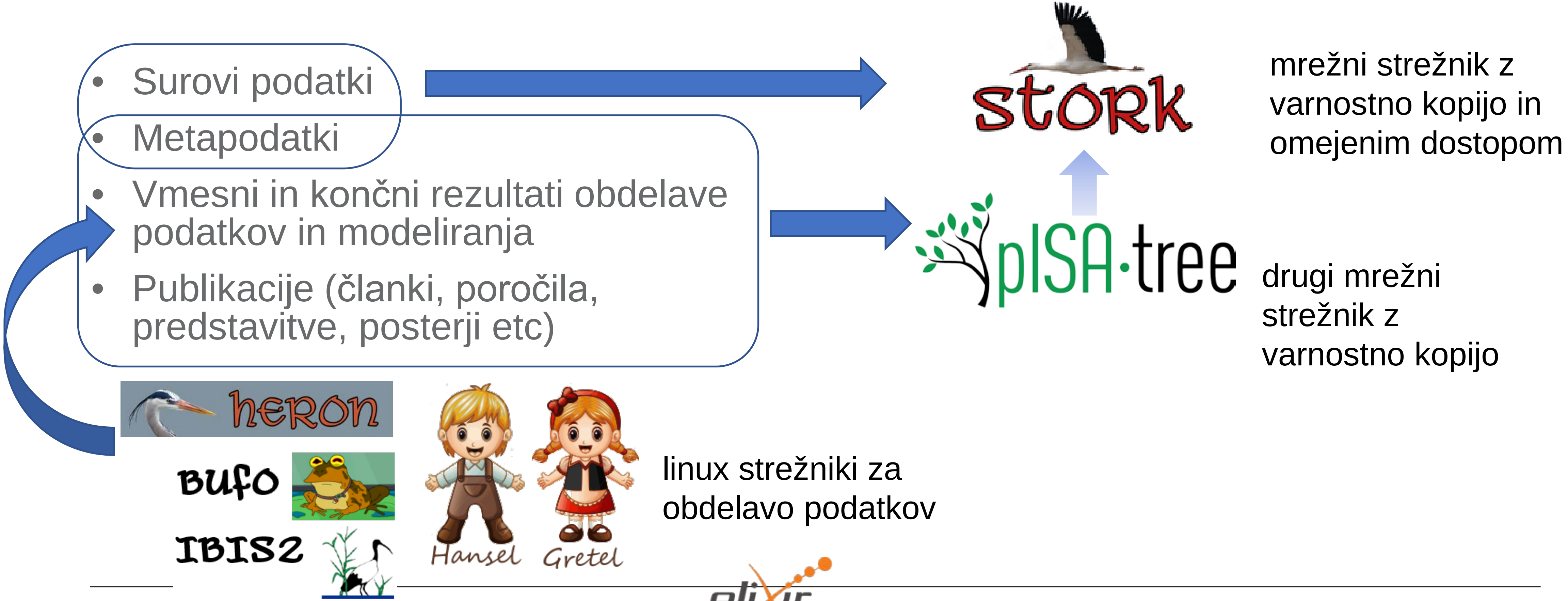
Kako (na NIB) ravnamo s podatki preden jih deponiramo v javnih repozitorijih?

- Surovi podatki
- Metapodatki
- Vmesni in končni rezultati obdelave podatkov in modeliranja
- Publikacije (članki, poročila, predstavitve, posterji etc)

mrežni strežnik z varnostno kopijo in omejenim dostopom

linux strežniki za obdelavo podatkov

Kako (na NIB) ravnamo s podatki preden jih deponiramo v javnih repozitorijih?



- Surovi podatki
- Metapodatki
- Vmesni in končni rezultati obdelave podatkov in modeliranja
- Publikacije (članki, poročila, predstavitve, posterji etc)

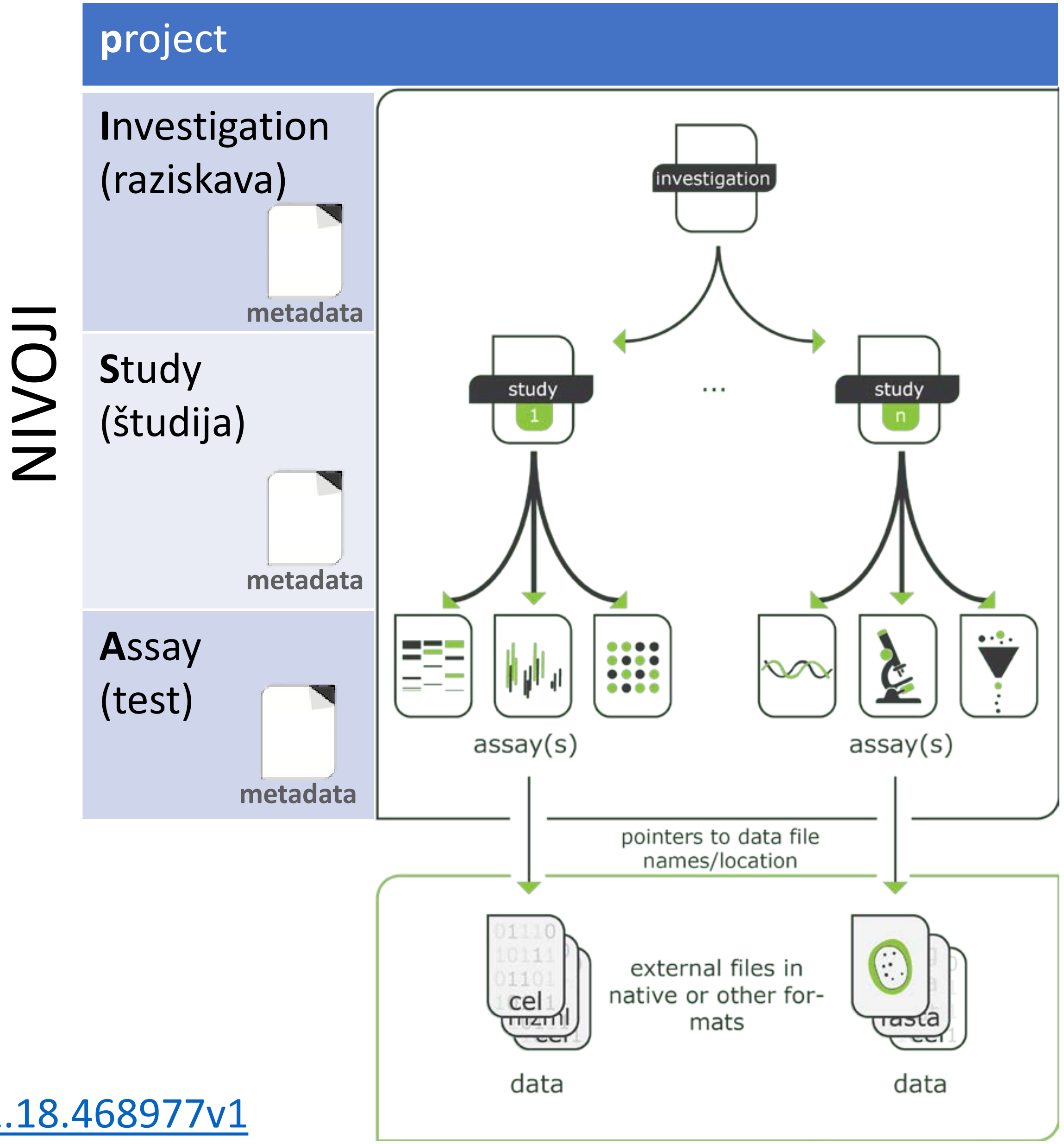
mrežni strežnik z varnostno kopijo in omejenim dostopom

drugi mrežni strežnik z varnostno kopijo

linux strežniki za obdelavo podatkov

pISA-tree

- Sistem za organizacijo projektnih podatkov (*.**bat** skripte)
- Datotečna drevesna struktura po **ISA specifikacijah**
- metapodatkovne *.txt datoteke v **ISA-Tab format** (sprotno beleženje metapodatkov)



NIVOJI

project

Investigation (raziskava)

metadata

Study (študija)

metadata

Assay (test)

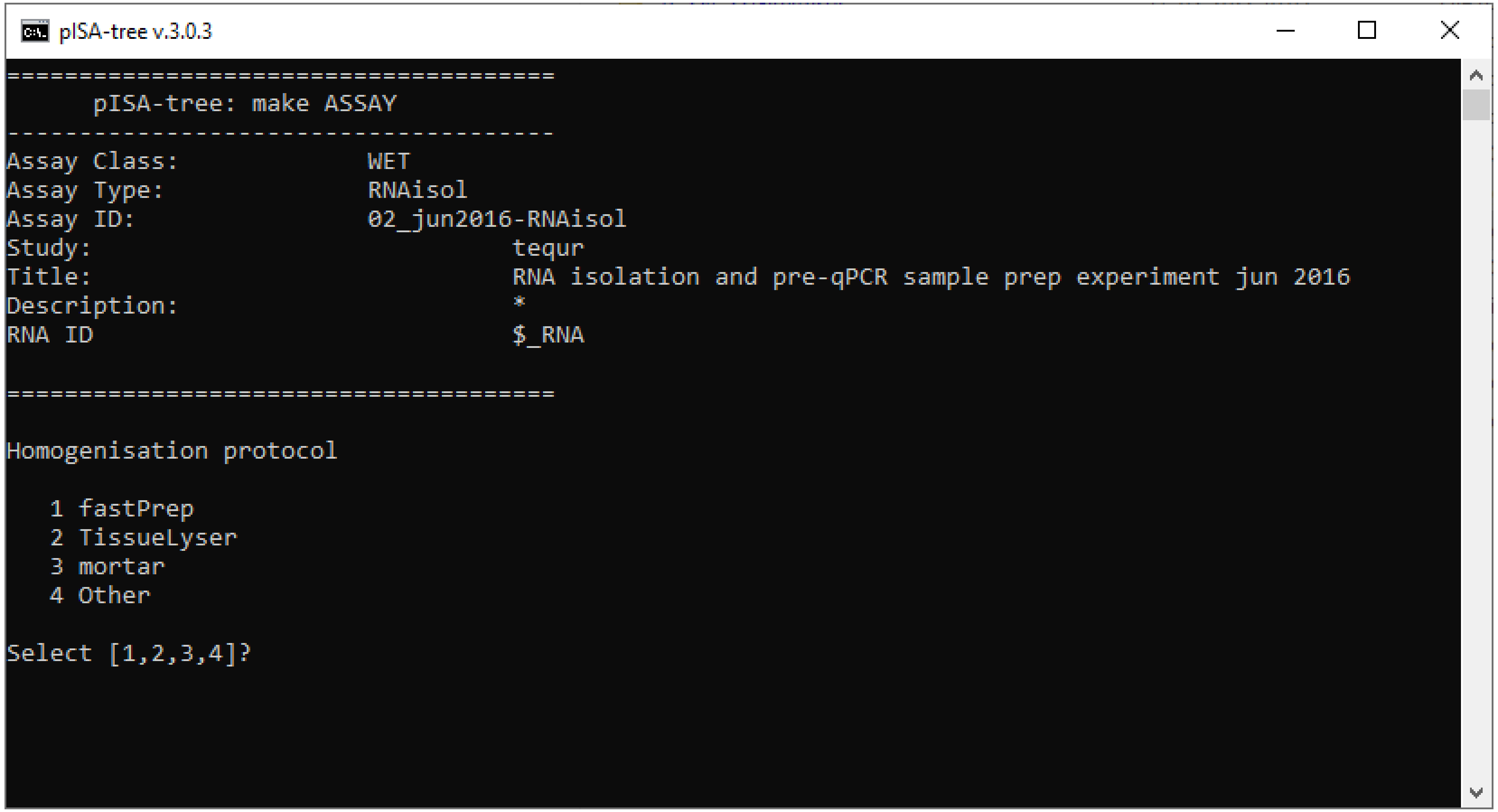
metadata

Image: https://isa-tools.org/format/specification.html

https://github.com/NIB-SI/pISA-tree

Prednatis članka:
https://www.biorxiv.org/content/10.1101/2021.11.18.468977v1

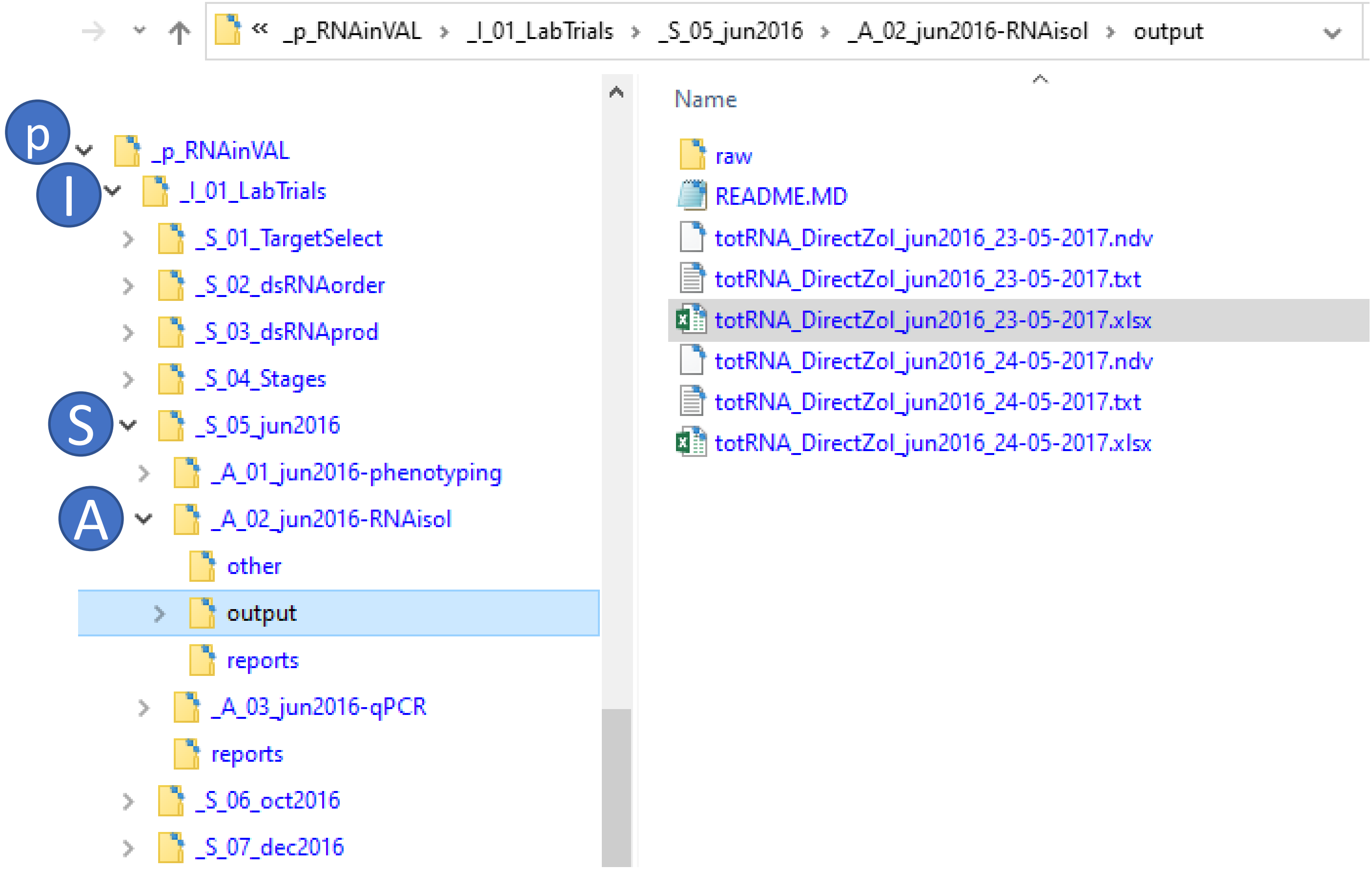
pISA-tree: ustvarjanje nivojev



```
======================================
      pISA-tree: make ASSAY
--------------------------------------
Assay Class:             WET
Assay Type:              RNAisol
Assay ID:                02_jun2016-RNAisol
Study:                             tequr
Title:                             RNA isolation and pre-qPCR sample prep experiment jun 2016
Description:                       *
RNA ID                             $_RNA

======================================

Homogenisation protocol

   1 fastPrep
   2 TissueLyser
   3 mortar
   4 Other

Select [1,2,3,4]?
```

https://github.com/NIB-SI/pISA-tree

pISA-tree: primer lokalne strukture



« _p_RNAinVAL › _I_01_LabTrials › _S_05_jun2016 › _A_02_jun2016-RNAisol › output

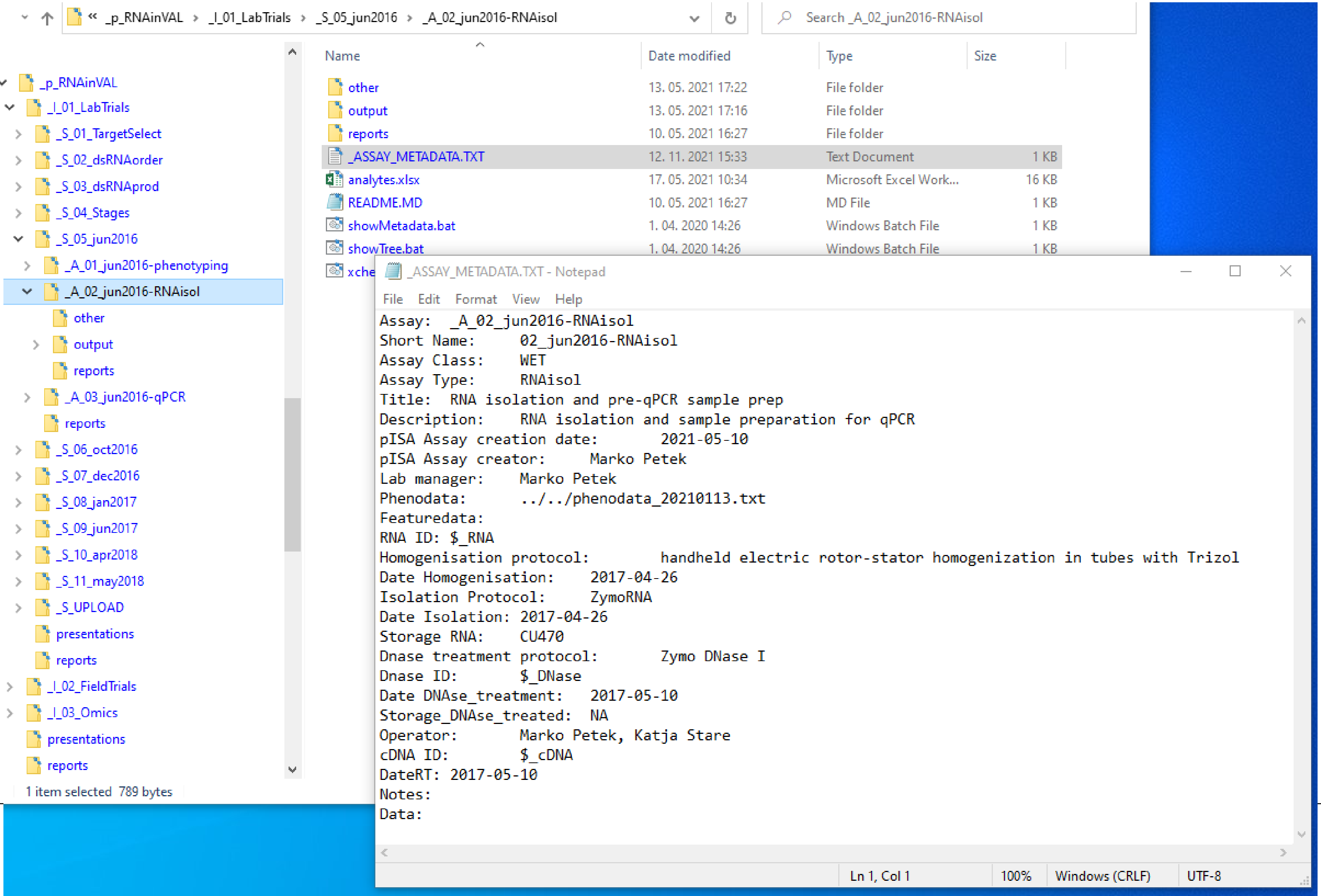
- _p_RNAinVAL
  - _I_01_LabTrials
    - _S_01_TargetSelect
    - _S_02_dsRNAorder
    - _S_03_dsRNAprod
    - _S_04_Stages
    - _S_05_jun2016
      - _A_01_jun2016-phenotyping
      - _A_02_jun2016-RNAisol
        - other
        - output
        - reports
      - _A_03_jun2016-qPCR
      - reports
    - _S_06_oct2016
    - _S_07_dec2016

Name

- raw
- README.MD
- totRNA_DirectZol_jun2016_23-05-2017.ndv
- totRNA_DirectZol_jun2016_23-05-2017.txt
- totRNA_DirectZol_jun2016_23-05-2017.xlsx
- totRNA_DirectZol_jun2016_24-05-2017.ndv
- totRNA_DirectZol_jun2016_24-05-2017.txt
- totRNA_DirectZol_jun2016_24-05-2017.xlsx

https://github.com/NIB-SI/pISA-tree

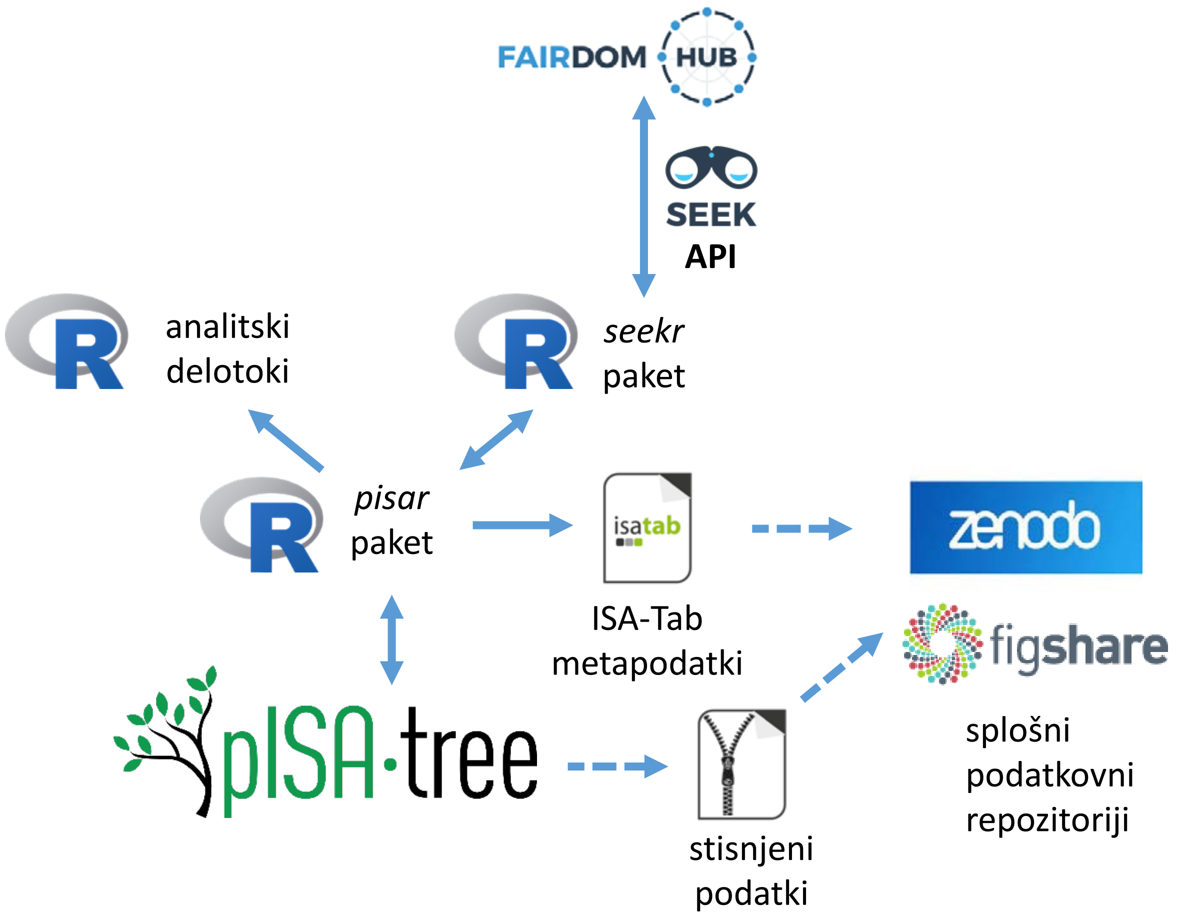
pISA-tree: primer lokalne strukture



Name	Date modified	Type	Size
other	13. 05. 2021 17:22	File folder	
output	13. 05. 2021 17:16	File folder	
reports	10. 05. 2021 16:27	File folder	
_ASSAY_METADATA.TXT	12. 11. 2021 15:33	Text Document	1 KB
analytes.xlsx	17. 05. 2021 10:34	Microsoft Excel Work...	16 KB
README.MD	10. 05. 2021 16:27	MD File	1 KB
showMetadata.bat	1. 04. 2020 14:26	Windows Batch File	1 KB
showTree.bat	1. 04. 2020 14:26	Windows Batch File	1 KB
xche			

```
Assay:  _A_02_jun2016-RNAisol
Short Name:     02_jun2016-RNAisol
Assay Class:    WET
Assay Type:     RNAisol
Title:  RNA isolation and pre-qPCR sample prep
Description:    RNA isolation and sample preparation for qPCR
pISA Assay creation date:       2021-05-10
pISA Assay creator:     Marko Petek
Lab manager:    Marko Petek
Phenodata:      ../../phenodata_20210113.txt
Featuredata:
RNA ID: $_RNA
Homogenisation protocol:        handheld electric rotor-stator homogenization in tubes with Trizol
Date Homogenisation:    2017-04-26
Isolation Protocol:     ZymoRNA
Date Isolation: 2017-04-26
Storage RNA:    CU470
Dnase treatment protocol:       Zymo DNase I
Dnase ID:       $_DNase
Date DNAse_treatment:   2017-05-10
Storage_DNAse_treated:  NA
Operator:       Marko Petek, Katja Stare
cDNA ID:        $_cDNA
DateRT: 2017-05-10
Notes:
Data:
```

pISA-tree: interoperabilnost



FAIRDOM HUB

SEEK API

analitski delotoki

seekr paket

pisar paket

ISA-Tab metapodatki

stisnjeni podatki

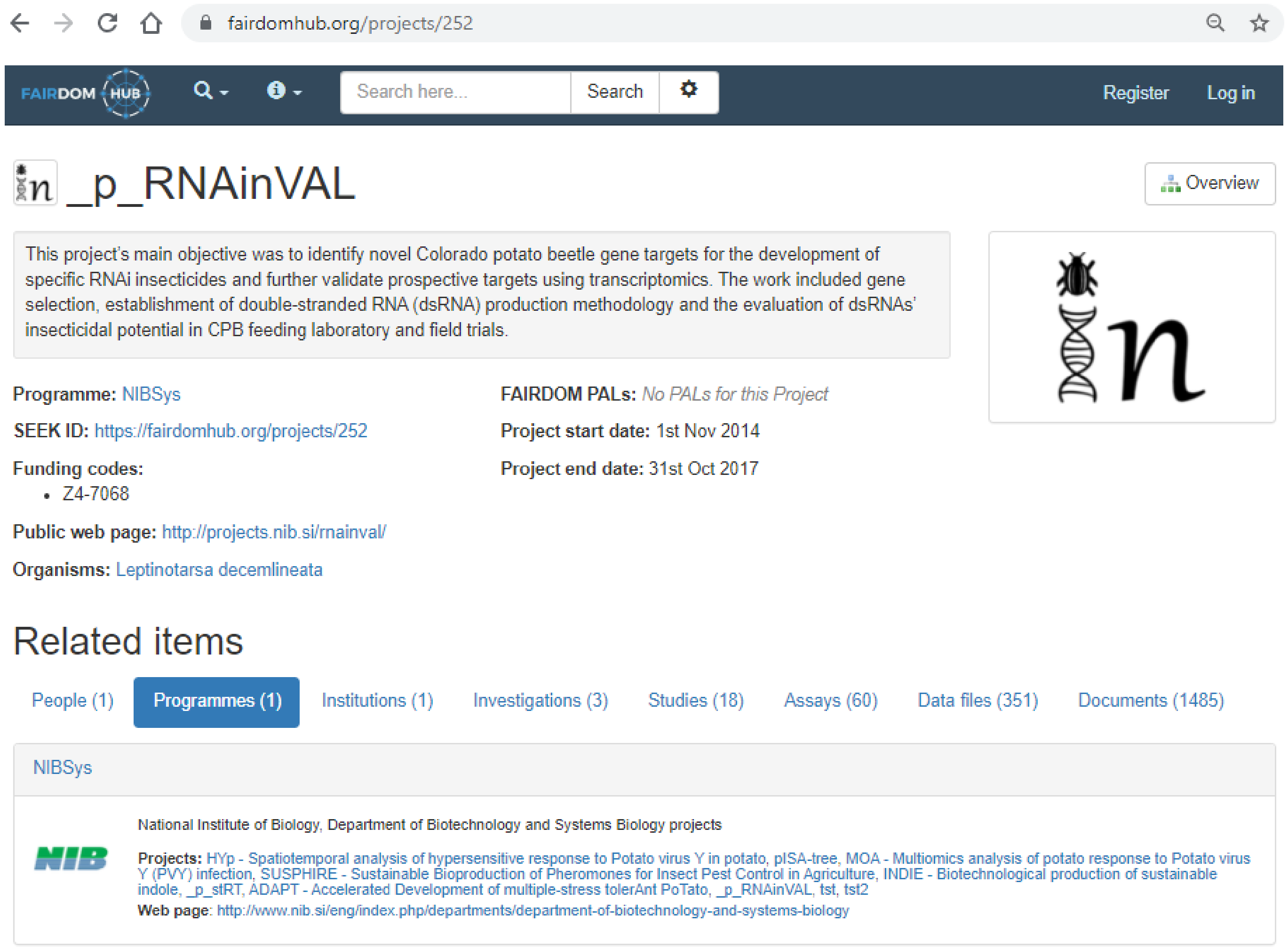
splošni podatkovni repozitoriji

https://github.com/NIB-SI/pISA-tree https://github.com/NIB-SI/seekr https://github.com/NIB-SI/pisar

pISA-tree --> FAIRDOMHub.org

- FAIRDOMHub
  - uporablja ISA format
  - na voljo večina licenc za odprto kodo in podatke (Creative Commons, Open Data Commons, …)
  - omogoča pridobitev DOI za podatke

- Prenos v programskem okolju R s paketom *seekr*(https://github.com/NIB-SI/seekr)



_p_RNAinVAL

This project's main objective was to identify novel Colorado potato beetle gene targets for the development of specific RNAi insecticides and further validate prospective targets using transcriptomics. The work included gene selection, establishment of double-stranded RNA (dsRNA) production methodology and the evaluation of dsRNAs' insecticidal potential in CPB feeding laboratory and field trials.

Programme: NIBSys

SEEK ID: https://fairdomhub.org/projects/252

Funding codes:
- Z4-7068

Public web page: http://projects.nib.si/rnainval/

Organisms: Leptinotarsa decemlineata

FAIRDOM PALs: *No PALs for this Project*

Project start date: 1st Nov 2014

Project end date: 31st Oct 2017

Related items

People (1) | Programmes (1) | Institutions (1) | Investigations (3) | Studies (18) | Assays (60) | Data files (351) | Documents (1485)

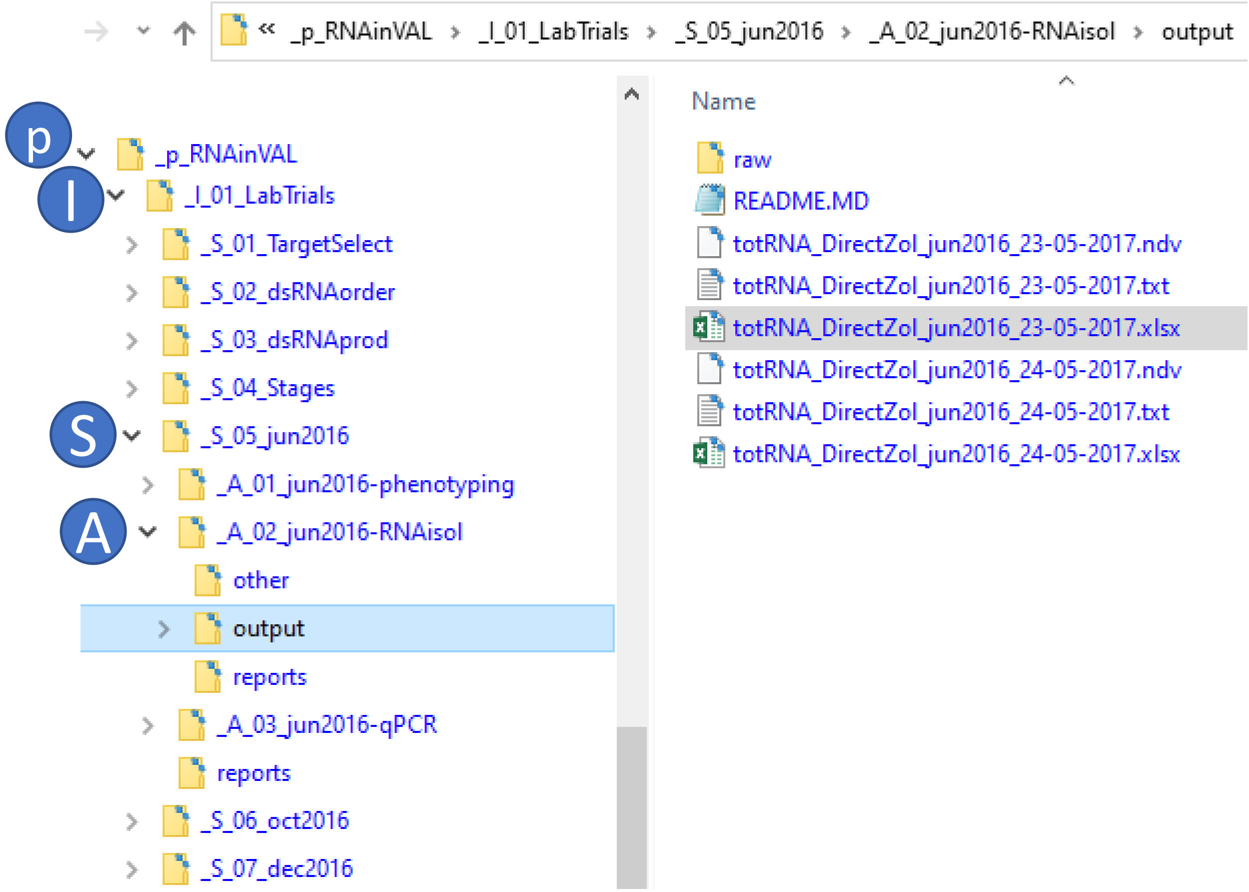
NIBSys

National Institute of Biology, Department of Biotechnology and Systems Biology projects

Projects: HYp - Spatiotemporal analysis of hypersensitive response to Potato virus Y in potato, pISA-tree, MOA - Multiomics analysis of potato response to Potato virus Y (PVY) infection, SUSPHIRE - Sustainable Bioproduction of Pheromones for Insect Pest Control in Agriculture, INDIE - Biotechnological production of sustainable indole, _p_stRT, ADAPT - Accelerated Development of multiple-stress tolerAnt PoTato, _p_RNAinVAL, tst, tst2

Web page: http://www.nib.si/eng/index.php/departments/department-of-biotechnology-and-systems-biology

pISA-tree ---> FAIRDOMHub.org

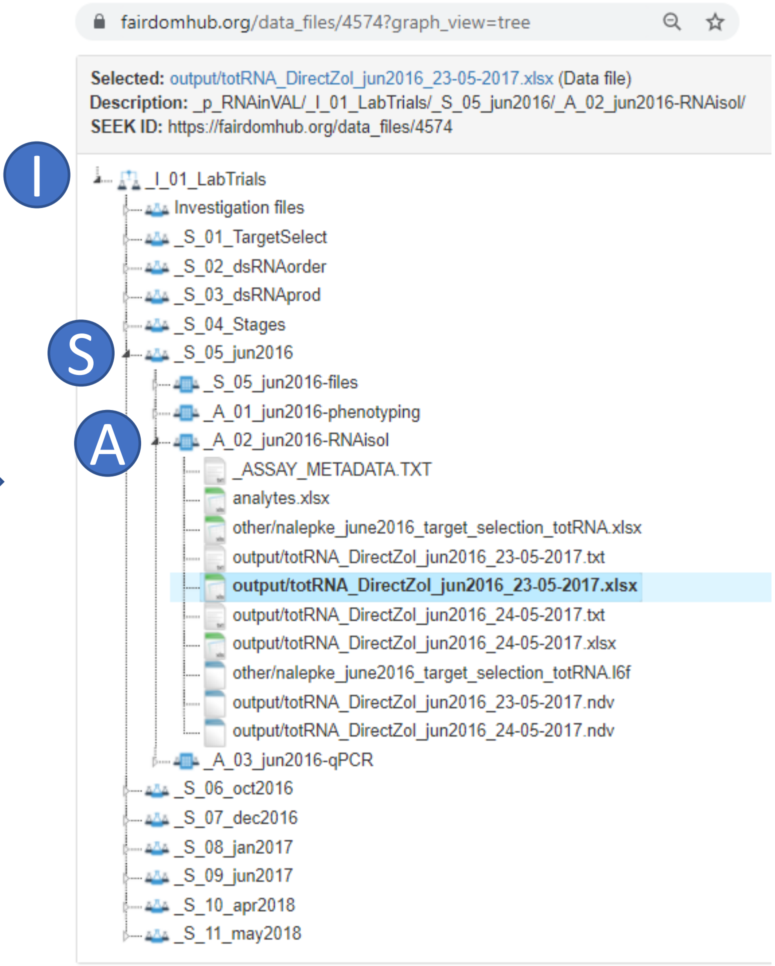


```
« _p_RNAinVAL › _I_01_LabTrials › _S_05_jun2016 › _A_02_jun2016-RNAisol › output
```

```
_p_RNAinVAL
  _I_01_LabTrials
    _S_01_TargetSelect
    _S_02_dsRNAorder
    _S_03_dsRNAprod
    _S_04_Stages
    _S_05_jun2016
      _A_01_jun2016-phenotyping
      _A_02_jun2016-RNAisol
        other
        output
        reports
      _A_03_jun2016-qPCR
      reports
    _S_06_oct2016
    _S_07_dec2016
```

```
Name
raw
README.MD
totRNA_DirectZol_jun2016_23-05-2017.ndv
totRNA_DirectZol_jun2016_23-05-2017.txt
totRNA_DirectZol_jun2016_23-05-2017.xlsx
totRNA_DirectZol_jun2016_24-05-2017.ndv
totRNA_DirectZol_jun2016_24-05-2017.txt
totRNA_DirectZol_jun2016_24-05-2017.xlsx
```

seekr



fairdomhub.org/data_files/4574?graph_view=tree

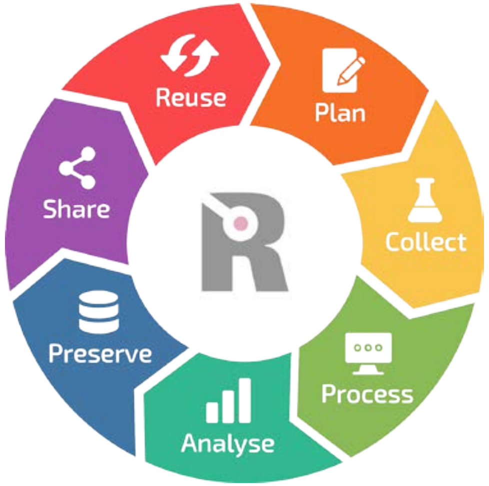
Selected: output/totRNA_DirectZol_jun2016_23-05-2017.xlsx (Data file)
Description: _p_RNAinVAL/_I_01_LabTrials/_S_05_jun2016/_A_02_jun2016-RNAisol/
SEEK ID: https://fairdomhub.org/data_files/4574

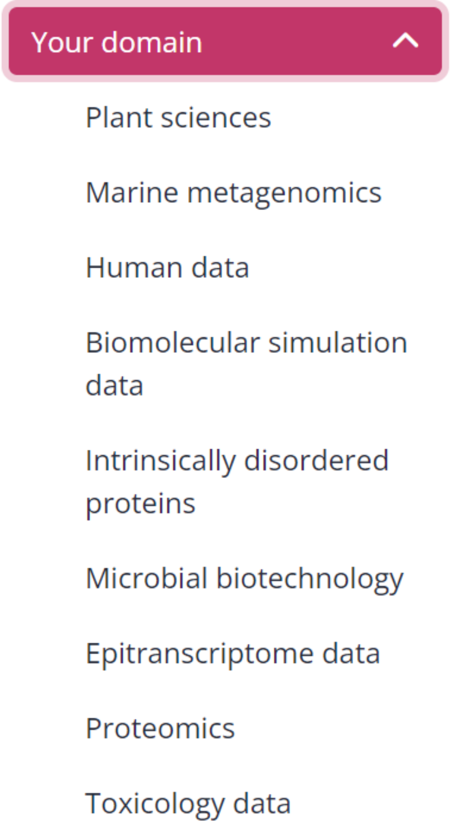
```
_I_01_LabTrials
  Investigation files
  _S_01_TargetSelect
  _S_02_dsRNAorder
  _S_03_dsRNAprod
  _S_04_Stages
  _S_05_jun2016
    _S_05_jun2016-files
    _A_01_jun2016-phenotyping
    _A_02_jun2016-RNAisol
      _ASSAY_METADATA.TXT
      analytes.xlsx
      other/nalepke_june2016_target_selection_totRNA.xlsx
      output/totRNA_DirectZol_jun2016_23-05-2017.txt
      output/totRNA_DirectZol_jun2016_23-05-2017.xlsx
      output/totRNA_DirectZol_jun2016_24-05-2017.txt
      output/totRNA_DirectZol_jun2016_24-05-2017.xlsx
      other/nalepke_june2016_target_selection_totRNA.l6f
      output/totRNA_DirectZol_jun2016_23-05-2017.ndv
      output/totRNA_DirectZol_jun2016_24-05-2017.ndv
    _A_03_jun2016-qPCR
  _S_06_oct2016
  _S_07_dec2016
  _S_08_jan2017
  _S_09_jun2017
  _S_10_apr2018
  _S_11_may2018
```

https://github.com/NIB-SI/pISA-tree https://fairdomhub.org/

ELIXIR Research Data Management Kit

- Spletni vodnik o dobrih praksah ravnanja s podatki za **celoten življenjski cikel podatkov** (know-how, orodja, primeri najboljše prakse)
- Namenjen raziskovalcem, upravljavcem podatkov in oblikovalcem politik




Your domain

Plant sciences

Marine metagenomics

Human data

Biomolecular simulation data

Intrinsically disordered proteins

Microbial biotechnology

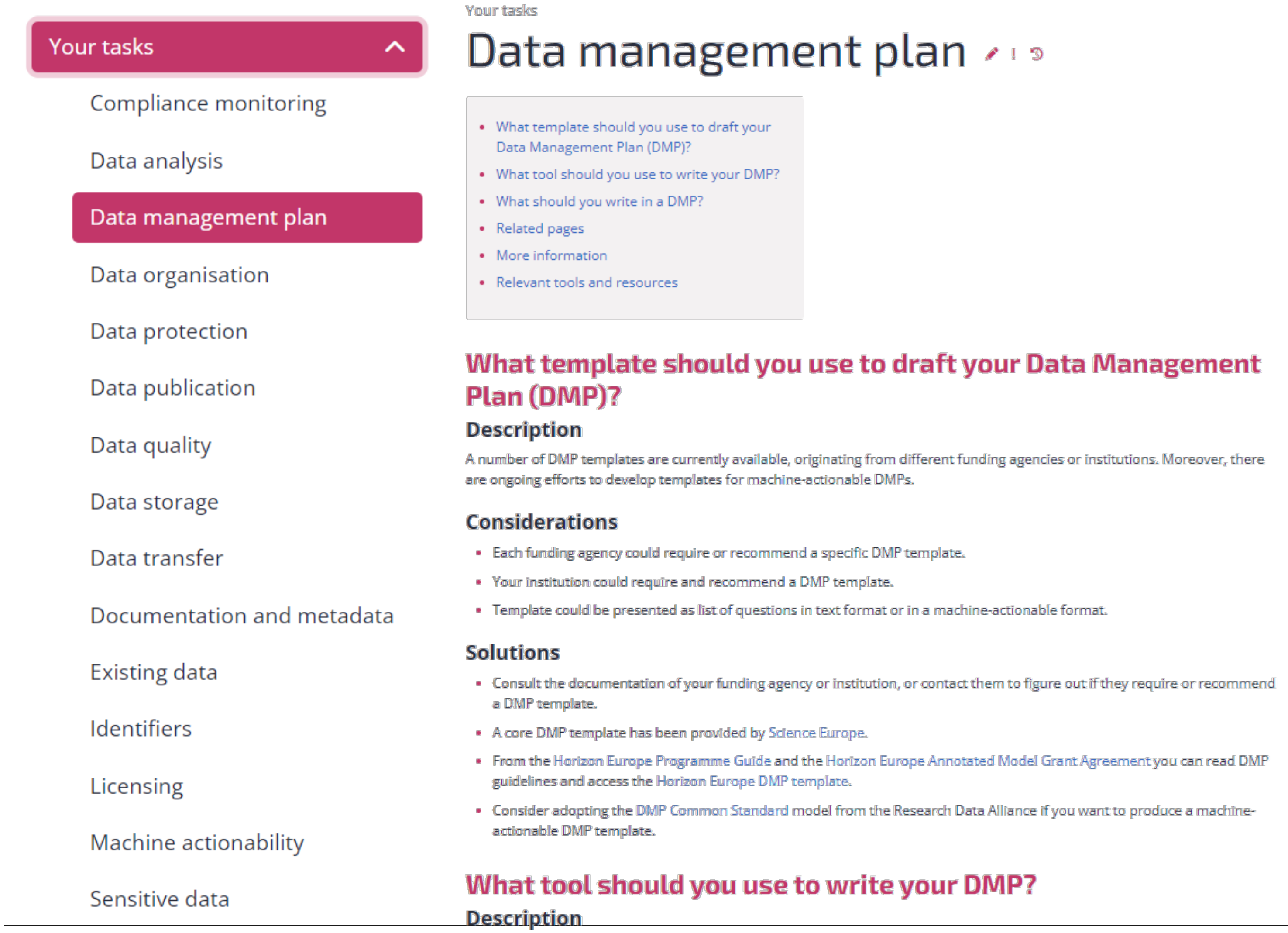
Epitranscriptome data

Proteomics

Toxicology data

https://rdmkit.elixir-europe.org/

ELIXIR Research Data Management Kit



Your tasks

- Compliance monitoring
- Data analysis
- Data management plan
- Data organisation
- Data protection
- Data publication
- Data quality
- Data storage
- Data transfer
- Documentation and metadata
- Existing data
- Identifiers
- Licensing
- Machine actionability
- Sensitive data

Your tasks

Data management plan

- What template should you use to draft your Data Management Plan (DMP)?
- What tool should you use to write your DMP?
- What should you write in a DMP?
- Related pages
- More information
- Relevant tools and resources

What template should you use to draft your Data Management Plan (DMP)?

Description

A number of DMP templates are currently available, originating from different funding agencies or institutions. Moreover, there are ongoing efforts to develop templates for machine-actionable DMPs.

Considerations

- Each funding agency could require or recommend a specific DMP template.
- Your institution could require and recommend a DMP template.
- Template could be presented as list of questions in text format or in a machine-actionable format.

Solutions

- Consult the documentation of your funding agency or institution, or contact them to figure out if they require or recommend a DMP template.
- A core DMP template has been provided by Science Europe.
- From the Horizon Europe Programme Guide and the Horizon Europe Annotated Model Grant Agreement you can read DMP guidelines and access the Horizon Europe DMP template.
- Consider adopting the DMP Common Standard model from the Research Data Alliance if you want to produce a machine-actionable DMP template.

What tool should you use to write your DMP?

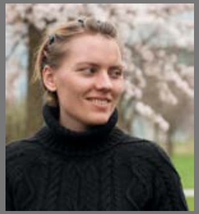
Description

https://rdmkit.elixir-europe.org/data_management_plan

Skupina razvijalcev pISA-tree

prof. dr. Kristina Gruden

prof. dr. Andrej Blejec



dr. Maja Zagorščak

doc. dr. Špela Baebler

dr. Živa Ramšak

dr. Anna Coll Rius

dr. Marko Petek

dr. Tjaša Lukan

Katja Stare

Valentina Levak

marko.petek@nib.si