Mokslo duomenų valdymas, FAIR' duomenys

lekt. dr. Andrius Kriščiūnas

Kauno Technologijos Universitetas

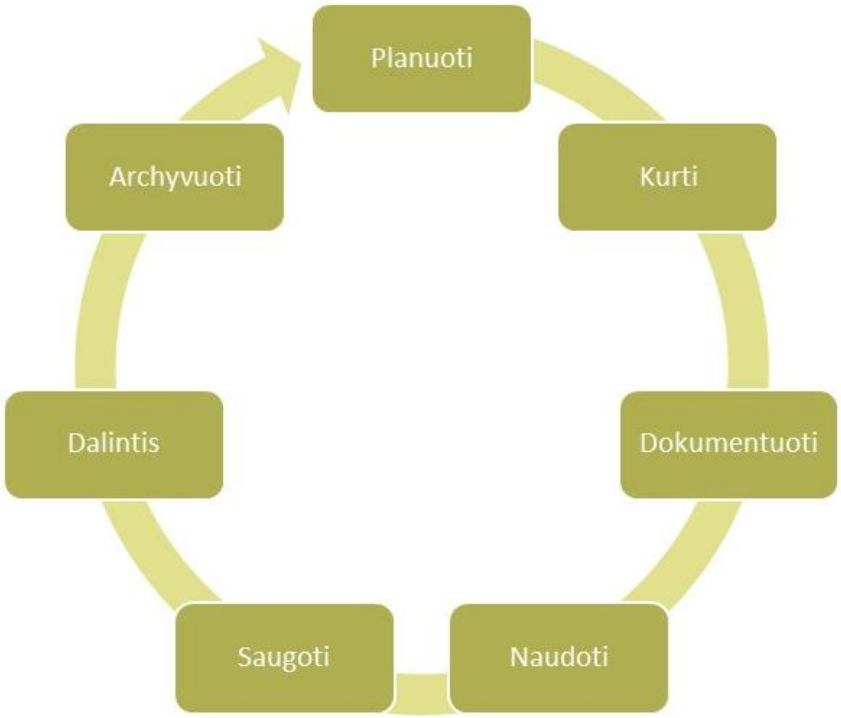
andrius.krisciunas@ktu.lt

2019 02 15

Mokslinių duomenų specifika

Yra per daug mokslo sričių, kad būtų galima vienareikšmiškai parinkti vieną duomenų saugojimo standartą, kuris tiktų joms visoms.

Svarbu kiekvienas periodas!



* mažai prasmės duomenis kurti, saugoti, archyvuoti, jei jų vėliau nebus galima **pakartotinai panaudoti**.

FAIR

FAIR' duomenys – rekomendacijos / koncepcija mokslinių duomenų saugojimui

https://www.go-fair.org/fair-principles/

Surandami	**Prieinami**	**Suderinami**	**Pakartotinai Panaudojami**
Findable	*Accessible*	*Interoperable*	*Reusable*

Findable, Accessible,

Surandami

F1. Duomenims priskiriamas unikalus ir nuolatinis identifikatorius

F2. Duomenims detaliai aprašomi metaduomenimis

....

Prieinami

A1. Duomenys, žinant jų identifikatorių, pasiekiami standartiniais protokolais

A2. Metaduomenys prieinami, net jeigu duomenys jau nebesaugomi

https://www.go-fair.org/fair-principles/

Findable, Accessible,...

F3. Metadata clearly and explicitly include the identifier of the data they describe

What does this mean?

This is a simple and obvious principle, but of critical importance to FAIR. The metadata and the dataset they describe are usually separate files. The association between a metadata file and the dataset should be made explicit by mentioning a dataset's globally unique and persistent identifier in the metadata. As stated in F1, many repositories will generate globally unique and persistent identifiers for deposited datasets that can be used for this purpose.

Example

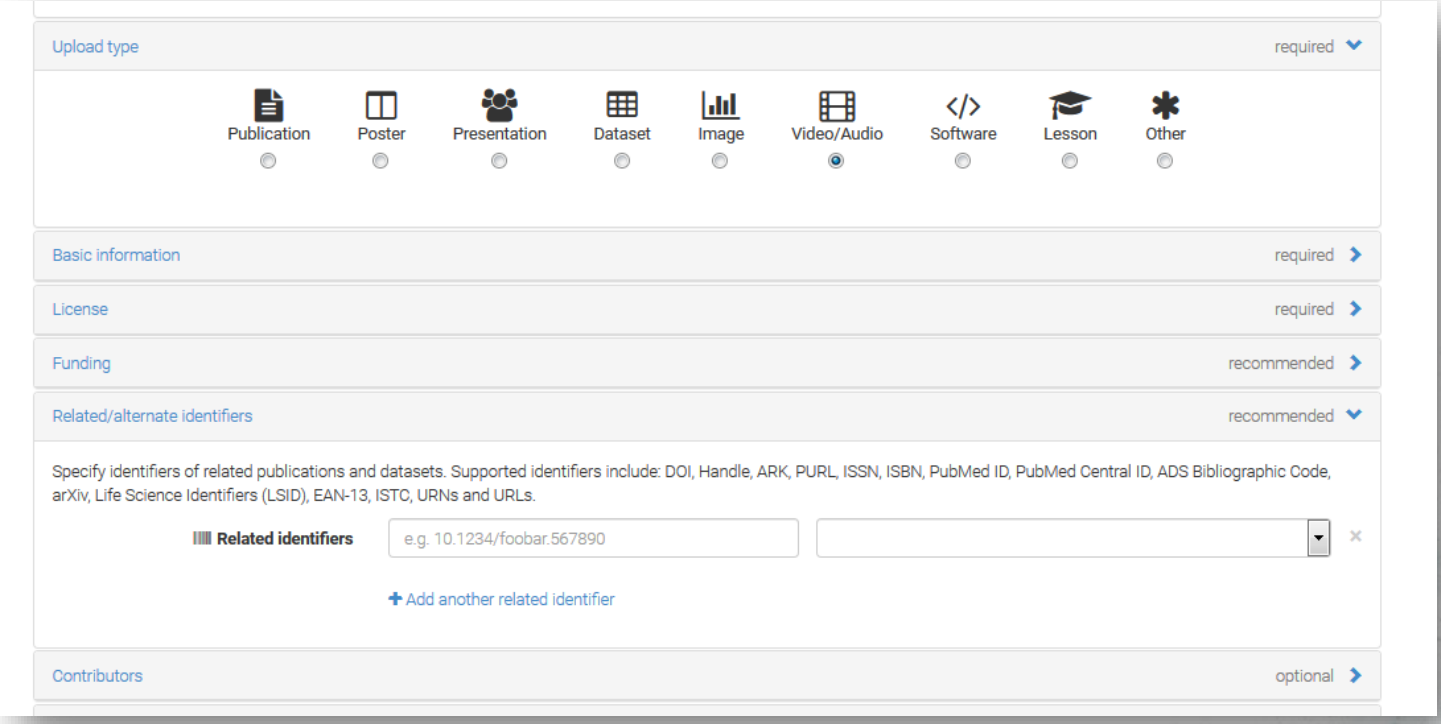
The connection should be annotated in a formal manner, for example using the foaf:primaryTopic predicate in the case of RDF metadata.

Links to Resources

The **DTL FAIRifier tool** guarantees F3.

https://www.go-fair.org/fair-principles/

Findable, Accessible,...



https://zenodo.org/deposit/new

FAIR

Surandami	**Prieinami**	**Suderinami**	**Pakartotinai Panaudojami**
Findable	*Accessible*	*Interoperable*	*Reusable*

1. Teisingos **duomenų talpyklos** pasirinkimas
2. Išsamiai aprašyti metaduomenys

1. Tinkamo **duomenų formato** pasirinkimas
2. Išsamiai aprašyti duomenys

…, Interoperable, Reusable

Vienas iš dažniausiai iškylančių klausimų rengiant duomenis ilgalaikiam išsaugojimui, yra: **kokį duomenų formatą parinkti?**

Failų formatai

Priklausomai nuo failo tipo juose saugoma informacija būna skirta:

- **Perskaityti žmogui**
  (angl. *Human readable files*)
- **Perskaityti kompiuteriui**
  (angl. *Machine readable files*)
- **Perskaityti žmogui ir kompiuteriui**
  (angl. *Human and Machine readable files*)

Failo_pavadinimas**.failo_tipas**

failo tipų pvz.: **.txt**, **.csv**, **.bin**, **.xml**, **.exe** ir kt.

(angl. *File Extension)*

Failų formatai

Kaip saugoma informacija faile?

kompiuteryje kiekvienas failas yra vienetų ir nulių (bitų) eilutė

...1 1 0 0 0 1 1 1 0 1 0 1 1 1 0 0 1 1 0 0 1 1 0...

Binariniai failai	Tekstiniai failai
Binariniai failai neturi būdingų apribojimų (gali būti bet kurių baitų sekų), tačiau turi būti **atidaromi su tinkama** programa, kuri žino konkretų failo formatą	Tekstiniai failų bitų sekos saugo informaciją apie tekstą, ir tokie failai gali būti atidaromi bet kurioje teksto redagavimo programoje.

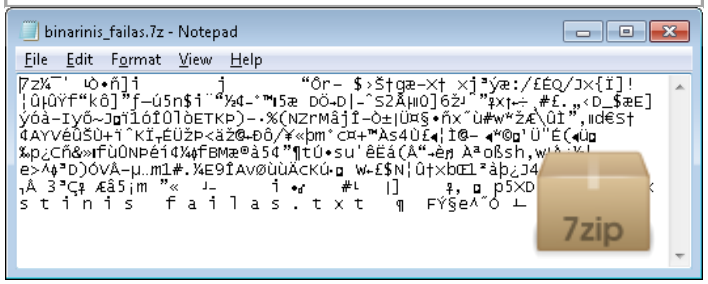
Failų formatai

Kaip saugoma informacija faile?

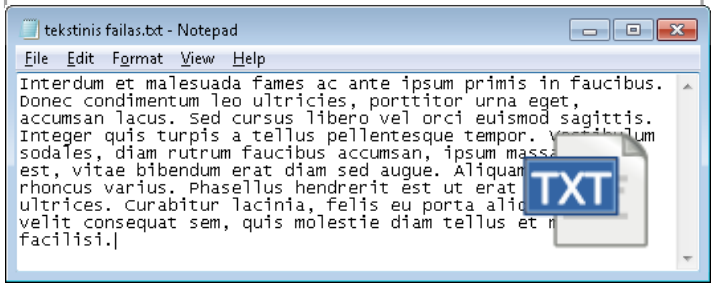
kompiuteryje kiekvienas failas yra vienetų ir nulių (bitų) eilutė

...1 1 0 0 0 1 1 1 0 1 0 1 1 1 0 0 1 1 0 0 1 1 0...

Binariniai failai

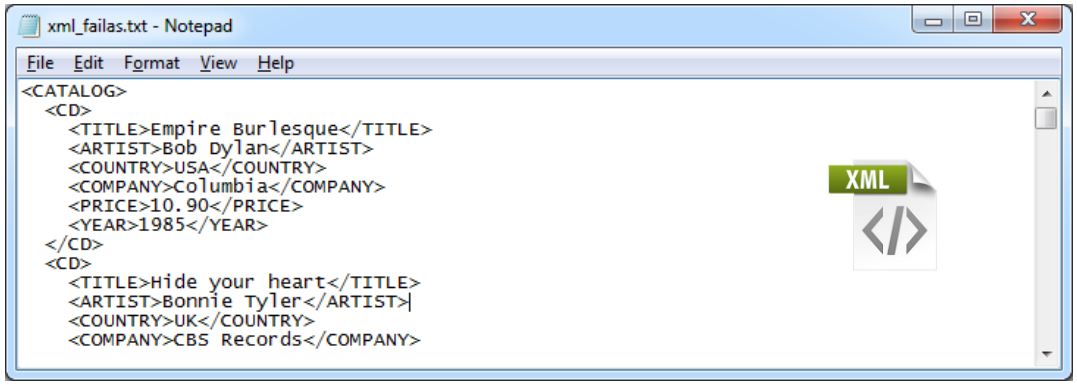


Tekstiniai failai



Failų formatai

Informacija skirta perskaityti **žmogui ir kompiuteriui**

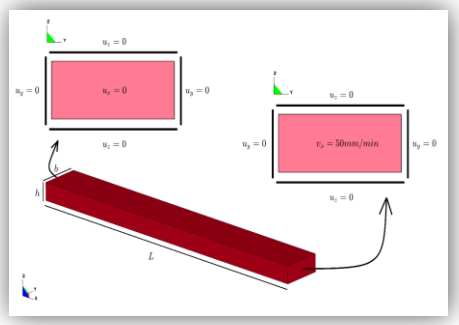


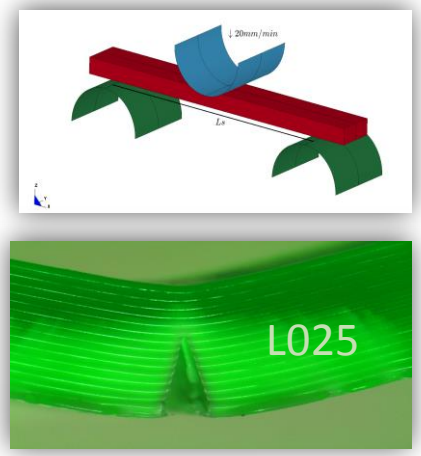
....

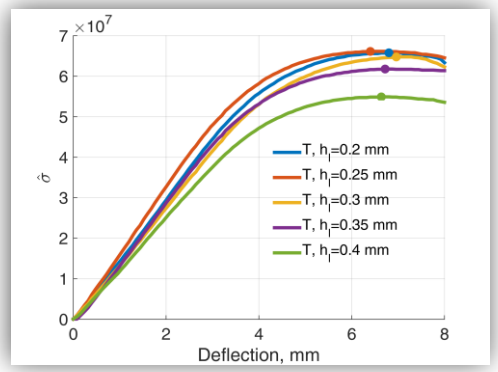
Ar visi duomenys gali būti išsaugoti šiais formatais?

Multi-scale Finite Element Modeling of 3D Printed Structures Subjected to Mechanical Loads

Modelio sudarymas → Modeliavimas / Mechaniniai tyrimai → Modeliavimo rezultatai ir jų analizė





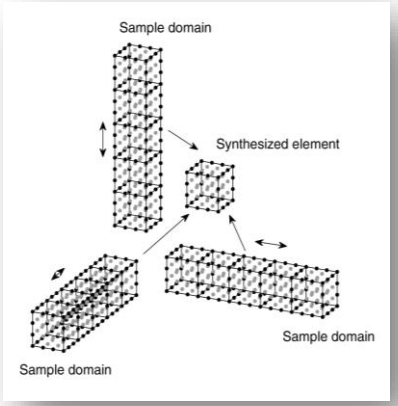


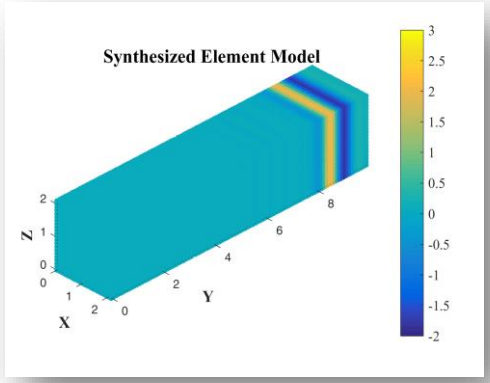
1. Kas yra tyrimo objektas?
2. Ar rezultatus galime palyginti su kitais tyrimais?
3. Kokie duomenys naudojami panašiuose tyrimuose? Kaip jie saugomi?

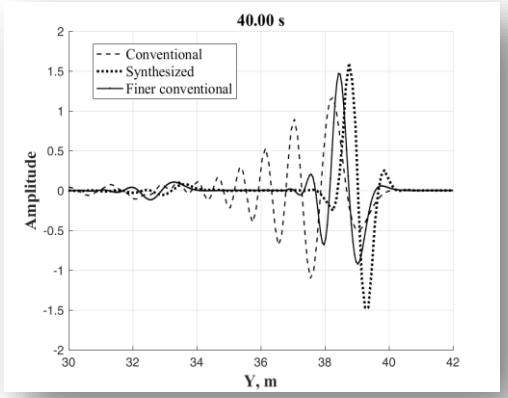
Synthesis of 2D and 3D High Order Finite Elements for Short Acoustic Wave Simulation

ktu 1922

Modelio sudarymas → Modeliavimas → Modeliavimo rezultatai ir jų analizė







1. Kas yra tyrimo objektas?
2. Ar rezultatus galime palyginti su kitais tyrimais?
3. Kokie duomenys naudojami panašiuose tyrimuose? Kaip jie saugomi?

Greedy algorithm based on the genetic optimization to construct a schedule for flight service staff

Uždavinio formalizavimas → Optimizavimo procesas → Optimizavimo rezultatų analizė

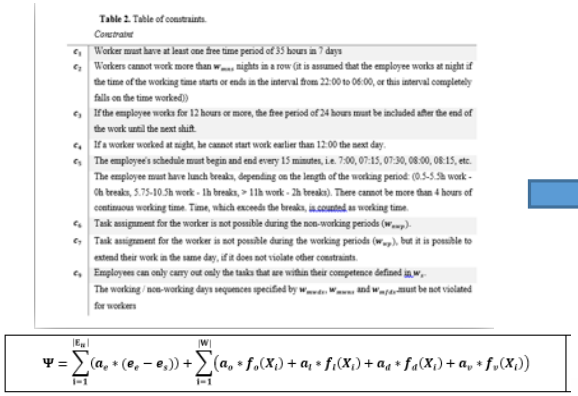
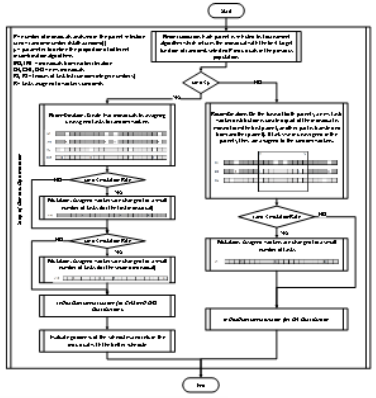
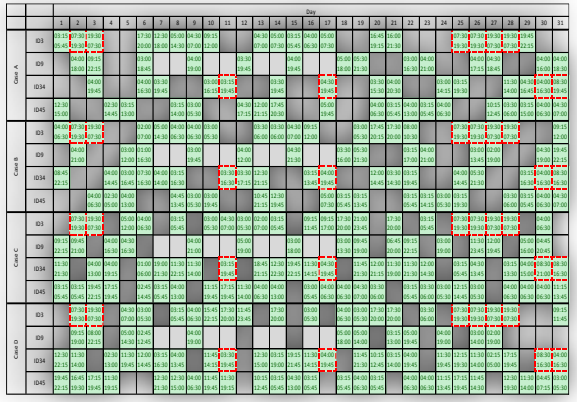


Table 2. Table of constraints.

	Constraint
c_1	Worker must have at least one free time period of 35 hours in 7 days
c_2	Workers cannot work more than w_{max} nights in a row (it is assumed that the employee works at night if the time of the working time starts or ends in the interval from 22:00 to 06:00, or this interval completely falls on the time worked))
c_3	If the employee works for 12 hours or more, the free period of 24 hours must be included after the end of the work until the next shift.
c_4	If a worker worked at night, he cannot start work earlier than 12:00 the next day.
c_5	The employee's schedule must begin and end every 15 minutes, i.e. 7:00, 07:15, 07:30, 08:00, 08:15, etc. The employee must have lunch breaks, depending on the length of the working period: (0.5-5.5h work - 0h breaks, 5.75-10.5h work - 1h breaks, > 11h work - 2h breaks). There cannot be more than 4 hours of continuous working time. Time, which exceeds the breaks, is counted as working time.
c_6	Task assignment for the worker is not possible during the non-working periods (w_{nwp}).
c_7	Task assignment for the worker is not possible during the working periods (w_{wp}), but it is possible to extend their work in the same day, if it does not violate other constraints.
c_8	Employees can only carry out only the tasks that are within their competence defined in w_e.
	The working / non-working days sequences specified by w_{mode}, w_{mmax} and w_{mfde} must be not violated for workers

$$\Psi = \sum_{i=1}^{|E_R|}(a_e * (e_e - e_s)) + \sum_{i=1}^{|W|}(a_o * f_o(X_i) + a_t * f_t(X_i) + a_{tt} * f_{tt}(X_i) + a_v * f_v(X_i))$$





1. Kas yra tyrimo objektas?
2. Ar rezultatus galime palyginti su kitais tyrimais?
3. Kokie duomenys naudojami panašiuose tyrimuose? Kaip jie saugomi?

Failų formatai

Tik savo **srities specialistai** susipažinę su jų aplinkoje naudojama duomenų specifika ir programine įranga gali nuspręsti, kaip duomenys turi būti saugomi.

1. Failų formatai **- ar reikalingos spec. programinė įranga, norint peržiūrėti failo turinį?**

2. Duomenų kompresija **- kompresija atliekama su duomenų praradimu / be duomenų praradimo?**

3. Duomenų transformavimas **- ar prireikus bus galima duomenis transformuoti iš vieno formato į kitą?**

Duomenų valdymo planas

Duomenų valdymo planas

H2020 Programme „**Guidelines onFAIR Data Managementin Horizon 2020**“

http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf#page=10

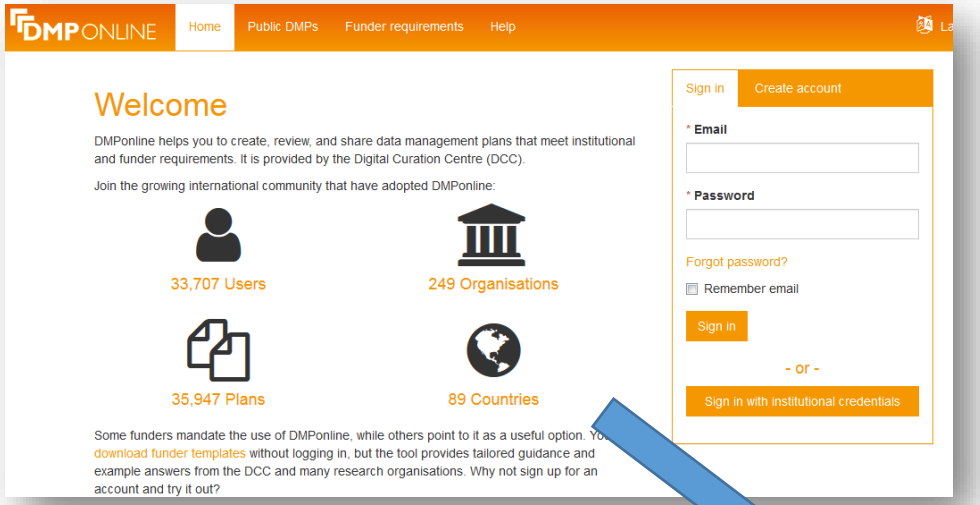
SUMMARY TABLE 1

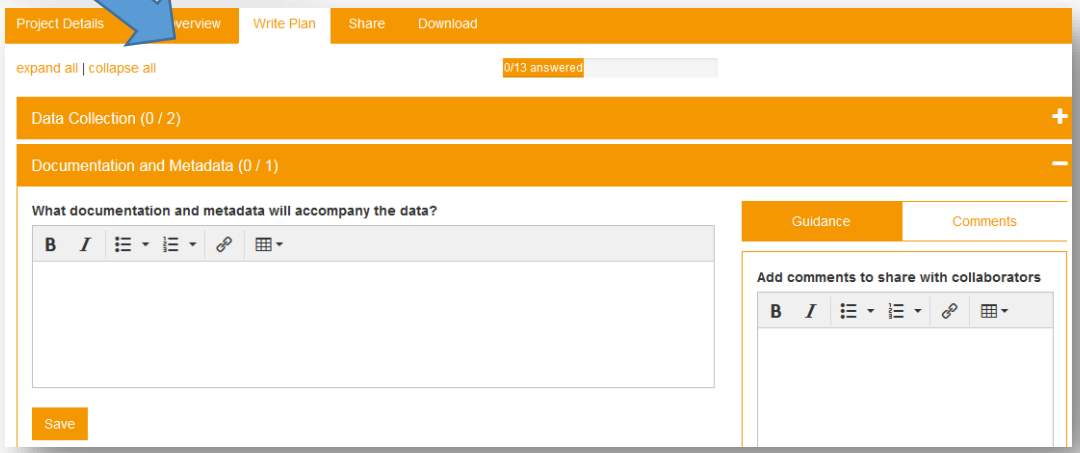
FAIR Data Management at a glance: issues to cover in your Horizon 2020 DMP

This table provides a summary of the Data Management Plan (DMP) issues to be addressed, as outlined in Annex I. You should refer to the annex and the main text of the guidelines for further guidance.

DMP component	Issues to be addressed
1. Data summary	• State the purpose of the data collection/generation • Explain the relation to the objectives of the project • Specify the types and formats of data generated/collected • Specify if existing data is being re-used (if any) • Specify the origin of the data • State the expected size of the data (if known) • Outline the data utility: to whom will it be useful

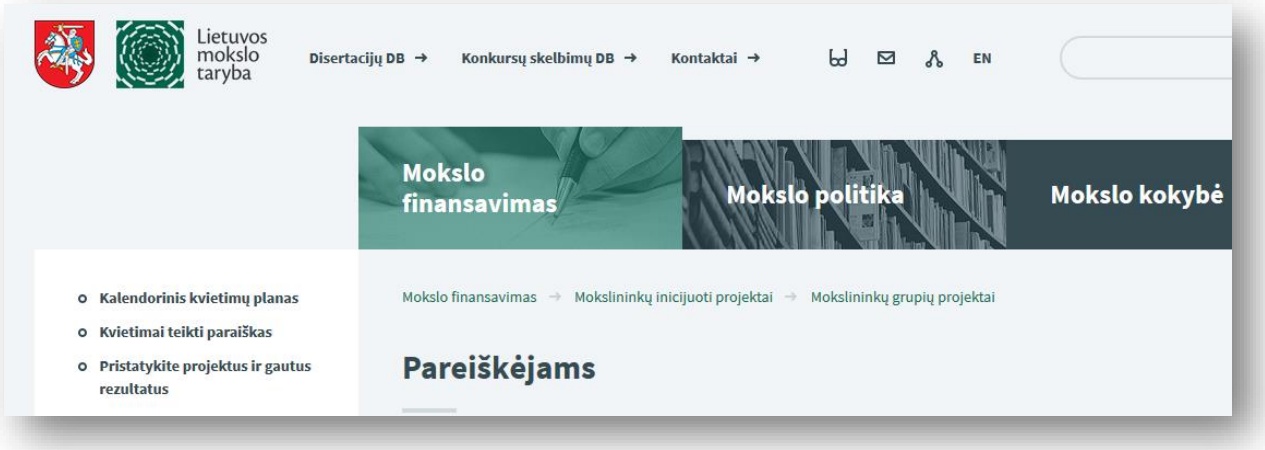
Duomenų valdymo planas





https://dmponline.dcc.ac.uk/

Duomenų valdymo planas



https://www.lmt.lt/lt/mokslininku-inicijuoti-projektai/mokslininku-grupiu-projektai/pareiskejams/2532

Rekomendacijos dėl mokslo ir sklaidos projektų duomenų valdymo plano

Rekomendacijos dėl mokslo ir sklaidos projektų duomenų valdymo plano

A. Duomenų rinkimas

1. Kokius duomenis planuojate rinkti ar sukurti?

1.1. Ar egzistuoja duomenys, kuriuos galėtumėte pakartotinai panaudoti?

1.2. Kokie numatomi duomenų tipai, formatas ir apimtys?

Rekomendacijos dėl mokslo ir sklaidos projektų duomenų valdymo plano

B. Duomenų ir jų atsarginių kopijų kaupimas

2. Kaip duomenys ir jų atsarginės kopijos bus kaupiami projekto metu?

2.1. Kur bus kaupiami duomenys?

2.2. Kokiu būdu planuojate atkurti duomenis, jei jie būtų pažeisti? Ar kursite duomenų atsargines kopijas

3. Kaip užtikrinsite sukauptų duomenų saugumą?

3.1. Kokia gali būti rizika duomenų saugumui ir kaip ši rizika bus valdoma?

3.2. Kaip užtikrinsite, kad projekto partneris (jei yra) galėtų saugiai pasiekti duomenis?

Rekomendacijos dėl mokslo ir sklaidos projektų duomenų valdymo plano

C. Duomenų atranka ir saugojimas

4. Kurie duomenys yra ilgalaikės vertės ir turi būti saugomi?

    4.1. Kurie duomenys turi būti saugomi ar sunaikinami dėl tam tikrų sutartinių nuostatų, teisinių ar kitų reikalavimų?

    4.2. Kokį laikotarpį duomenys bus saugomi?

Rekomendacijos dėl mokslo ir sklaidos projektų duomenų valdymo plano

D. Prieiga prie duomenų

5. Kaip užtikrinsite duomenų prieinamumą ir galimybę jais naudotis?

5.1. Kada duomenys taps prieinami?

5.2. Kaip potencialūs naudotojai sužinos apie duomenis?

5.3. Kam bus suteikta galimybė naudotis duomenimis ir kokiomis sąlygomis?

Rekomendacijos dėl mokslo ir sklaidos projektų duomenų valdymo plano

E. Atsakomybė ir ištekliai

6. Kam bus priskirta atsakomybė už duomenų tvarkymą bei valdymą?

6.1. Kas atsakingas už DVP įgyvendinimą bei periodišką peržiūrą ir koregavimą?

6.2. Ar nuostatos dėl duomenų nuosavybės ir atsakomybės už mokslinių tyrimų duomenų tvarkymą / valdymą bus aptartos su partneriu (jei yra)?

Rekomendacijos dėl mokslo ir sklaidos projektų duomenų valdymo plano

E. Atsakomybė ir ištekliai

7. Kokie žmogiškieji ir kiti ištekliai bus reikalingi rengiant bei įgyvendinant DVP?

7.1. Ar reikės įdarbinti specialias kvalifikacijas turintį darbuotoją?

7.2. Ar bus reikalinga speciali, papildoma įranga, įskaitant programinę?

7.3 Ar įvertinote, kad duomenų bankai ar saugyklos gali taikyti mokesčius už duomenų kaupimą, saugojimą ar atvėrimą?

Duomenų valdymo planas, DVP

1. Individualaus duomenų valdymo plano rašymas

~15 min

2. Duomenų valdymo plano aptarimas grupelėse

~15 min

3. Bendras rezultatų / iškilusių klausimų aptarimas

~15 min