Datan organisoinnin ABC -työpajan ohjeistus

(in English below)

Datan organisoinnin ABC -työpajan ohjeistus 1
Tehtävien ohjeistus 2
Tehtävä 1: Projektin kansiorakenteen suunnittelu sekä tiedostojen ja kansioiden nimeäminen 2
Tehtävä 2: Taulukkomuotoisen datan organisointi ja sudenkuopat 4
ABC for Organizing data - workshop guidance 7
Exercise guidance 8
Task 1: Designing the folder structure of the project and naming files and folders 8
Task 2: Organization of tabular data and pitfalls 10

Datan organisoinnin ABC -työpajan tarkoituksena on kiinnittää huomiota järkeviin tapoihin käsitellä ja dokumentoida dataa. Työpajassa käydään läpi muun muassa tiedostojen ja kansioiden nimeämiskäytänteitä, versionhallintaa sekä muita dokumentaatioon liittyviä käsitteitä. Työpajan materiaalit on julkaistu CC BY 4.0 lisenssillä.

Työpajan esimerkit ja tehtävät ovat luonnontiedepainotteisia, mutta asia on yleisluontoista ja sitä voi hyvin käyttää alasta riippumatta. Muokkaamalla tehtävien aiheita, voit helposti räätälöidä materiaalia alakohtaiseksi.

Työpaja jakautuu kahteen osioon, joista (1) ensimmäinen keskittyy (alustus + tehtävä) datan organisoinnin perusteisiin, kuten kansiorakenne, tiedostojen nimeäminen, versionhallinta ja readme-tiedostot. Toinen osa (2) ja tehtävä pureutuu taulukkomuotoisen datan järjestämiseen ja hyvien käytänteiden läpikäymiseen. Alla on tarkemmat ohjeet molempien osioiden tehtävien tekemiseen ja fasilitointiin.

Työpajan voi pitää yhdessä osassa, jolloin aikaa kannattaa varata 2,5 tuntia (aikatauluesimerkki A alla). Työpaja on myös mahdollista jakaa kahteen tilaisuuteen (aikatauluesimerkki B alla), jossa ensimmäinen tilaisuus keskittyy datan organisoinnin perusteisiin ja käy osan 1 alustuksesta sekä tehtävän numero 1 läpi, ja toinen tilaisuus pureutuu taulukkomuotoisen datan organisointiin ja tehtävään numero 2.

Aikatauluesimerkki A työpajan kulusta:

- 30min - 45min Alustus osa 1: Intro, kansiorakenne, tiedostonimet, versiointi, read me
- 30min - 60min Tehtävä 1: kansiorakenne, tiedostojen nimet
- 30min - 45min Alustus osa 2: taulukkomuotoisen datan järjestäminen
- 30min - 60min Tehtävä 2: taulukkomuotoinen data
- 20min - 30min Tehtävien purku & loppukeskustelu

Aikatauluesimerkki B työpajan kulusta:

Tilaisuus 1

- 30min - 45min Alustus osa 1: Intro, kansiorakenne, tiedostonimet, versiointi, read me
- 30min - 60min Tehtävä 1: kansiorakenne, tiedostojen nimet
- 10-30min Tehtävän purku & loppukeskustelu

Tilaisuus 2

- 30min - 45min Alustus osa 2: taulukkomuotoisen datan järjestäminen
- 30min - 60min Tehtävä 2: taulukkomuotoinen data
- 10min - 30min Tehtävän purku & loppukeskustelu

Tehtävien ohjeistus

Tehtävä 1: Projektin kansiorakenteen suunnittelu sekä tiedostojen ja kansioiden nimeäminen

- **Kesto:** 30 min fasilitaattorin kanssa, 45-60 min ilman fasilitaattoria.
- **Yhteistyöalusta:** Tehtävän tekoon voi käyttää yhteistyöalustaa, jolla voi liikuttaa laatikoita esimerkiksi: Conceptboard, Miro, Flinga, Mural, mutta tehtävän voi tehdä myös post-it -lappuja käyttäen tai esimerkiksi Power pointilla! Kouluttajan täytyy tehdä pohja (tarvittavat laatikot, mahdollisesti kansiorakenneaihio ym.) käytössä olevaan yhteistyökäyttöalustaan.
    - Mallimme on erillisillä dioilla ja kuvakaapattuna Conceptboardilta (alla)
    - Kouluttajan tulee tehdä yhtä monta tehtäväpohjaa, kun ryhmiä on, jotta kaikki ryhmät voivat työstää omaa pohjaa.
- **Ryhmät:** Tehtävää varten jakaudutaan 3-6 hengen pienryhmiin.
    - Ryhmissä voi olla fasilitaattori tai ryhmät voivat toimia itsenäisesti. Fasilitaattori voi nopeuttaa ryhmän toimintaa toimimalla kirjurina ja ohjata tehtävän tekoa esimerkiksi: tehdään ensin kansiorakenne, toiseksi siirretään tiedostot oikeisiin kansioihin ja kolmanneksi nimetään kansiot ja tiedostot uudelleen.

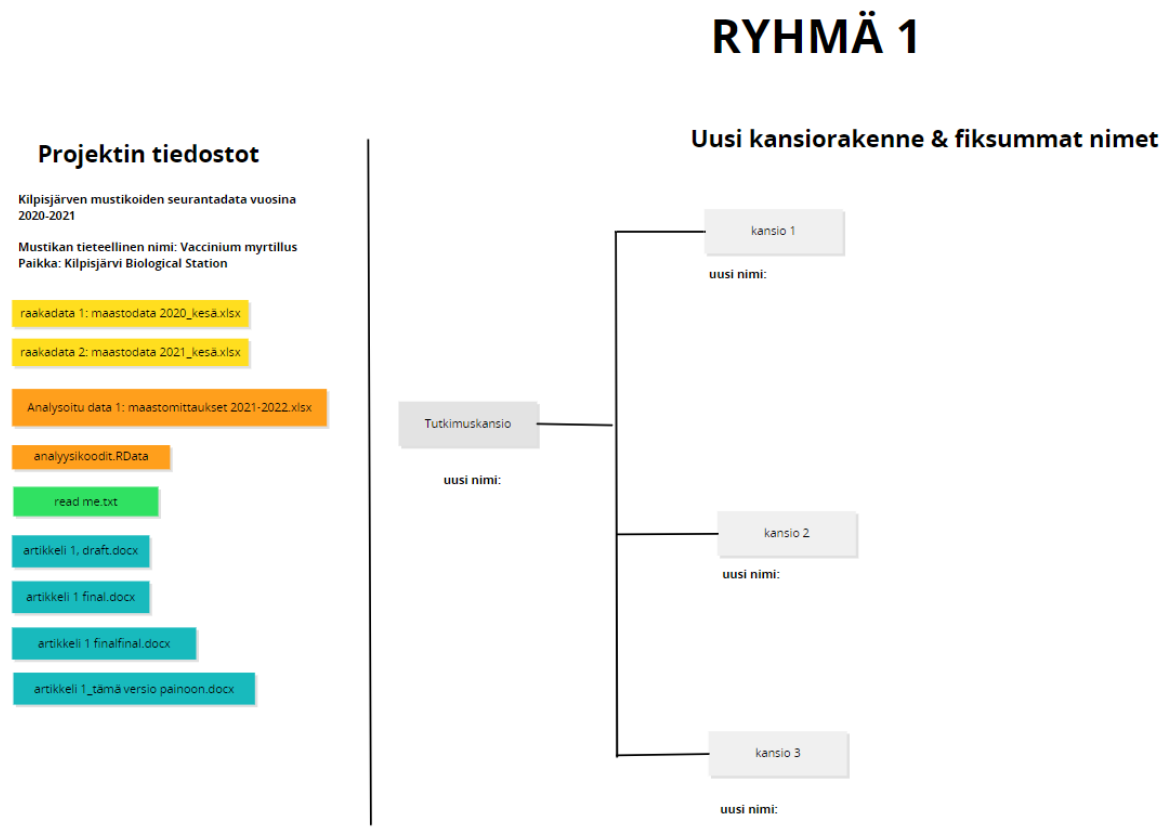
    - Jos ryhmässä ei ole fasilitaattoria, ryhmän kannattaa sopia työnjaosta ja ajan seuraamisesta.
- **Tehtävä:** Ryhmien tulee rakentaa tutkimusprojektille selkeämpi kansiorakenne sekä nimetä kansiot ja tiedostot. Tehtävässä ei ole yhtä oikeaa ratkaisua vaan tarkoitus on miettiä hyviä käytänteitä.
    - Tarkoitus on siirtää/raahata vasemmalle listatut (kuva 1) huonosti nimetyt tiedostot (laatikot) oikealle, johon ryhmäläiset voivat tehdä haluamansa kansiorakenteen. Lisäksi uudet kansiot ja projektin tiedostot tulee nimetä ymmärrettävästi ja yhdenmukaisesti hyvin nimeämiskäytänteiden mukaisesti (kuva 2).
    - Työjärjestys voi olla seuraavanlainen
        - 1) Suunnitelkaa uusi kansiorakenne
        - 2) Siirtäkää tiedostolaatikot kansioiden alle suunnittelemallanne tavalla
        - 3) Nimetkää kansiot & tiedostot yhdenmukaisesti
- Lopuksi käydään eri ryhmien tekemät rakenteet ja nimet läpi yhdessä.

Tehtävän projektin taustatiedot:

- Kilpisjärven mustikoiden seurantadataa vuosina 2020-2021, useampi tiedosto
- Mustikan tieteellinen nimi: *Vaccinium myrtillus*
- Tutkimuksen tekopaikka: Kilpisjärvi Biological Station

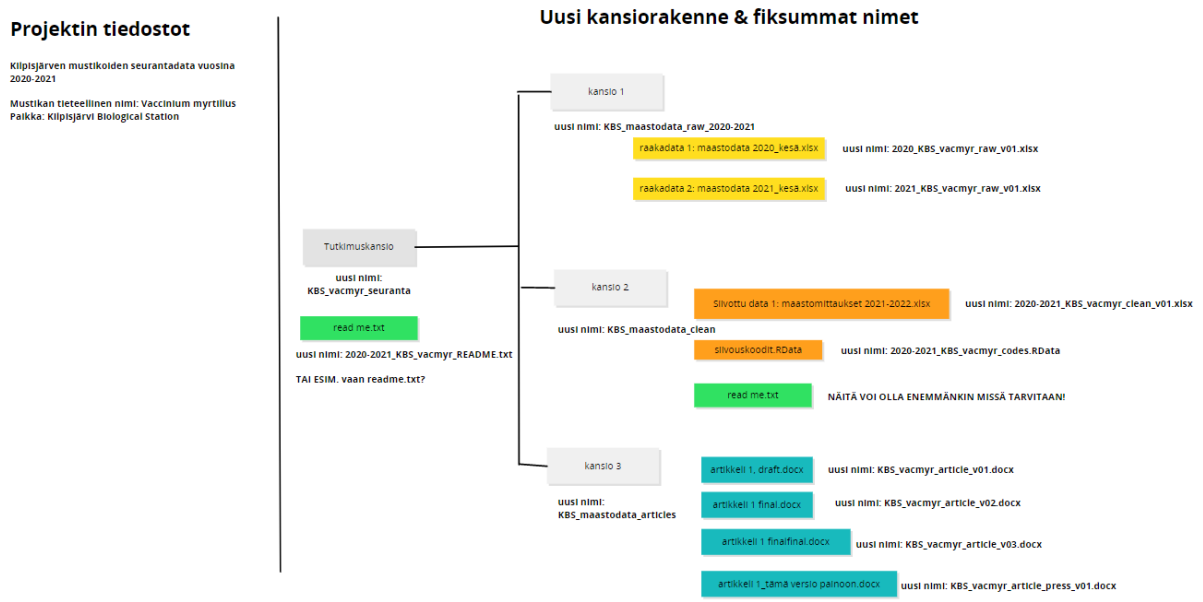
Tehtävän mallipohja erillisillä dioilla.

Esimerkki Conceptboardilla



Kuva 1: Esimerkkipohja Conceptboardilla. Vasemmalla on projektin tiedot sekä epäselvästi nimetyt tiedostot eri värisinä laatikoina. Laatikoissa voi käyttää värejä, joka nopeuttaa

tiedostojen hahmottamista tai vain yhtä väriä. Vasemmalla on alustava kansiorakenne (pääkansio ja kolme alikansiota), mutta ryhmäläiset voivat muokata pohjaa haluamallaan tavalla. Tehtävässä on tarkoitus 1) suunnitella kansiorakenne, 2) raahata tiedostot (=laatikot) vasemmalta oikealle haluamaansa kansioon. Tämän jälkeen 3) kansiot ja tiedostot tulee nimetä järkevästi.



Kuva 2: Esimerkkivastaus, jossa tiedostot (laatikot) on siirretty oikealle kansioihin ja kansiot sekä tiedostot on nimetty uudelleen. Kansiorakenteeseen ja tiedostojen nimeämiseen ei ole yhtä oikeaa vastausta.

Tehtävä 2: Taulukkomuotoisen datan organisointi ja sudenkuopat

- **Kesto:** 30 min fasilitaattorin kanssa, 45-60 min ilman fasilitaattoria.
- **Yhteistyöalusta:** Tehtävän tekoon voi käyttää mitä tahansa taulukkolaskentaohjelmaa, esim. google sheet tai excel. Kouluttajan täytyy tehdä taulukko etukäteen ja jakaa linkki tai tiedosto osanottajille.
    - Mallimme on erillisessä taulukossa.
    - Kouluttajan tulee tehdä yhtä monta tehtäväpohjaa (esim. sheet-välilehtiä), kuin ryhmiä on, jotta kaikki ryhmät voivat työstää tehtävää omalle pohjalle.
- **Ryhmien koko:** Tehtävää varten jakaudutaan samoihin 3-6 hengen pienryhmiin, joita käytettiin tehtävässä 1.
    - Ryhmissä voi olla fasilitaattori tai ryhmät voivat toimia itsenäisesti. Fasilitaattori voi nopeuttaa ryhmän toimintaa toimimalla kirjurina ja ohjata tehtävän tekoa. Jos ryhmässä ei ole fasilitaattoria tai siitä huolimatta, ryhmän kannattaa sopia työnjaosta (kuka kirjaa) ja ajan seuraamisesta.
- **Tähtävä:** Tehtävässä ryhmien tulee tarkastella taulukkomuotoista dataa ja pohtia
    - 1) mitä virheitä, puutteita tai huonoja käytänteitä löydät ohessa olevasta taulukosta (datasta),
    - 2) miksi käytänne on huono, (kerro miksi kyseinen toteutus/käytänne on huono)

    - 3) miten korjaisit/parantaisit taulukkoa ja,
    - 4) miten muokkaisit taulukkoa, esimerkiksi miten muokkaisit päivämäärät vastaamaan ISO 8601 standardia. Käyttäsitkö koodia vai jotain taulukkolaskennan keinoa. Tähän osaan ei ole yhtä oikeaa vastausta, vaan kysymyksen tarkoitus on herättää keskustelua tutkijoiden keskuudessa ja vaihtaa käytänteitä.
    - Tätä tehtävää voi käyttää soveltuvin osin riippuen kouluttajan taustasta. Jos jokin osa tuntuu kouluttajasta vaikealle, sen voi jättää pois tehtävästä.
- Lopuksi käydään eri ryhmien löytämät puutteet ja virheet läpi ja keskustellaan paremmista vaihtoehdoista.

Tehtävän projektidatan taustatiedot:

- Sarvaksen koealan tunturikoivun siementuotannon seuranta Kilpisjärvellä
- Koealalla on 10 siemensuppiloa tunturikoivuissa (II A - R).
- Suppilot tyhjennetään 4 kertaa kesäkuukausien aikana ja pussiin kertyneiden siementen määrä lasketaan.
- Joku on muuttanut **osan** nollista 1 luvuiksi (0 = 1).

Mallipohja erillisessä taulukossa.

Mallivastauksia

	VIRHE / PUUTE	MIKSI?	PARANNUSEHDOTUS	MITEN TEKISIT (KOODI KÄSIN)
1	Useampi taulukko yhdessä taulukossa	Kone ei ymmärrä useita taulukoita. Jokaisen rivin tulisi olla havainto, jotta kone datan lukee.	Taulukot yhdistettävä - (yhdistetään siten, että päivämäärät juoksevat riveillä).	Käsin vai koodeilla?
2	Taulukossa on otsikkorivi	Otsikkorivi ei ole dataa	Otsikko rivi poistetaan. Otsikkorivin teksti voisi olla tiedoston nimi	
3	Taulukossa on negatiivisia arvoja -1	0 arvoista käytetään aina 0 arvoja. kone lukee tämän joko -1 arvona tai jopa tyhjänä arvona ohjelmistosta riippuen.	nollat korjattava takaisin raakadatan	

4	Kaikkia nolla-arvoja ei ole merkitty	Kone lukee tyhjät solut puuttuviksi arvoiksi. Tällä saattaa olla suuri vaikutus analyysituloksiin.	Täytä tyhjät ruudut nolla-arvoilla. Ota kuitenkin huomioon, että jotkin puuttuvat arvot ovat oikeasti puuttuvia arvoja.	
5	Taulukossa on kommentteja puuttuvista tiedoista ja poikkeuksista	Datassa voi olla kommentteja mutta niiden tulee olla omassa sarakkeessa.	Tee oma sarake kommenteille.	
6	Tyhjät rivit taulukoiden välissä	Kun yhdistät taulukot älä jätä väliin tyhjiä rivejä, koska...	Kun yhdistät taulukot älä jätä tyhjiä rivejä erottamaan esimerkiksi eri vuosia.	
7	Sarakkeet ja rivit osin väärinpäin	Data ei lukeudu oikein. Päivämäärä on dataa ei otsikko	Päivämäärät sarakkeisiin	
8	Muuttujia ei ole nimetty	Päivämäärä ei kerro mitään datasta	Anna muuttujille kuvaavat nimet. Kuten BirdSeedAbu tms. Yksikkö eli määrä kannattaa liittää jo muuttujan nimeen.	
9	Roomalaiset numerot koodauksessa	Koealojen nimet eivät välttämättä lukeudu oikein	Parempi nimeämiskäytäntö koealoille	
10	Päivämäärien muoto	/ merkit lukeutuvat hieman eri tavalla ja useasti suhteellisen helposti koodattavissa auki. Datan lukua ja käsittelyä helpottaa kun päivämäärät noudattavat ISO 8601 standardia ja ne vuodet, kuukaudet ja päivät on erotettu omiin sarakkeisiin	päivämäärät standardimuotoon sekä eroteltu omiin sarakkeisiin	

11	Paikannimet	Koordinaattien puuttuminen ei vaikuta datan luettavuuteen, mutta moneen muuhun asiaan kyllä.	koordinaatit paikannimiin	

ABC for Organizing data - workshop guidance

The purpose of the data organisation ABC workshop is to raise awareness of good data organisation and documentation practices. The workshop will cover file and folder naming practices, version control, and other documentation-related concepts. The materials for the workshop have been released under the CC BY 4.0 licence.

The examples and tasks in the workshop are based on the life sciences, but as the topic is generic, the materials can be used regardless of scientific discipline. By editing the topics of your tasks, you can easily adapt the material to be discipline specific.

The workshop is divided into two parts, the first (1) of which (introduction + assignment) focuses on the basics of data organisation, such as folder structure, file naming, version control and readme files. The second part (2) and the assignment focus on organising tabular data and going through good practices. Below are more detailed instructions on how to carry out and facilitate the tasks for both sections.

The workshop can be held in one session, and we recommend 2 hours and 30 minutes for this (timetable example A below). It is also possible to split the workshop into two sessions (schedule example B below), where the first session focuses on the basics of data organisation, going through part 1 of the introduction and task number 1, and the second session focuses on organising tabular data (part 2) and task number 2.

Schedule example A for the workshop:

- 30min - 45min Introduction part 1: Intro, folder structure, file names, versioning, read me - files
- 30min - 60min Task 1: folder structure, file names
- 30min - 45min Introduction part 2: arranging tabular data
- 30min - 60min Task 2: tabular data
- 20min - 30min Going through the tasks & final discussion

Schedule example B for the workshop:

Event 1

- 30min - 45min Introduction part 1: Intro, folder structure, file names, versioning, read me
- 30min - 60min Task 1: folder structure, file names
- 10-30min Going through the tasks & final discussion

Event 2

- 30min - 45min Introduction part 2: arranging tabular data
- 30min - 60min Task 2: tabular data
- 10min - 30min Going through the tasks & final discussion

Exercise guidance

Task 1: Designing the folder structure of the project and naming files and folders

- **Duration**: 30 min with a facilitator, 45-60 min without a facilitator.
- **Collaboration platform**: To do the exercise, you can use a collaboration platform where you can move boxes, for example: Conceptboard, Miro, Flinga, Mural, but the exercise can also be done with post-it notes or with Powerpoint, for example! The trainer must create the basis (necessary boxes, possibly a folder structure blank, etc.) for the collaboration platform in used.
    - Our model is on separate slides and screenshots from Conceptboard are below.
    - The trainer should make the same number of exercise bases as there are groups, so that each group can work on its base.

- **Groups**: For the exercise, we will divide into small groups of 3-6 people.
    - The groups can have a facilitator, or the groups can work independently. The facilitator can speed up the group's work by acting as a secretary and guiding the group through the exercise, for example: first create a folder structure, second move the files into the correct folders, third rename the folders and files.
    - If the group does not have a facilitator, the group should agree on the division of work and keep track of time.
- **Task**: The groups must create a clearer folder structure for the research project and name the folders and files. There is no single correct solution to this task, but the aim is to think about good practice.
    - The aim is to move/drag the poorly named files (boxes) listed on the left (Figure 1) to the right where group members can create the folder structure they want. In addition, new folders and project files must be comprehensively and consistently named according to naming conventions (Figure 2).
    - The exercise can be carried out in the following order
        - 1) Plan a new folder structure
        - 2) Move the file boxes under the folders as planned
        - 3) Give consistent names to the folders and files.
- Finally, go through the structures and names created by the different groups together.

Background information of the assignment project:

- Kilpisjärvi blueberry monitoring data in 2020-2021, several files
- Blueberry scientific name: *Vaccinium myrtillus*
- Place of research: Kilpisjärvi Biological Station

Example on separate slides.

Example from Conceptboard

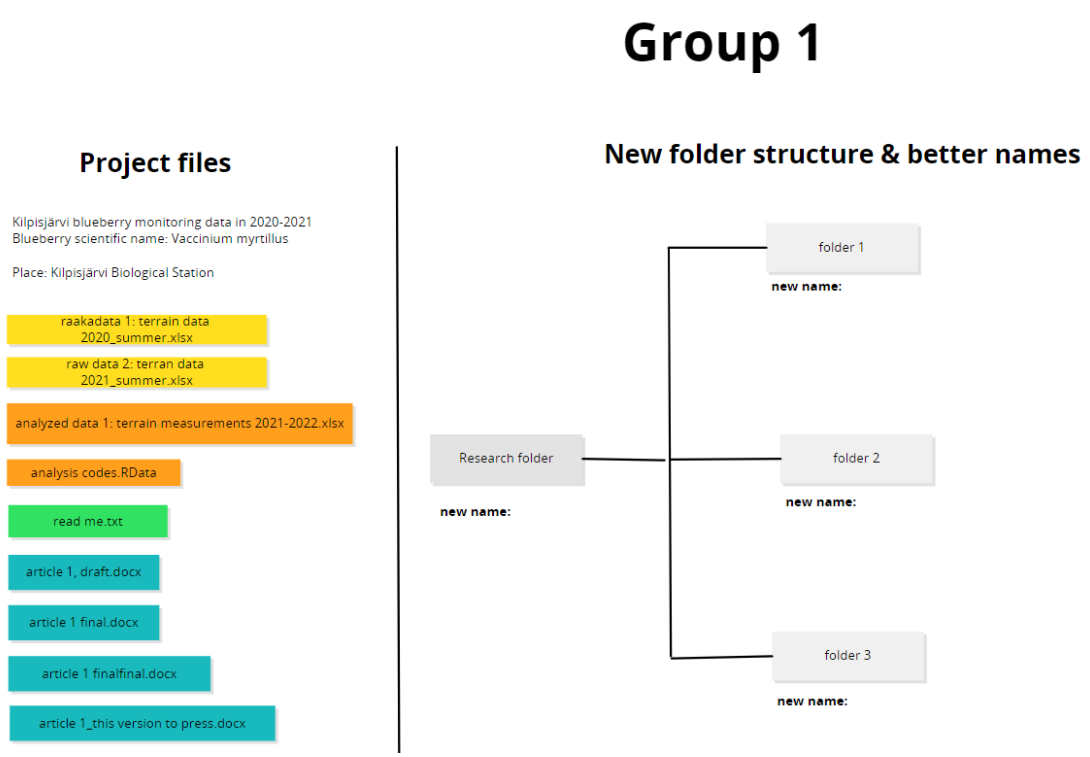


Figure 1: An example template with Conceptboard. On the left are the project information and ambiguously named files in different coloured boxes. You can use colours in the boxes, which speeds up the presentation of the files, or just one colour. On the left is the initial folder structure (main folder and three subfolders), but group members can modify the base as they wish. The task is to 1) design a folder structure, 2) drag the files (=boxes) from left to right into the desired folder. Then 3) give the folders and files meaningful names.

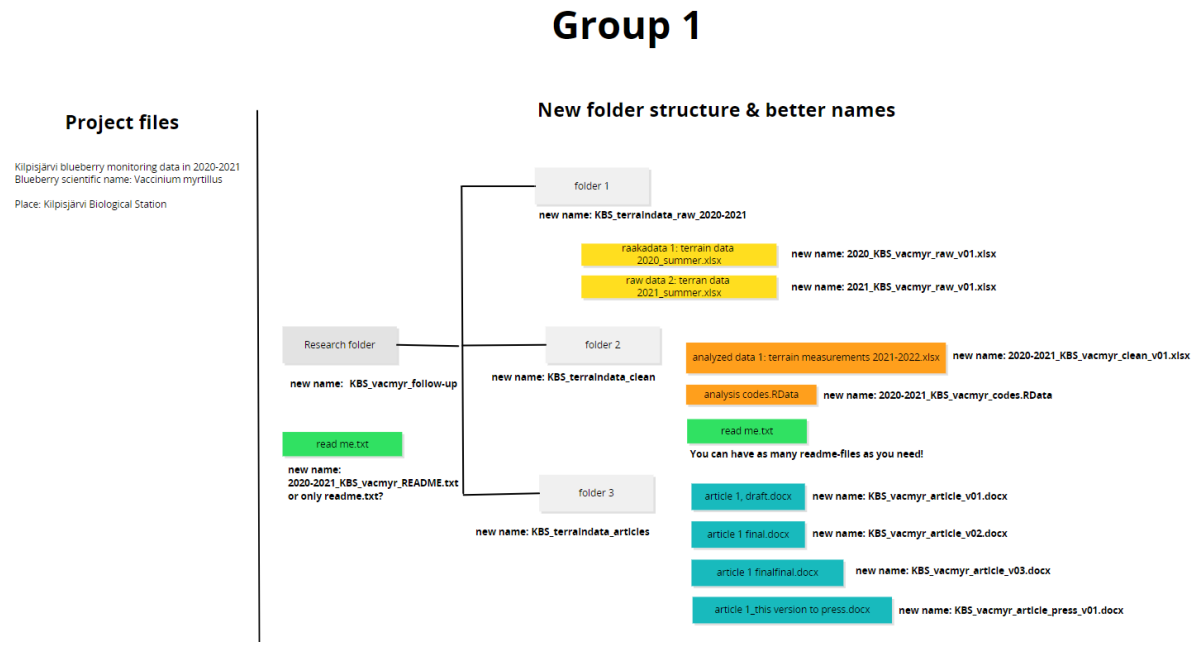


Figure 2: An example answer where the files (boxes) have been moved to the right in the folders and the folders and files have been renamed. There is no single correct answer to folder structure and file naming.

Task 2: Organization of tabular data and pitfalls

- **Duration**: 30 min with a facilitator, 45-60 min without a facilitator.
- **Collaboration platform**: You can use any spreadsheet program to do this exercise, e.g. google sheet or excel. The trainer must prepare the spreadsheet in advance and distribute the link or file to the participants.
    - Our model is on a separate sheet.
    - The trainer should make the same number of exercise bases (e.g., sheet tabs) if there are groups, so that each group can work on its own base.
- **Size of the groups:** For this task, divide into the same small groups of 3-6 people as for Task 1.
    - The groups can have a facilitator, or the groups can work independently. The facilitator can speed up the work of the group by acting as a secretary and guiding the work.

    - If there is no facilitator in the group, or even if there is one, the group should agree on the division of work (who will take notes) and time recording.
- **Objective**: In the task, the groups must look at the tabular data and think about
    - 1) what errors or bad practices do you find in the table (data),
    - 2) why is the practice bad (tell why the implementation/practice in question is bad)
    - 3) how would you correct/improve the table and,
    - 4) how would you edit the table, e.g., how would you edit the dates to comply with the ISO 8601 standard. Would you use code or some form of spreadsheet calculation? There is no single correct answer to this part, but the purpose of the question is to stimulate discussion among researchers and to share practices.
    - This task can be used as appropriate depending on the trainer's background. If any part seems difficult to the trainer, it can be left out of the task.
- Finally, go through the exercise with the different groups and discuss the better alternatives found for the tabular data.

Background information of the project data for the task:

- Sarvas research plot for birch seed production in Kilpisjärvi
- In the research area there are 10 funnels (II A - R) located on different trees in the Sarvas area to gather bich seeds.
- Funnels are emptied 4 times during summer months and the number of seeds is counted.
- Someone has changed **some** of the zeros into 1 (0=1)

Example on separate sheet

Example answers

	ERROR / DEFICIENCY	**Why?**	**SUGGESTION FOR IMPROVEMENT**	HOW WOULD YOU DO IT (CODE MANUALLY)
1	Multiple tables in one table	False associations to computer. Here each row is not an observation.	Need to combine the data in one table	Manually or with a code?
2	There is a unnecessary heading	Heading is not data	Delete the heading. This could be the name of the file or included in the name.	

3	Using -1 as a zero value	zeros as zeros. many programs might interpret -1 as a null value or an actual -1	Need to change this back! Remember never to modify your raw data!	
4	Not filling in the zeros	Computer probably interprets this as a null value hence your analysis are not correct	Fill in the zeros! And remember to do this!	
5	There are comments in the table	Comments can be in data, but as an individual column.	Should make an extra column for comments. This way you can separate them from the numeric values.	
6	Empty rows between tables.	Also, empty rows confuse the computer...	While combining the data, do not leave empty spaces for example to separate different years.	
7	Rows and columns are partly in wrong order	Date is data, not a heading	Bring the dates to columns	
8	Real variables have no names	Date is not a heading and tells nothing about the data	Give variables descriptive names. Such as BirdSeedAbu. Include the unit in the name, here abundance (Abu).	
9	Roman numbers in names.	There might be a problem when reading the research plot names as they are now.	Give better names, replace roman numbers and remove spaces.	

10	Date format	/ might not read properly into different programs. It is quite easy to code in for example in R, but when dates follow ISO 8601 standard and year, month and day are in different columns	Change dates in standard format and separate year, month, and day in separate columns.	
11	No coordinates	Missing coordinates have no affect how data reads, but should always be included	Add coordinates	