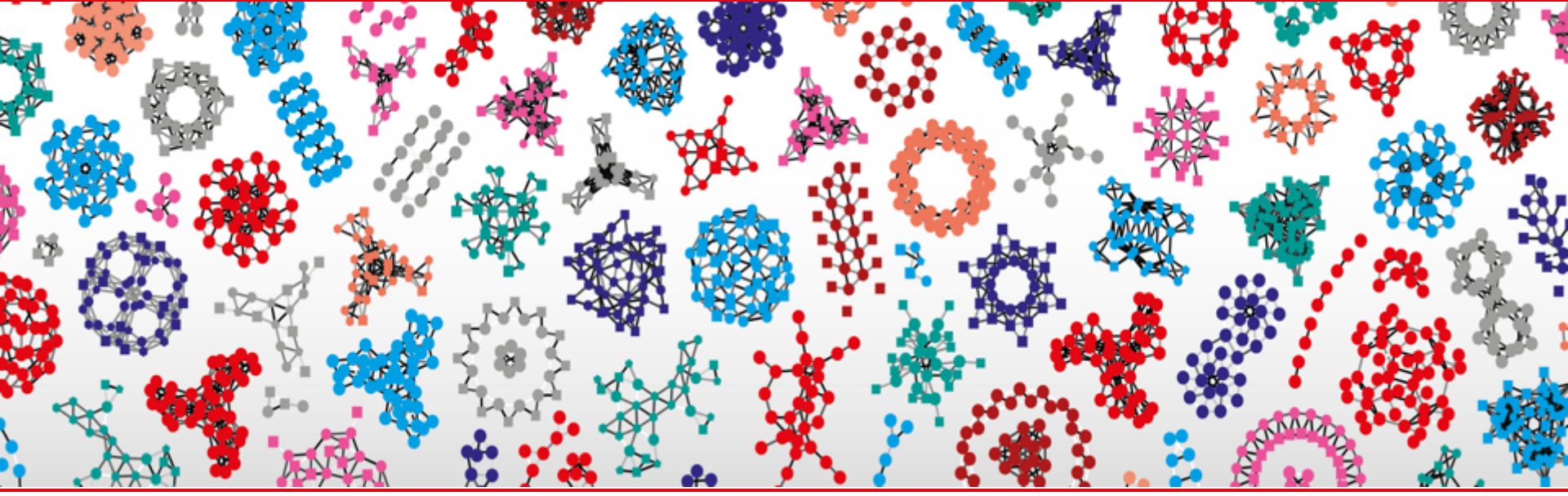


#09 Guidelines to ease DMP template creation

Swiss Open Research Data Hackathon 2021

Anne-Laure Kaufman – Alexandre Cotting – Gregoire Rossier - Clemens Trautwein – Karine Villettaz- Daniela Subotic

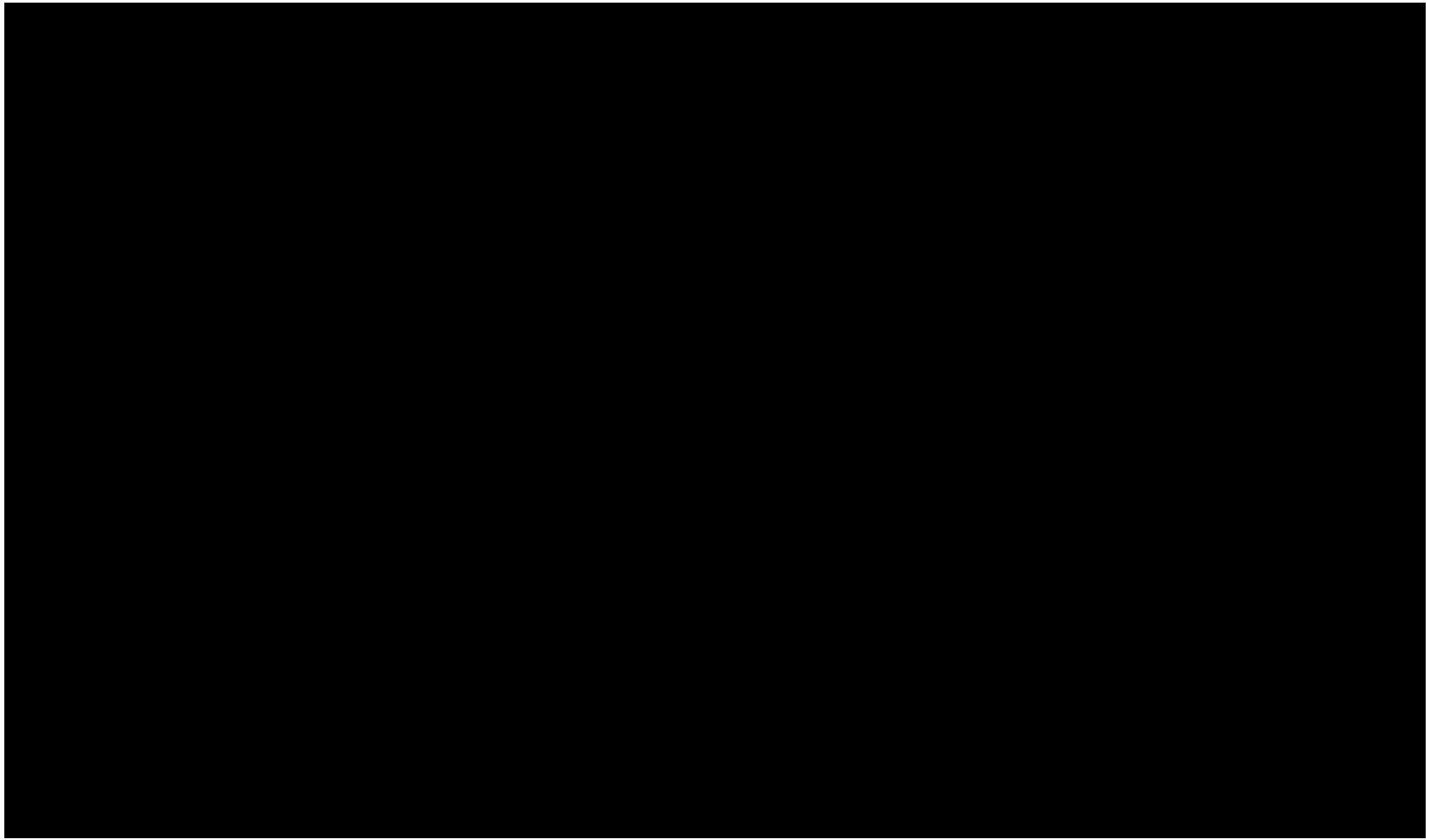
Problem description

DMP Generators are a great help nowadays for researchers and DMP in general could help to promote Open Science culture. However to be fully useful, they would need to be adapted to each domain and institution.

For this hackathon, we have used the *SIB DMP Canvas Generator* * and start brainstorming about providing a process with guidelines to ease DMP template creation for specific domains and institutions.

* https://dmp.vital-it.ch

Questionnaire



DMP Template generated by Canvas 1.0

ORCID (generic)

1 Data collection and documentation

1.1 What data will you collect, observe, generate or reuse?

[Questions you might want to consider: What type and format of data will you collect/generate?]

Genomic data: Genomic data from genome-wide association studies (GWAS) will be stored in PLINK binary format. Processed exome sequencing data will be stored in VCF (variant call format). Other genomic sequencing data, for example whole genome bisulfite sequencing (WGBS) and ChIP-Seq will be stored as raw sequence (FASTQ) and alignments stored as compressed binary (BAM) files.

Microscopy data: Microscopy data will be collected as raw Zeiss LSM (confocal) or standard TIF (JEOL, TEM). Images can be exported in a variety of data formats (e.g. TIF, JPG).

Transcriptomic data: Affymetrix data will be stored as CEL files which will be described in an associated table with sample name, description and other information relevant to the sample such as RNA purity. RNA-Seq data will be stored in FASTQ format and alignments to genomic sequence in compressed binary (BAM) format.

We estimate that a total data volume ranging between 100 GB and 2 TB will be generated during the course of the project.

We plan on reusing Metabolomic data third-party datasets. [Please mention the owner of the third-party data, the format and the content of the datasets] For the aforementioned, third-party datasets, we have open access usage rights.

We plan on reusing Transcriptomic data datasets produced in our group.

[Questions to consider: What type of data, patient data management software (e.g. soariane/data warehouse, secutrial), format and volume of data will you collect/generate?].

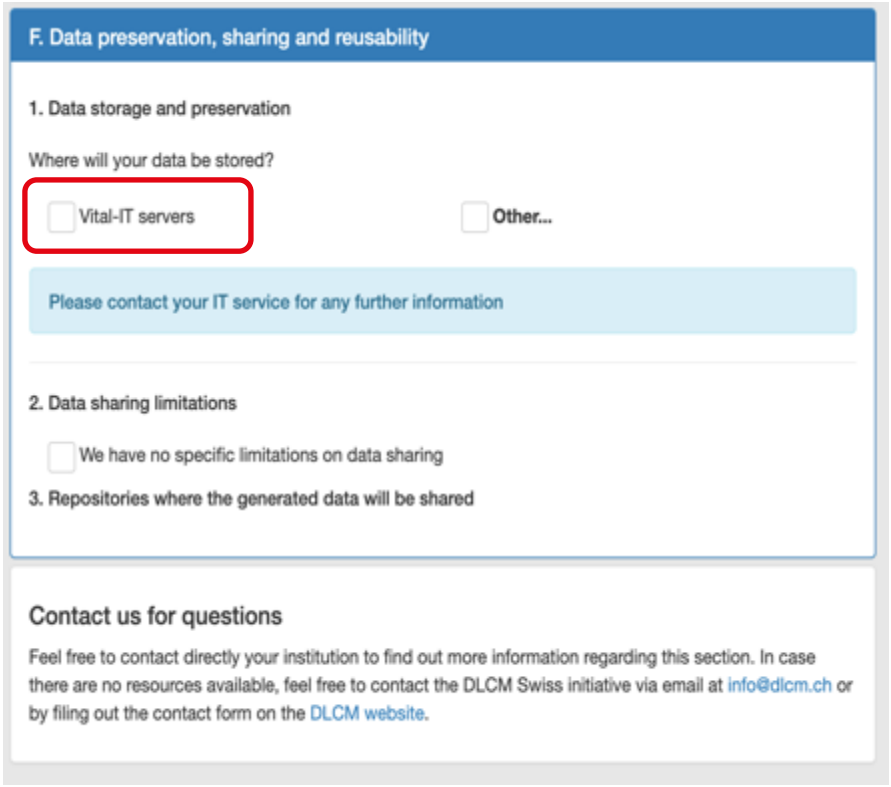
1.2 How will the data be collected, observed or generated?

[Questions you might want to consider. What tool(s) will be used for the analysis of your raw data? How will you ensure reproducibility?]

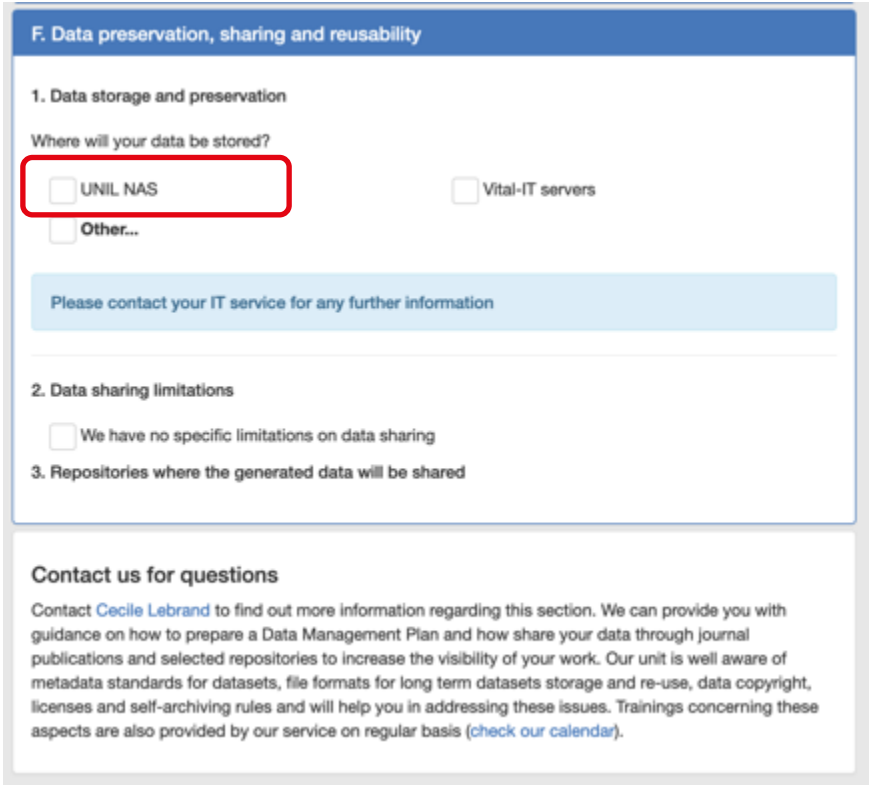
The produced raw data will be processed using the following tools: R and Custom script.

Questionnaire: Generic vs Specific

Generic



Specific (UNIL)



Template: Generic vs Specific

Generic

ORCID (generic)

1 Data collection and documentation

1.1 What data will you collect, observe, generate or reuse?

[Questions you might want to consider: What type and format of data will you collect/generate?]

Genomic data: Genomic data from genome-wide association studies (GWAS) will be stored in PLINK binary format. Processed exome sequencing data will be stored in VCF (variant call format). Other genomic sequencing data, for example whole genome bisulfite sequencing (WGBS) and ChIP-Seq will be stored as raw sequence (FASTQ) and alignments stored as compressed binary (BAM) files.

Microscopy data: Microscopy data will be collected as raw Zeiss LSM (confocal) or standard TIF (JEOL, TEM). Images can be exported in a variety of data formats (e.g. TIF, JPG).

Transcriptomic data: Affymetrix data will be stored as CEL files which will be described in an associated table with sample name, description and other information relevant to the sample such as RNA purity. RNA-Seq data will be stored in FASTQ format and alignments to genomic sequence in compressed binary (BAM) format.

We estimate that a total data volume ranging between 100 GB and 2 TB will be generated during the course of the project.

We plan on reusing Metabolomic data third-party datasets. [Please mention the owner of the third-party data, the format and the content of the datasets] For the aforementioned, third-party datasets, we have open access usage rights.

We plan on reusing Transcriptomic data datasets produced in our group.

[Questions to consider: What type of data, patient data management software (e.g. soariane/data warehouse, secutrial), format and volume of data will you collect/generate?].

1.2 How will the data be collected, observed or generated?

[Questions you might want to consider. What tool(s) will be used for the analysis of your raw data? How will you ensure reproducibility?]

The produced raw data will be processed using the following tools: R and Custom script.

Specific (UNIL)

UNIL

1 Data collection and documentation

1.1 What data will you collect, observe, generate or reuse?

[Questions to consider: What type of raw data, secondary data or analyzed data and volume of data will you collect/generate? What type of data acquisition and analysis software and file formats will you use? Also specify the acquisition equipment used to generate raw data?]

Genomic data: Genomic data from genome-wide association studies (GWAS) will be stored in PLINK binary format. Processed exome sequencing data will be stored in VCF (variant call format). Other genomic sequencing data, for example whole genome bisulfite sequencing (WGBS) and ChIP-Seq will be stored as raw sequence (FASTQ) and alignments stored as compressed binary (BAM) files.

Microscopy data: This project will generate raw data images [Please specify: For example, 3D stacks of confocal microscopy cell images, time lapse video microscopy images] acquired using [Please specify: For example, Zeiss LSM 710 Quasar, Leica SP5 equipment] with imaging software [Please specify: For example, ZEN lite software, LAS AF Lite 4.0.11706]. All data will be stored in digital form, either in the format in which it was originally generated [Please specify: For example, .lsm, .liff, .czi], or will be converted to [Please specify: For example, tiff uncompressed, jpeg2000 files, .Avi, .Mov] files. The raw data files will be processed and analyzed using various software [Please specify: For example, Imaris 7.2.1 software, Fiji/ImageJ, Metamorph software 6.0, Adobe Photoshop CS5]. Quantification analyses will be collected in [Please specify: For example, CSV or Excel files].

Transcriptomic data: Affymetrix data will be stored as CEL files which will be described in an associated table with sample name, description and other information relevant to the sample such as RNA purity. RNA-Seq data will be stored in FASTQ format and alignments to genomic sequence in compressed binary (BAM) format.

We estimate that a total data volume ranging between 100 GB and 2 TB will be generated during the course of the project.

We plan to reuse third-party Metabolomic data datasets. [Please mention the owner of the third-party data and the format and content of the datasets]. We have open access usage rights for the specified, third-party datasets.

DMP CANVAS GENERATOR "2.0"

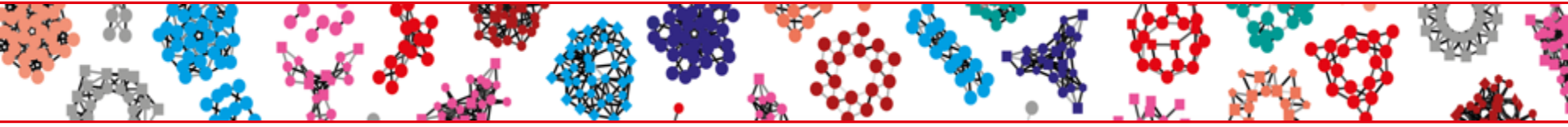
Integration of new templates (**X = Done;** X = Potential)

** Will integrate medical data and be renamed Biomedical*

Dimension 2: discipline

Dimension 1: institution	Life Sciences*	Theology	Geosciences	Social Sciences
GENERIC	**X**	X	X	X
UNIL	**X**			X
CHUV	**X**			
EPFL	X			
ETHZ	X	X		
UNIGE	X		X	X
UNIBE	X	X		X
UNIZH	X			
UNIFR		X		
SPHN	X			

Process followed during the hackathon



Starting point
Generic and specific DMP template for life sciences

Brainstorming
How to create generic templates for other domains (Generic – Domain-specific – Institution specific)

Creation of generic templates
Definition of the process

Analysis of DMP generator
Vital-IT : https://dmp.vital-it.ch/#/form

Functionalities of tool
How to improve/develop it ?

Existing solutions analysis

Analysis of existing solutions

Developer	Solution	Link
Canvas	DMP generator	https://dmp.vital-it.ch
DCC	DMP online	https://dmponline.dcc.ac.uk/
DCC	DMP Tool	https://dmptool.org
DCC	DMP online UNIL	https://dmp.unil.ch

Comparison of results (very short analysis!!)

DMP Canvas Generator (https://dmp.vital-it.ch)

+ login by Swiss Institution with Switch AAI
+ preselected option with branching logic
+ generation of documentation according to selected option
+ export on word
+ in conformity with SNSF Guidelines
+ specific links for more information

- only for one discipline: life science
- no description of administrative reference
- no text field

DMP online / DMP tool /DMPonline UNIL (same infrastructure)

+ Specific guide according to the institution
+ Additional tab for different contributors (only dmptool.org)
+ tab for sharing
+ Quick Start Guide, Data Management General Guidance
+ download in different format and different Font
+ final doc with date of version
+ access to public DMPs

- only text field

Definitions & Roles in the Canvas Generator

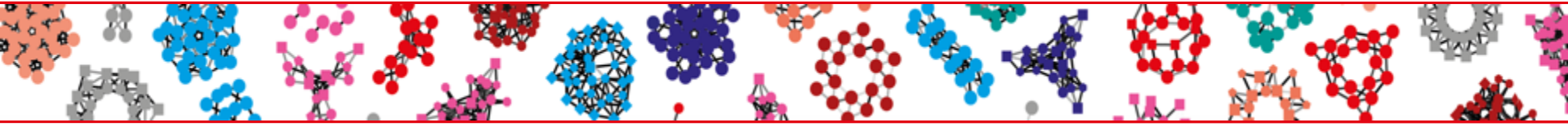
Definitions

- **DMP guidelines :** Doc from SNF
- **DMP template :** Final word document generated by the canvas generator (editable by template editors)
- **DMP questionnaire :** coded questions in the canvas generator done by a programmer
- **Generic template :** questions and answers specific to a domain but not to an institution (ex. *Life Sciences generic template* for all institutions).
- **Specific template :** questions and answers specific to a domain and an institution (ex. *UNIL Life Sciences specific template*)

Roles

- **User:** the person(s) doing the DMP
- **Editor:** a Data Curator in charge of creating an institution-specific template for his/her domain(s)
- **Admin:** it will probably be the Project Manager.
- **Developer:** in charge of the developments
- **Project Manager:** a person in charge of a project (e.g. develop a generic template for social sciences). It might be someone at SIB/Vital-IT or one of the data curators or a manager in a library.
- **Data curator:** A specialist knowing about data research, data types and formats, tools, methodologies and best practices

Topics we decideded to work on



Topic 1:

What is needed to create a generic template, whatever the domain?

Topic 2:

What is needed to create a specific template?

Topic 3:

Canvas Generator improvement

Topic 1: Process proposal to create a generic template

Methodology followed

- Define direction for hackathon: **What is needed to create a generic template, (whatever the domain)?”**

- Analyse the needs of a DMP with specialists

- Analyse the logic and philosophy behind https://dmp.vital-it.ch/#/form

- Create Word document containing questions from SNF guidelines and all options given by https://dmp.vital-it.ch/#/form

Topic 1: Process proposal to create a generic template

1. Preparation phase (PM)

- Selects a domain & identify local data curators from several institutions

- Creates guideline for data curators

2. Collection phase (DC)

- Visits labs, interviews researchers to gather information about data and tool usage and needs (generic name and specific brand)

3. Aggregation phase (PM+DCs)

- Collective work to have consensus information

- adapts the branching logic questionnaire by specifying between GENERIC and SPECIFIC.

- prepare two documents for the developer:
  - one with information for the questionnaire (website) -> Questions, options branching logic
  - one with information for the template (generic text in the Word template document) -> Generic content

Topic 2: What is needed to create a specific template

4. Additional phase (PM+ 1DC)

- Extract the **SPECIFIC** tasks and elements collected in labs (data types, analysis devices, software, storage…)

- Prepare a document with information for the institution-specific template (text in red in the Word template document)

TOPIC 1: PROCESS PROPOSAL TO CREATE A GENERIC TEMPLATE

Phase 1: prepare

- Project manager (PM) and data curator (DC) of institution identify a domain
- PM creates guideline for DC
- DC screens existing generic templates for reusable elements
- Collecting information and needs of the institution (PM and DC)

Phase 2: collect

- DC gathers information from labs on data and toolsdescription
- Identify what is for GENERIC and for SPECIFIC template

Phase 3: aggregate

- PM and DC put information together and find consensus
- Adapt branching logic
- prepare questionaire and template for the developer

TOPIC 2: PROCESS PROPOSAL TO CREATE A SPECIFIC TEMPLATE

Phase 1: prepare

- Project manager (PM) and data curator (DC) of institution identify a domain
- PM creates guideline for DC
- DC screens existing generic templates for reusable elements
- Collecting information and needs of the institution (PM and DC)

Phase 2: collect

- DC gathers information from labs on data and toolsdescription
- Identify what is for GENERIC and for SPECIFIC template

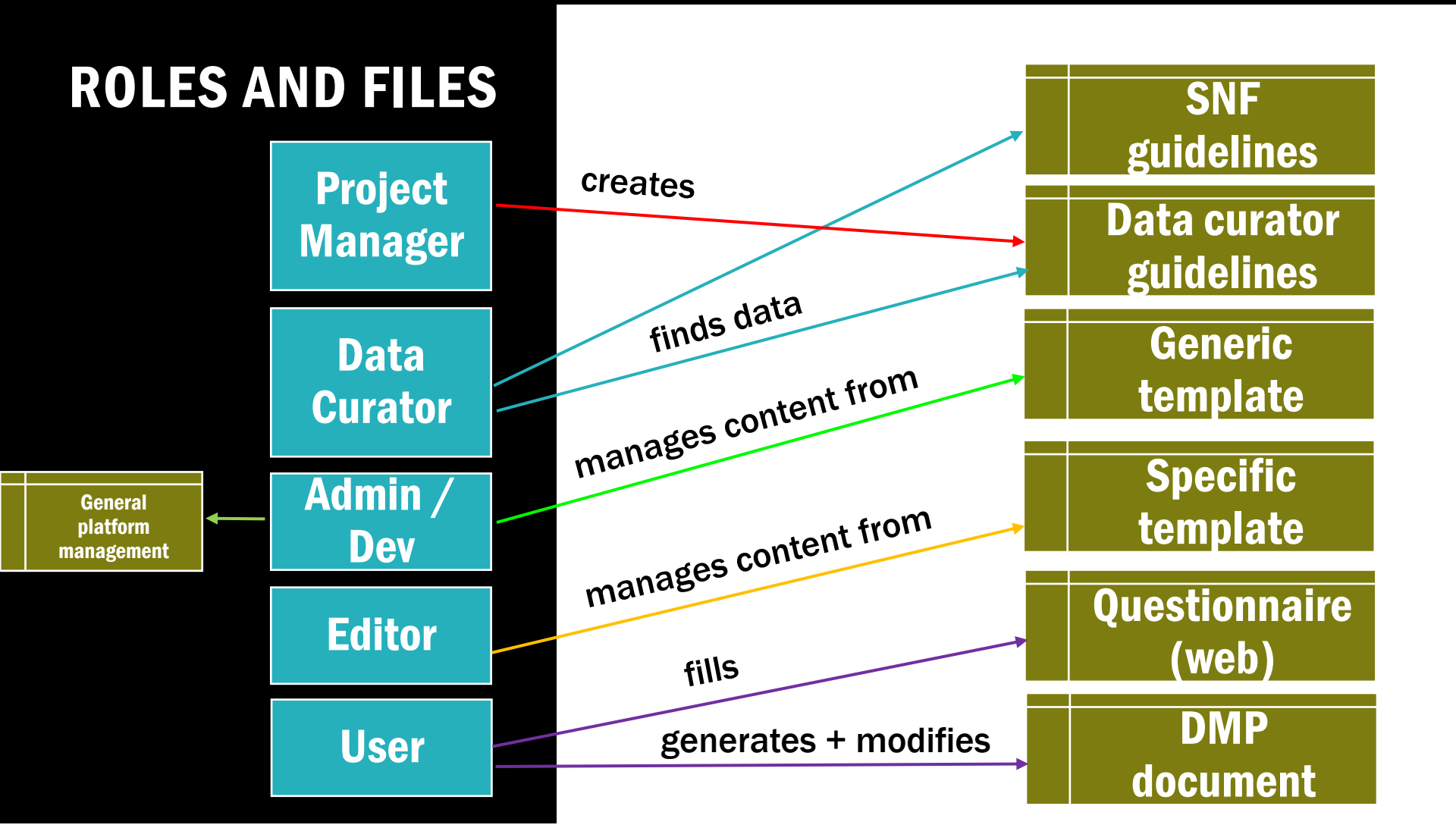
Phase 3: aggregate

- PM and DC put information together and find consensus
- Adapt branching logic
- prepare questionaire and template for the developer

Additional phase

- Extract institution-specific information
- Prepare document for specific questionnaire/template

ROLES AND FILES



Topic 3: Canvas 2.0 new functionalities

What we did

- Analyse the architecture and functionalities behind https://dmp.vital-it.ch/#/form

- Put ourselves at the place of the different roles to envision their needs

- Create a Product Backlog with User Stories for the different roles

Next steps

- Validate the PB with real users

- Build a student project to create a new version …. or …. find some funds to directly develop the new version

- Discuss with FORS about how to join their internal ThinkTank on the subject

Topic 3: CANVAS GENERATOR “2.0”: improvements

For the users we plan to add functionalities in the following categories

- As a User I can select my domain to load the correct template
- As a User I want specific information for each option (or each section) in order to understand what content is mandatory (institution and domain decision) and what is free for the user
- As a User I can create several DMPs at the same time
- As a User I can see the history of DMPs for reusing previous ones
- As a User I can save unfinished questionnaires
- As a User I can preview the effects of an option of the questionnaire on the template, with highlighted text modifications
- As a User I can finalize my template directly in the questionnaire interface (currently the word document) before generating the final version.
- As a User, given the previous point, I can export my final document in several formats, in particular Exchange API (in case SNSF portal is later compliant with this)
- As a User I can update my DMP during the project by using the saved version before generating an updated document.
- As a User I would like to work collaboratively on a DMP (2 users work on the same questionnaire)
- As a User I can report adaptations to bring to the system (functionalities, questions, content, …)

Topic 3: CANVAS GENERATOR “2.0”: improvements

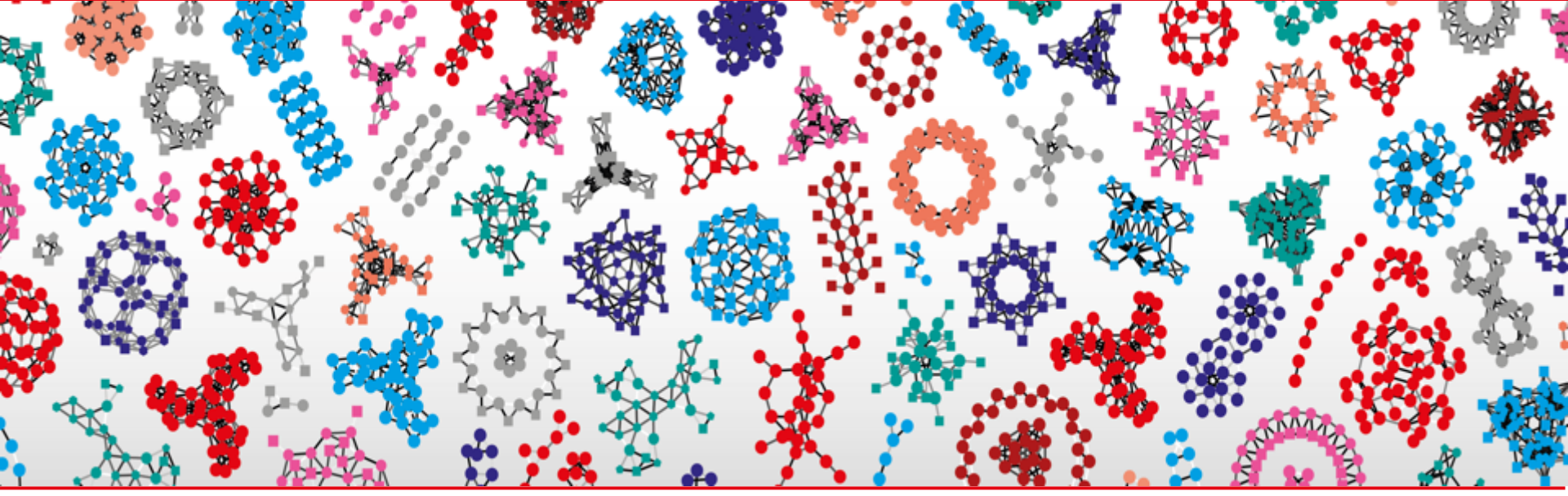
For the Editor we plan to add those functionalities about....

- As an Editor I can see the users of my institutions, but cannot modify their status to become editors
- As an Editor I can draft my template before publishing it and I should be warned if another editor is working on the template
- As an Editor I would be able to edit the introductory texts in each section for my specific template.
- As an Editor I would have access to logs and statistics for my institution/domain

Topic 3: CANVAS GENERATOR “2.0”: improvements

For the Admin we plan to add those functionalities about....

- As an Admin I can edit the questions, more precisely the options to click and the introductory texts (generic template)
- As an Admin I can report to the developer for modification of the questions and logic and contents if necessary, for other domains (second dimension)
- As an Admin, I would have access to logs and statistics
- As SNSF I can send notifications to the user when I need some answers to my questions



Thank you!

Anne-Laure Kaufman – Alexandre Cotting – Gregoire Rossier - Clemens Trautwein – Karine Villettaz- Daniela Subotic

Overview of Canvas DMP Generator 1.0

WHEN?

- Developed during DLCM project phase 1 (2015-2018)

WHAT?

- Web tool, which helps **researchers in Life Sciences** to generate DMPs for SNSF-funded projects.

HOW?

- Web questionnaire -> generation of a Word document (template) to be completed

ROLES

- User: fills the questionnaire to generate a DMP word doc
- Editor: generation of a template specific to his/her institution