Data Management Tool

Lecture: Data Management in Quanitative Biology

Names: Sven Fillinger, Simon Heumos, Judith Neukamm

June 11, 2015

1 Introduction and Motivation

The increasing amount of data and the participation of several institutions in a project makes it important to document well how to handle essential aspects like the storage, analysis and integration of data during the project. This can be realized with a Data Management Plan [3]. Moreover, this plan ensures before the data collection starts that data is in the correct format, well-organized and well annotated [4]. The documentation of the different steps throughout the data's life cycle helps data users to understand and use the data in the future.

Furthermore, the data management plan also makes the data available to other researchers upon project completion, which can impinge positive on the whole work, concerning discovery and relevance [4].

There exists no standardized guidance how to create a data management plan however the DAMA Data Management Body of Knowledge [2] provides a good orientation of essential aspects which should be part of the plan [3].

During the scope of the project, it was our task to develop a Data Management Planning Tool. The tool should be able to creates automatically a DMP based on an experimental design given as a .tsv file. This file was generated by *QWizard* [1], a portlet to input experimental data. The tool also offers users the possibility to add project information which are not included in the .tsv file.

The following chapters give an overview about the tool *CMPcreator* which was implemented during the scope of this project. The last chapters compare our tool with the already existing tool *DMPTool* developed by the University of California[1] followed by an outlook how our tool can be extended.

[1] `https://dmptool.org/`

2 Experimental Design Knowledge

One of the project's task was the implementation of results from an web-based experimental design tool called QWizard [1]. It is part of QBiC's web-based science gateway QPortal which handles scientific experimental projects and data. Customers are using the QWizard for setting up their experimental design and therefore provide information which already can be used for the data management plan. To name some of the information that will be available from the QWizard are detailed description of the biological entity, such as treatment or phenotype, as well as the species. Also the sample extraction is described, listing the type of tissue used as well as defined conditions used. Moreover, the QWizard will provide information of the sample size and technologies used for retrieving the data.

This information also belongs into the data management plan as it is part of the experimental design and therefore essential for the project planning and execution. We can also use the information for supporting the user during the process of generating a data management plan by suggesting storage amount and backup solutions for example.

In order to use the information given by the QWizard, we will provide an upload section, where the user can upload the file with all the content generated from QWizard and a background parser will retrieve the information as well as automatically implementing it in the report template.

3 System Architecture

3.1 Java Framework

For the setup of an own data management planning tool we decided to use the `Vaadin Framework`[2]. `Vaadin` is a single-page web framework for Java developers that provides powerful functions for creating rich Internet applications (RIAs) without the knowledge of classical web-languages as HTML5 or CSS3. It also provides a huge library of functional components that can be included in the project and already satisfy out needs for usability and navigation. The `Vaadin` Framework takes care of browser incompatibilities and automatically designs ajax communication protocols, which we evaluated as advantage in terms of time efficiency during the development. It is also open-source but still provides a support service.

As we are all quite familiar with Java, the availability of a huge component library, good documentation and support as well as the capacity to build a rich Internet application convinced us to give it a chance and use it for our own creation of a data management planning tool.

3.2 Java Application Server

In order to run the project, local or on a web-server, we needed to set-up a Java application server. There are numerous servers available, most of them are free and open-source. For

[2] `https://vaadin.com/home`

our needs, we thought it does not make a big difference which server we select as we will not use the full capacities anyway. Apache's Tomcat was already known to us and was used before in other projects. To get to know a new application server which also comes with a nice documentation and web interface is JBoss' Wildfly[3]. In our case the installation and set-up of Wildfly was easy under Linux OS. So we configured local instances on every development environment and used the server's default settings.

3.3 Vaadin Theme

The `Vaadin Framework` also comes with additional themes, that apply a different layout and style to the graphical interface elements. We decided to choose the theme 'Valo', as it comes with a complete set of designed components, is responsive and ensures a pleasant user experience.

4 Result: DMPcreator

The result of our implementation efforts is the user friendly webinterface `DMPcreator`. It enables scientists to create a data management plan based on the .tsv file produced by `QWizard`. The tool is structured into five slides. A progress bar on top of every slide gives an excellent overview of the current advances. Furthermore, the user can navigate through the slides by clicking at the buttons at the bottom on the right (present in every slide). Moreover, every slide contains a detailed description about what, where and why the user has to fill in here. Note, that every field in every slide will be parsed in the end and the information is stored in a human readable PDF file.

The first slide, the *General Information Slide* 4.1, provides fields letting the user enter general information about his project, for example the project name, the person in charge of the project and contact data. Furthermore, the user can upload his .tsv file created by `Q-Wizard`. From this uploaded data experiment details like species, experiment types, number of experiment per experiment type, and so forth are extracted for later usage (slide *Storage & Backup*4.4 needs this information in order to calculate required storage correctly).

One important topic that needs to be covered when creating a data management plan is *Roles & Responsibilities* 4.2 of every project member. This second slide allows the user to assign roles to persons. The chosen values are added two a responsibilities list later being parsed by the PDF creator.

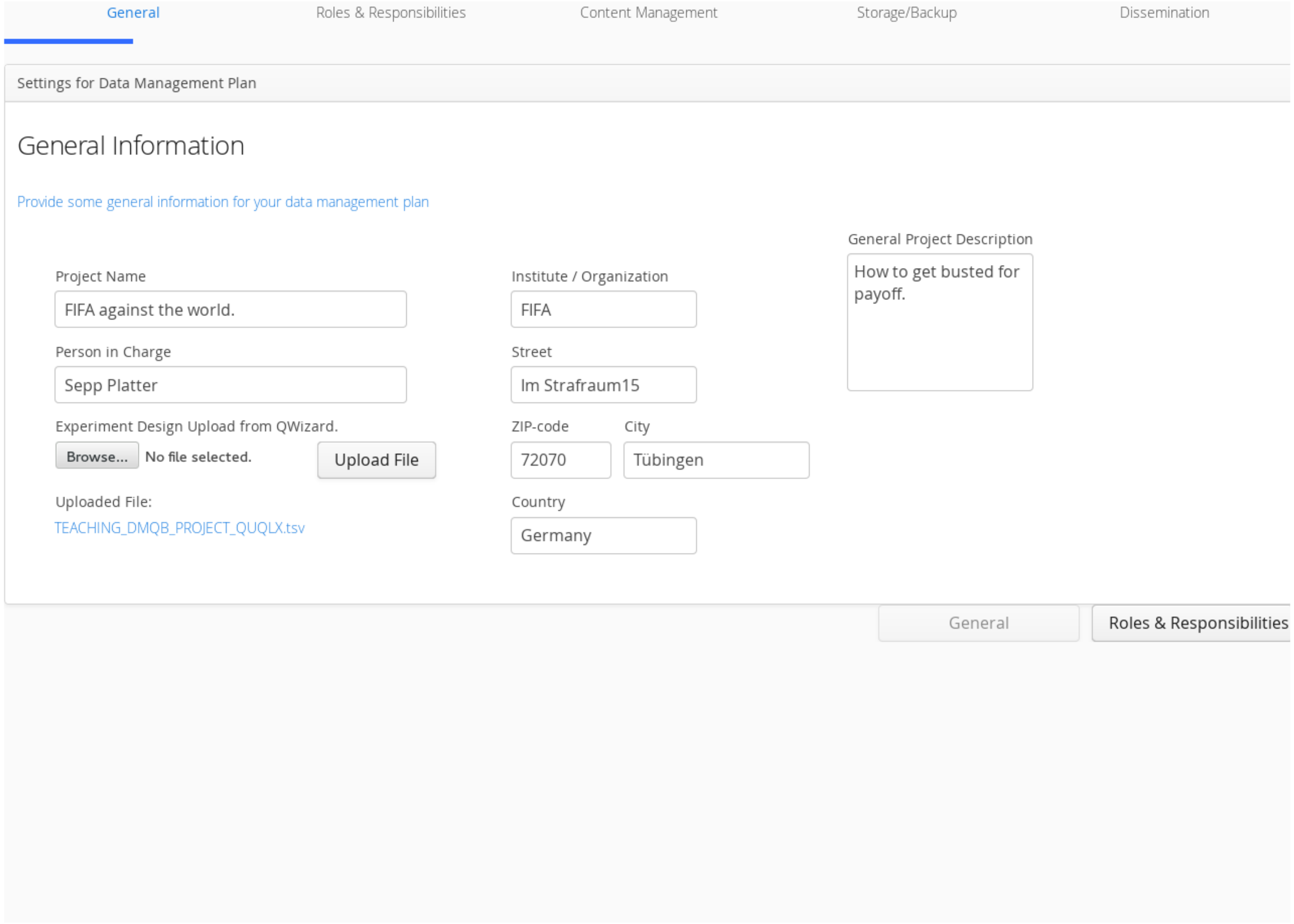
Having specified who is responsible for which data, the user still has to decide, how the data is stored. Here steps the third slide *ContentManagement*4.3 in. The user can assign file types to an associated description building a content. PDF creator parsed these user created contents, too.

How to store data efficiently is one key component in a data management plan. In the slide *Storage & Backup*4.4 the user is able to fill in the storage location, the RAID backup solution, the archive solution and how much data (in GB) one experiment will produce

[3] `http://wildfly.org/`

approximately. Changing the needed disk space for an experiment results in an update of the displayed total required space. So the user gets an immediate feedback when filling in necessary fields.

The last slide covers the topic of *Dissemination*4 of data. This mask provides fields for the user to generate dissemination methods. Furthermore, by clicking on the 'Generate Report' button, the PDF creator produces the PDF. Another button ('Download Report') then allows the user to download the created report.



Settings for Data Management Plan

General Information

Provide some general information for your data management plan

Project Name: FIFA against the world.

Person in Charge: Sepp Platter

Experiment Design Upload from QWizard.

Uploaded File: TEACHING_DMQB_PROJECT_QUQLX.tsv

Institute / Organization: FIFA

Street: Im Strafraum15

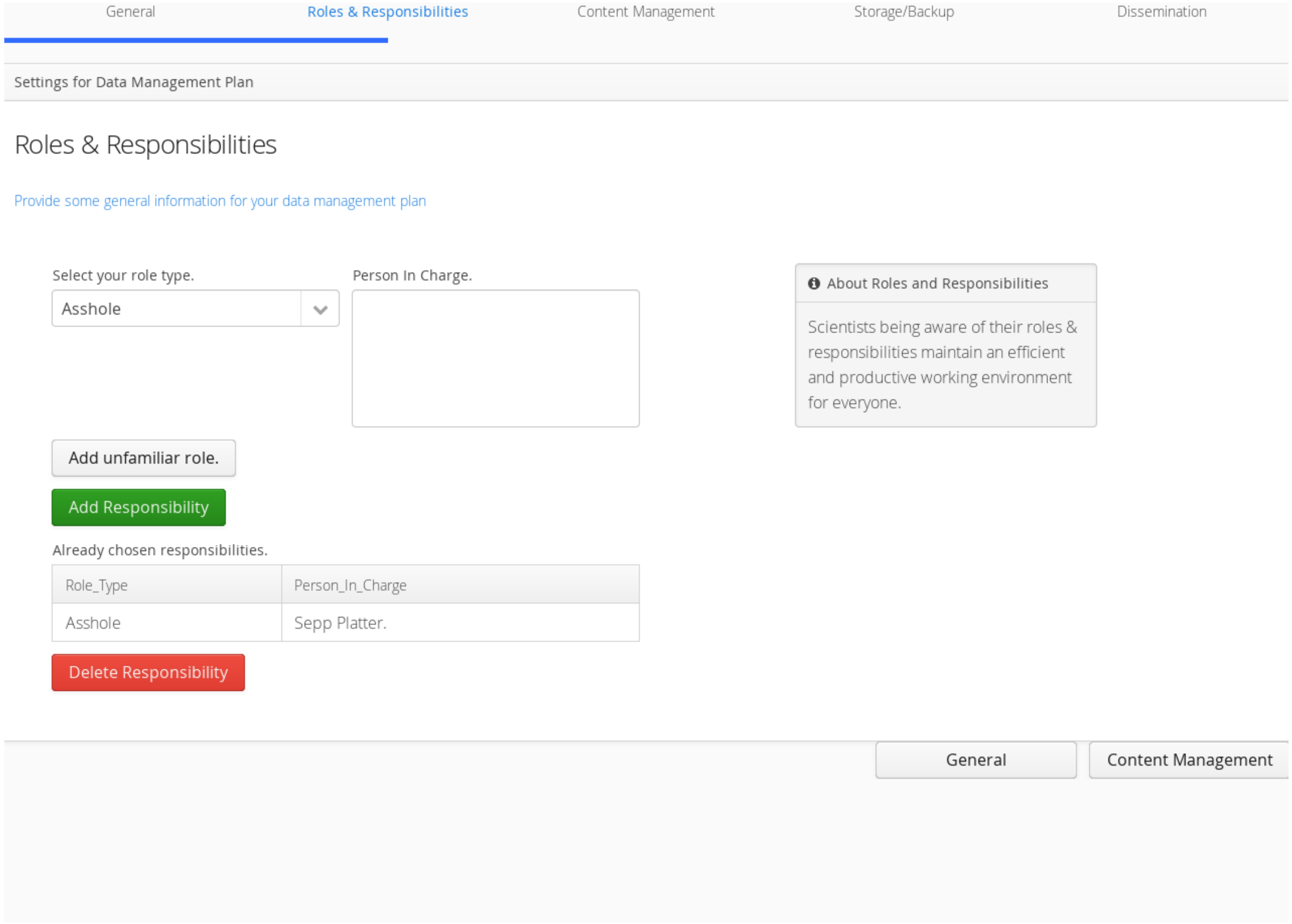
ZIP-code: 72070

City: Tübingen

Country: Germany

General Project Description: How to get busted for payoff.

Figure 4.1: *General Information* Slide of `DMPcreator`. The progress bar is placed on the top. Fields that are fillable by the user can be seen below. Note, a special upload field for the .tsv file from `Q-Wizard` is visible on the left bottom.



Settings for Data Management Plan

Roles & Responsibilities

Provide some general information for your data management plan

Select your role type.

Asshole

Person In Charge.

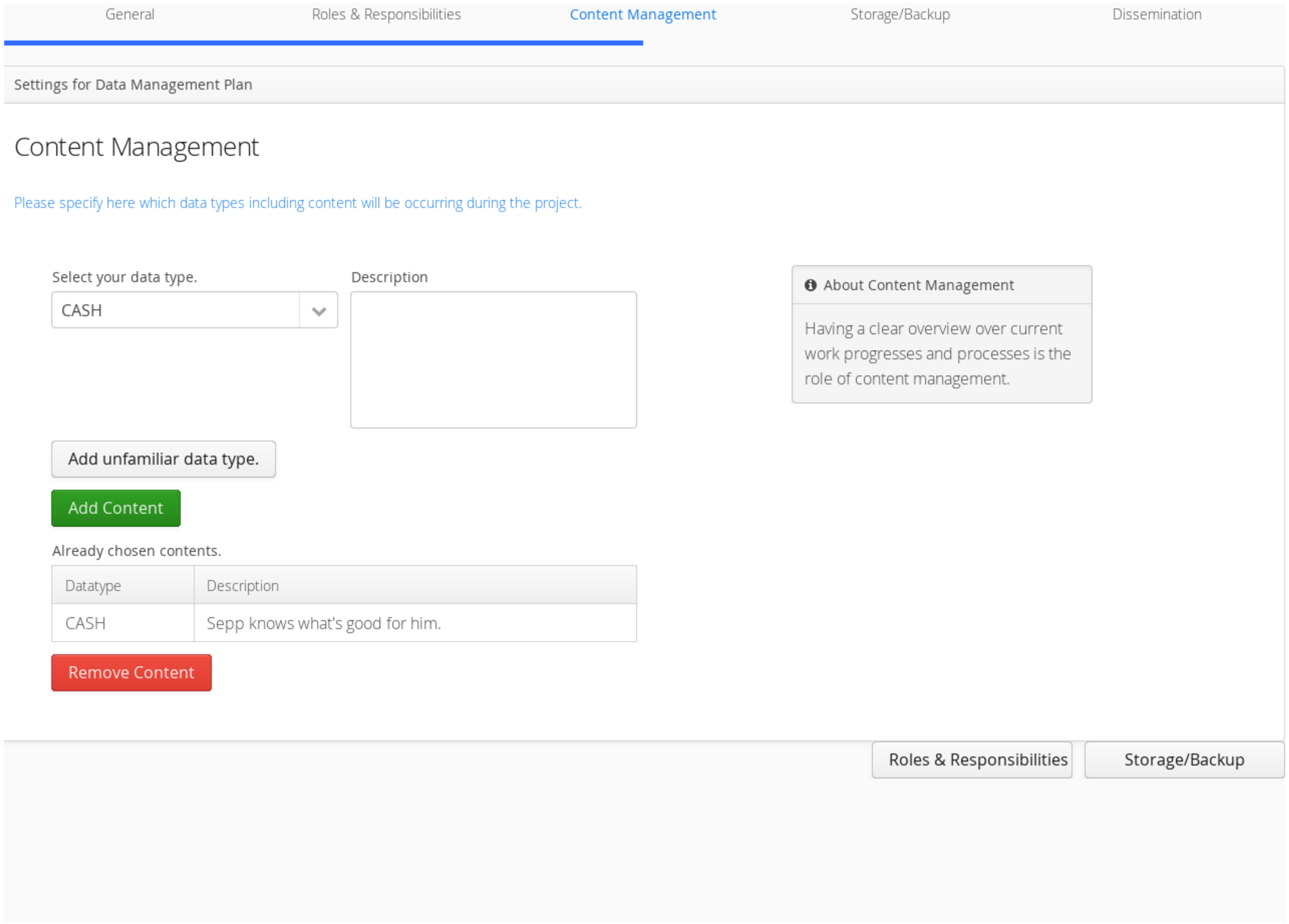
About Roles and Responsibilities

Scientists being aware of their roles & responsibilities maintain an efficient and productive working environment for everyone.

Already chosen responsibilities.

Role_Type	Person_In_Charge
Asshole	Sepp Platter.

Figure 4.2: *Roles & Responsibilities* Slide of `DMPcreator`. The progress bar is placed on the top. Fields that are fillable by the user can be seen below.



Settings for Data Management Plan

Content Management

Please specify here which data types including content will be occurring during the project.

Select your data type.

CASH

Description

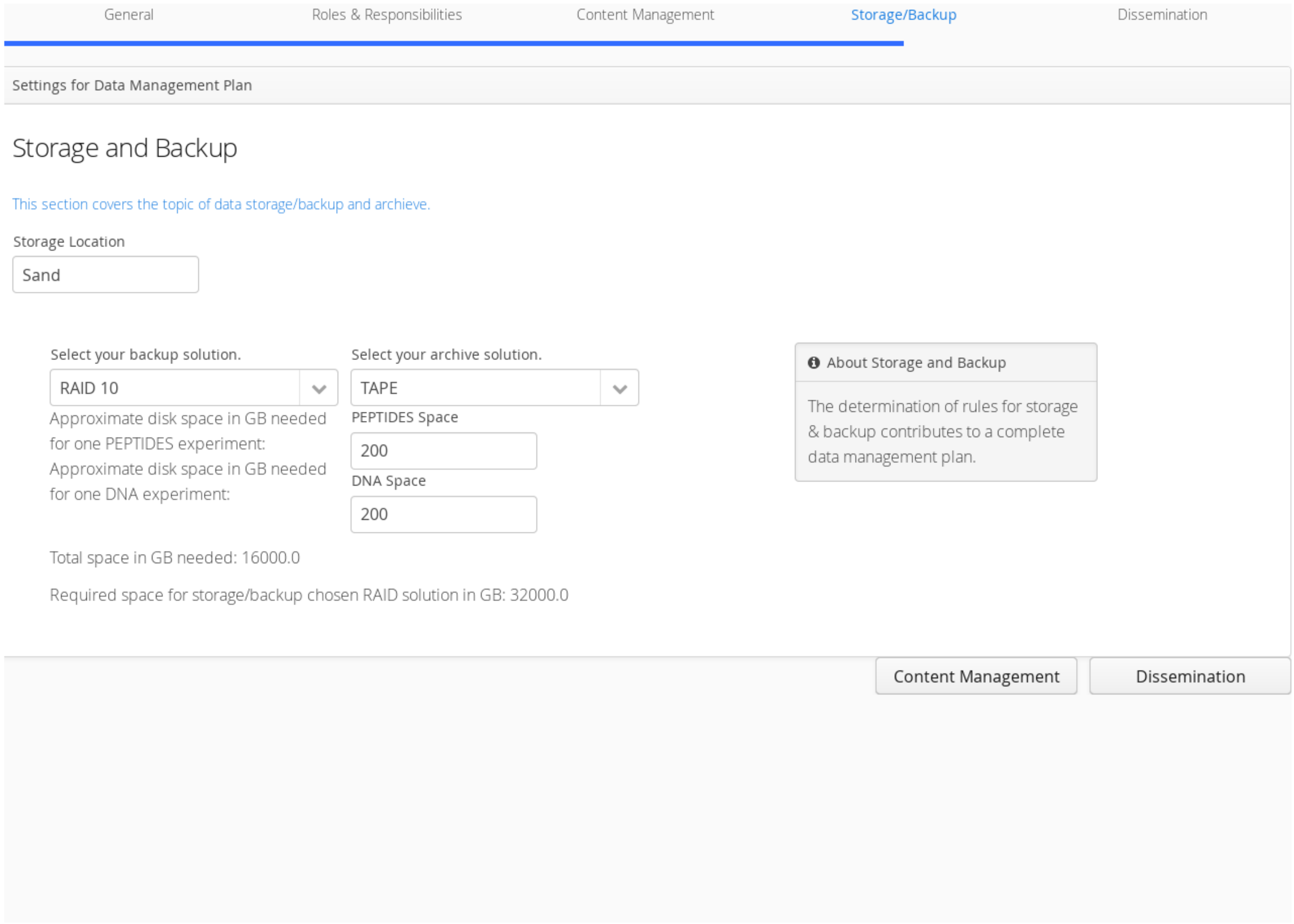
About Content Management

Having a clear overview over current work progresses and processes is the role of content management.

Already chosen contents.

Datatype	Description
CASH	Sepp knows what's good for him.

Figure 4.3: *Content Management* Slide of `DMPcreator`. The progress bar is placed on the top. Fields that are fillable by the user can be seen below.



Settings for Data Management Plan

Storage and Backup

This section covers the topic of data storage/backup and archieve.

Storage Location: Sand

Select your backup solution.: RAID 10

Select your archive solution.: TAPE

Approximate disk space in GB needed for one PEPTIDES experiment:

PEPTIDES Space: 200

Approximate disk space in GB needed for one DNA experiment:

DNA Space: 200

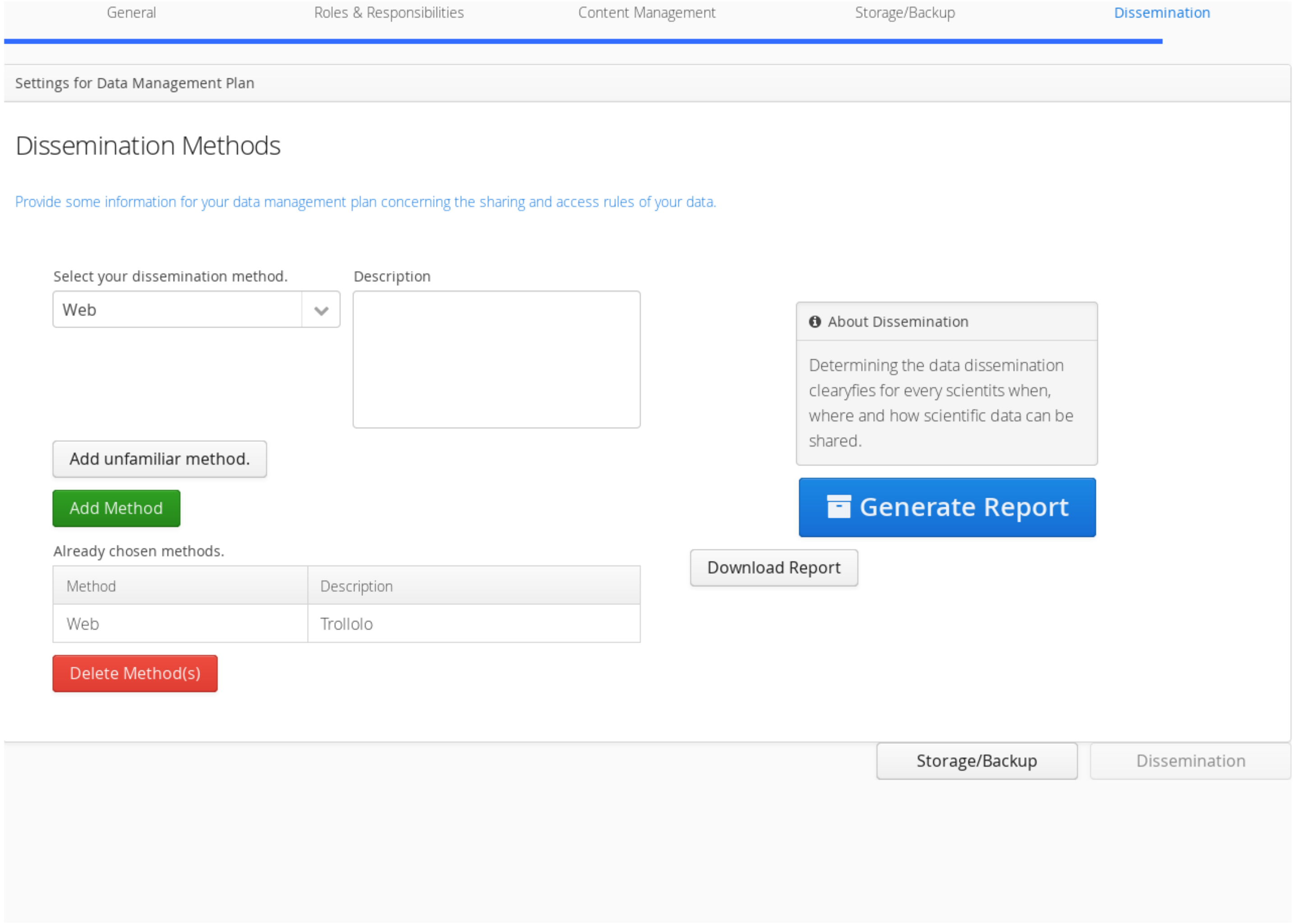
Total space in GB needed: 16000.0

Required space for storage/backup chosen RAID solution in GB: 32000.0

About Storage and Backup

The determination of rules for storage & backup contributes to a complete data management plan.

Figure 4.4: *Storage & Backup* Slide of `DMPcreator`. The progress bar is placed on the top. Fields that are fillable by the user can be seen below.



Settings for Data Management Plan

Dissemination Methods

Provide some information for your data management plan concerning the sharing and access rules of your data.

Select your dissemination method.
Web

Description

About Dissemination

Determining the data dissemination clearyfies for every scientits when, where and how scientific data can be shared.

Already chosen methods.

Method	Description
Web	Trollolo

Figure 4.5: *Dissemination* Slide of `DMPcreator`. The progress bar is placed on the top. Fields that are fillable by the user can be seen below.

5 Discussion

Proper data management will become more and more important in the future, as new developments of technologies are going to produce an expanding amount of data. This is in particular true for scientific applications. The recent developments in next-generation sequencing and mass spectrometer analysis have shown, that there is a flood of data that is created during a project's lifetime. Initiatives have been started over the past years, to formulate standards in data management. The DAMA International Foundation is an example for such an initiative, that took steps towards an important direction. It has defined the Data Management Book of Knowledge (DMBOK), which is a guideline for proper data management and categorizes the task in eleven different fields.

During the lecture, we got to know the Data Management Plan Tool (DMPtool) from the University of California, which provides a guided web-interface for generating a data management plan. We experienced the tool as a step in the right direction, but remark, that it is not standardized at all, concerning the input the user can do in order to create such a plan. There are a lot of text-fields, which allow the input of free text, but this raises issues at the same time. The user can input what he wants, and there is no standard for that, nor do unexperienced user have a glue what to fill in. Though the tool provides templates, which are plans created from other users, the quality of the input varies a lot.

With our tool DMPcreator, we want to present a data management plan tool that follows the DMBOK guideline and controls some of the standards mentioned there. Additionally, we provide an integration for projects created with QWizard, a web-based experimental design tool from the Quantitative Biology Center (University of Tuebingen). The user can upload the project description as tabbed-separated file (tsv), and this content will be parsed automatically and information such as storage requirement derived from the technology type will be integrated in the web-interface as well as in the report automatically. The user can still change the input, if necessary. The taxonomy ID in the tsv-file will be queried automatically from the NCBI web-server and translated in the organism species. We used Vaadin as a Java framework and provide a user-friendly interface with intuitive and customizable components such as flexible drop-down menus with predefined content and dynamic tables that will display the user input. Because of the short project time, we only implemented topics from the DMBOK such as 'Storage and Backup', 'Documentation Management', 'Dissemination', 'Roles and Responsibilities' and 'General Information' for contact data. Of course, in order to provide a detailed data management plan a lot more chapters and options have to be added. We designed the software as generic as possible, developers can add more user slides and extend the class for parsing the content into the report as pdf.

We have shown, that a more user-friendly and more standardized data management plan tool is possible, and has to be the further development. The DMBOK provides the ideal template for such a tool and frameworks offer the components and functionality to build a detailed and guided interface to generate a content-rich and standardized data management plan.

6 Outlook

The *DMPcreator* covers just a small fraction of the aspects of a data management plan. Therefore, it could be extended by adding more user slides concerning data security and data sharing for example. Furthermore, the .tsv file could be parsed more detailed. It also contains information about the experiment structure, the treatment of the individuals and the tissues, which are not yet used by generating the data management plan.

Moreover, additional fields for user input could be added. This provides the user the possibility to create a more specific and adapted data management plan.

References

[1] Andreas Friedrich, Erhan Kenar, Oliver Kohlbacher, and Sven Nahnsen. Intuitive web-based experimental design for high-throughput biomedical data. *BioMed Res Int*, 2015:958302, 2015.

[2] DAMA International. *The DAMA Guide to the Data Management Body of Knowledge - DAMA-DMBOK*. Technics Publications, LLC, USA, 2009.

[3] Sven Nahnsen. ’lecture notes in data management in quantitative biology’, April 2015.

[4] MASSACHUSETTS INSTITUTE OF TECHNOLOGY. Data Management. `http://libraries.mit.edu/data-management/plan/why/`. [Online; accessed 11-June-2015].