LSDF-Portal: Project administration for the Large Scale Data Facility

Felix Karg

Wednesday, February 16th, 2022

Steinbuch Centre for Computing (SCC)

What is the LSDF-Portal

Introduction to Django

Project Results

Conclusion

What is the LSDF-Portal

Overview

Project Creation

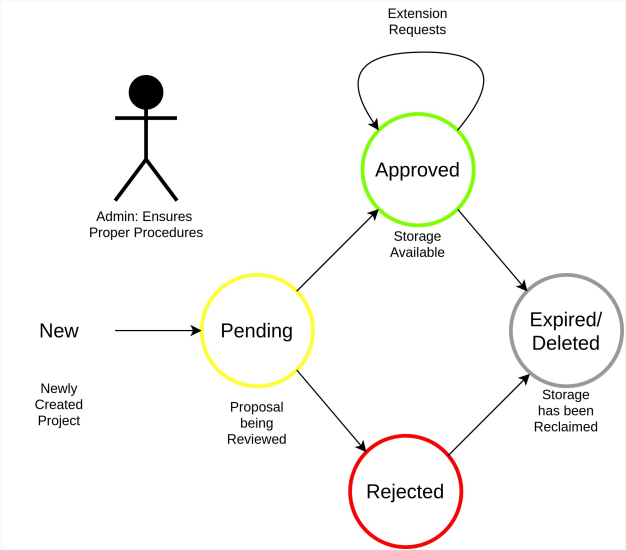
What is the Large Scale Data Facility?

- Cloud-based storage for scientific research
- About 18 PetaByte (1PB = 1000 TB)
- All commonly used Storage Protocols
- Direct Integration with e.g. HoReKa HPC-systems
- For KIT, Baden-Württemberg, and worldwide partnered researchers
- LSDF-Portal: *administration* of storage projects

What is the LSDF-Portal?

- LSDF-Portal: *administration* of storage projects
  - Research groups can request storage capacity for research projects
  - Allows for structured interaction with admins
  - Transparent overview of project status
- Internally available at `https://lsdf.kit.edu`
- My test instance available at `https://lsdf.fkarg.de`
  - Play around with `user:test` or `admin4:lsdf` (both admin)
  - (Cannot properly route from within KIT for some reason)

Project Lifecycle



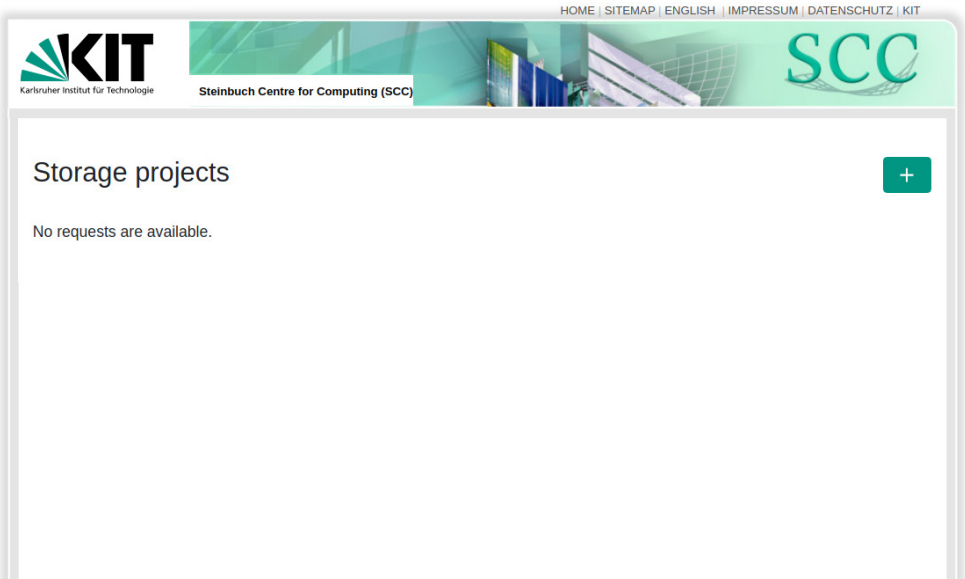
Live Demo!

What is the LSDF-Portal

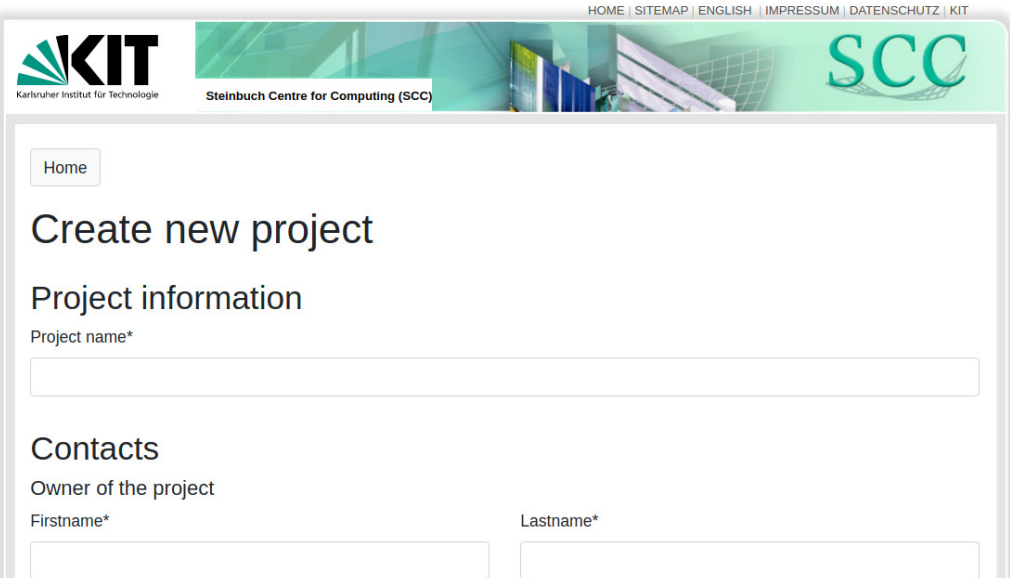
Overview

Project Creation

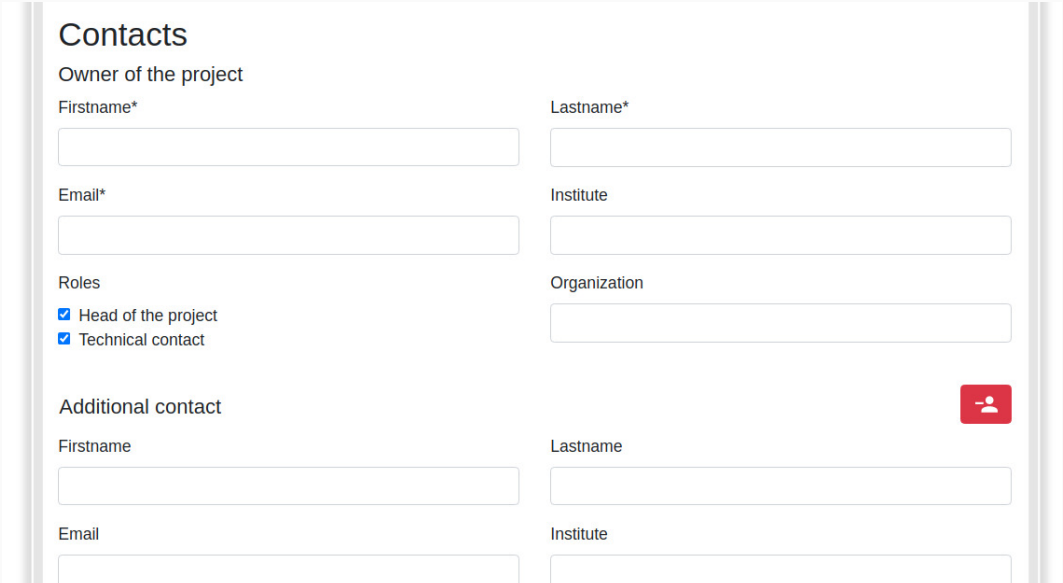
First Login


HOME | SITEMAP | ENGLISH | IMPRESSUM | DATENSCHUTZ | KIT
Karlsruher Institut für Technologie
Steinbuch Centre for Computing (SCC)
SCC
Storage projects
+
No requests are available.

Project Creation



Project Creation: Add Contacts



Available Fields

Please Specify where your project is mainly located according to the DFG Fachsystematik:

DFG Discipline*

DFG Review Board*

DFG Subject Area*

End of the project*

06.02.2023

How long you need the storage. You will be able to request an extension

Capacity*

Expected storage capacity in TB (1 TB = 1000 GB)

Directory name*

Fields filled out

Please Specify where your project is mainly located according to the DFG Fachsystematik:

DFG Discipline*

Natural Sciences (3)

DFG Review Board*

Mathematics (312)

DFG Subject Area*

Mathematics (312-01)

End of the project*

06.02.2023

How long you need the storage. You will be able to request an extension

Capacity*

15

Expected storage capacity in TB (1 TB = 1000 GB)

Submit Proposal

Protocols

- [x] SSH, SFTP, SCP, HTTPS/Web are enabled for all storage projects
- [ ] CIFS
- [ ] NFS V3 (Client needs to be connected to KIT-IDM)

Access control

Owner name*

Who should be the owner the project directory? The owner can be a KIT user (e.g. ab1234) or a KIT service account (e.g. OE-ProjectName-0001). Please, contact your ITB to create a service account.

Group name*

Which group should get access to your project directory (e.g. OE-ProjectName-LSDF)? Please contact your ITB to create a group.

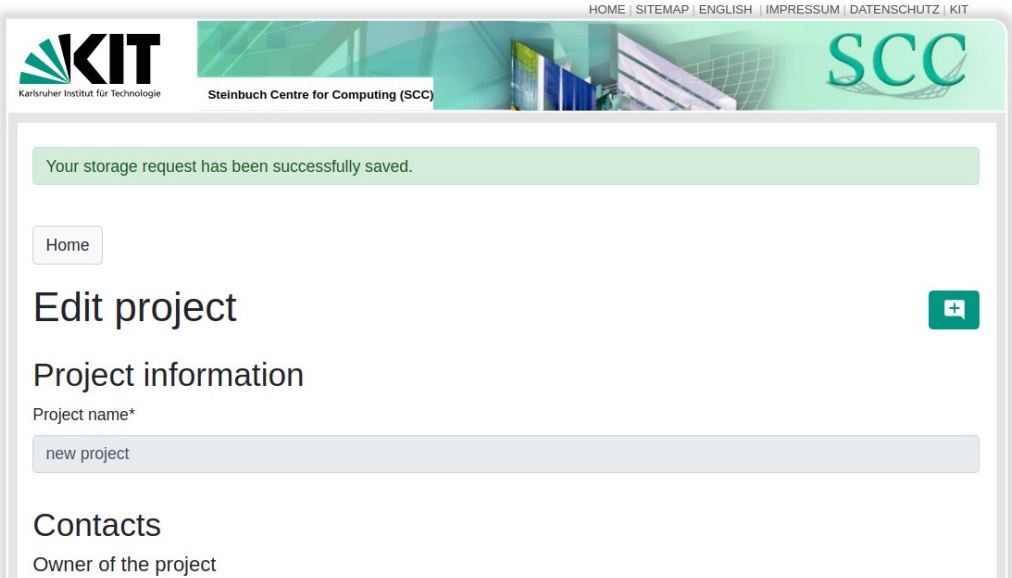
Group permission*

No permissions

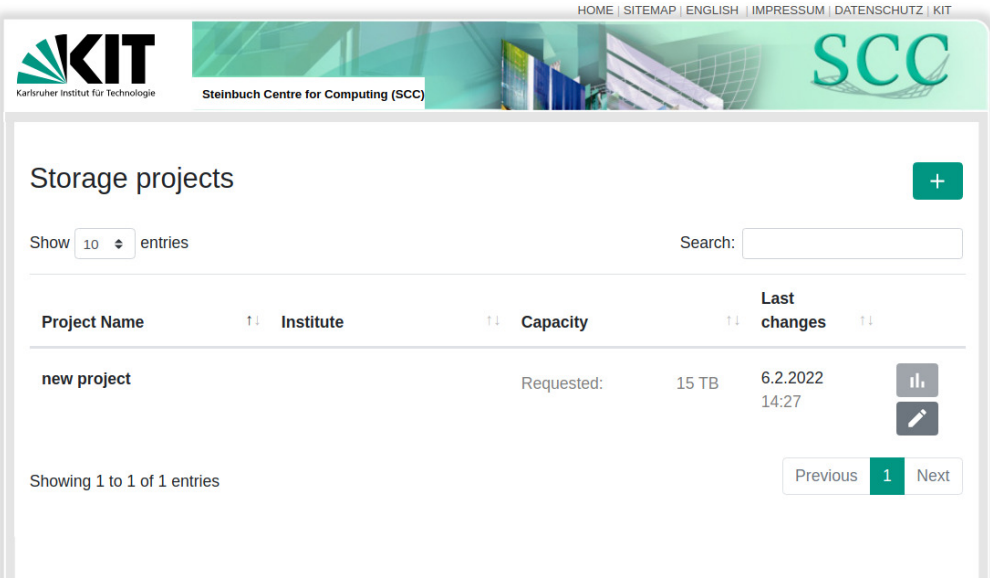
- [ ] Extended group permissions (ACLs)

Submit Storage Proposal

Succesful Submission



Project in List



Storage Use Histogram

(TODO: Screenshot: Histogram of storage usage.)

Django is a Popular Framework

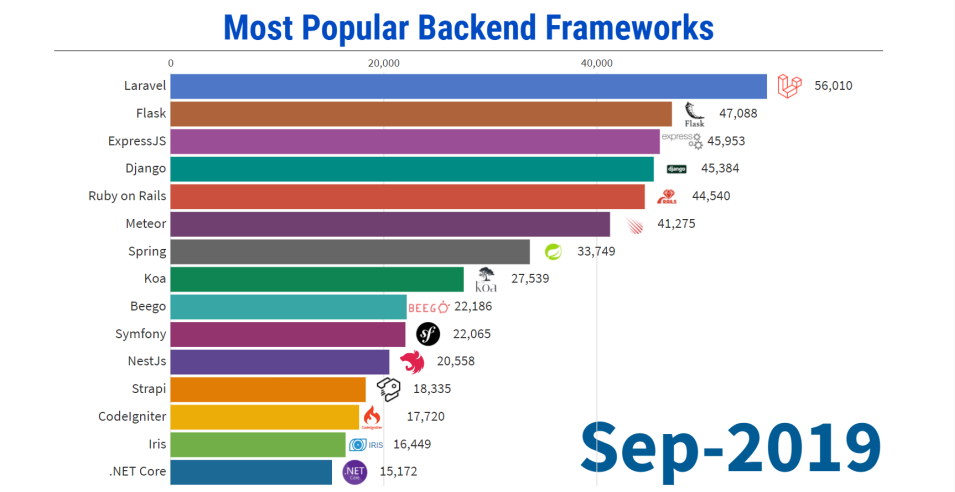


Image source: [1]

What is Django?

Meet Django

Django is a high-level Python web framework that encourages rapid development and clean, pragmatic design. Built by experienced developers, it takes care of much of the hassle of web development, so you can focus on writing your app without needing to reinvent the wheel. It’s free and open source.

Ridiculously fast.

Django was designed to help developers take applications from concept to completion as quickly as possible.

Exceedingly scalable.

Some of the busiest sites on the web leverage Django’s ability to quickly and flexibly scale.

Django Overview

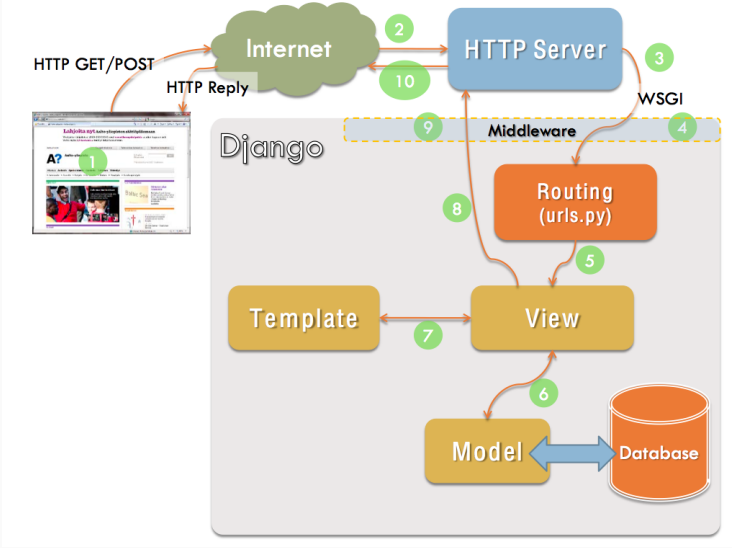


Image source: [2]

Project Results

Proper Logging

Automatically Generate Documentation

What Research is the Storage used for?

Diff CSV to DFG Schema in Database

Attempt to properly modularize PersonForm

Extension Requests

Changes to Logging

- Used to be done in inconsistent format over individual `print` statements
- Good to get overview of Codebase and touch many files
- Now: Properly defined Logging levels (debug, info, warn, error, critical)
- Now: Log messages are sent to console, syslog and saved in a file locally
- Now: file and console formatting differs (timestamps, colors, ...)
- Now: Rotating files: Keep last five days, overwrite after

Logging Now: Example

```
INFO     [basehttp.log_message]: "GET / HTTP/1.1" 302 0
INFO     [basehttp.log_message]: "GET /admin/login?next=/ HTTP/1.1" 302 0
INFO     [basehttp.log_message]: "GET /admin/login/?next=/admin/login%3Fnext%3D/ HTTP/1.1" 200 2245
INFO     [basehttp.log_message]: "GET /static/admin/js/nav_sidebar.js HTTP/1.1" 200 1360
INFO     [basehttp.log_message]: "GET /static/admin/css/nav_sidebar.css HTTP/1.1" 200 2271
INFO     [basehttp.log_message]: "GET /static/admin/css/login.css HTTP/1.1" 200 939
INFO     [basehttp.log_message]: "GET /static/admin/css/base.css HTTP/1.1" 200 19513
INFO     [basehttp.log_message]: "GET /static/admin/css/responsive.css HTTP/1.1" 200 18545
INFO     [basehttp.log_message]: "GET /static/admin/css/fonts.css HTTP/1.1" 200 423
INFO     [basehttp.log_message]: "GET /static/admin/fonts/Roboto-Light-webfont.woff HTTP/1.1" 200 85692
INFO     [basehttp.log_message]: "GET /static/admin/fonts/Roboto-Regular-webfont.woff HTTP/1.1" 200 85876
WARNING  [log.log_response]: Not Found: /favicon.ico
WARNING  [basehttp.log_message]: "GET /favicon.ico HTTP/1.1" 404 7000
INFO     [basehttp.log_message]: "POST /admin/login/?next=/admin/login%3Fnext%3D/ HTTP/1.1" 200 2405
INFO     [basehttp.log_message]: "GET /static/admin/fonts/Roboto-Bold-webfont.woff HTTP/1.1" 200 86184
INFO     [basehttp.log_message]: "GET /admin/login/?next=/ HTTP/1.1" 200 2205
INFO     [basehttp.log_message]: "POST /admin/login/?next=/ HTTP/1.1" 200 2365
INFO     [basehttp.log_message]: "GET /admin/login/ HTTP/1.1" 200 2204
INFO     [basehttp.log_message]: "POST /admin/login/ HTTP/1.1" 200 2364
INFO     [basehttp.log_message]: "GET / HTTP/1.1" 302 0
INFO     [basehttp.log_message]: "GET /admin/login?next=/ HTTP/1.1" 302 0
INFO     [basehttp.log_message]: "GET /admin/login/?next=/admin/login%3Fnext%3D/ HTTP/1.1" 200 2245
INFO     [basehttp.log_message]: "POST /admin/login/?next=/admin/login%3Fnext%3D/ HTTP/1.1" 200 2405
INFO     [basehttp.log_message]: "GET /persons/ HTTP/1.1" 200 13898
```

Project Results

- Proper Logging
- **Automatically Generate Documentation**
- What Research is the Storage used for?
- Diff CSV to DFG Schema in Database
- Attempt to properly modularize PersonForm
- Extension Requests

Why attempt automatic documentation?

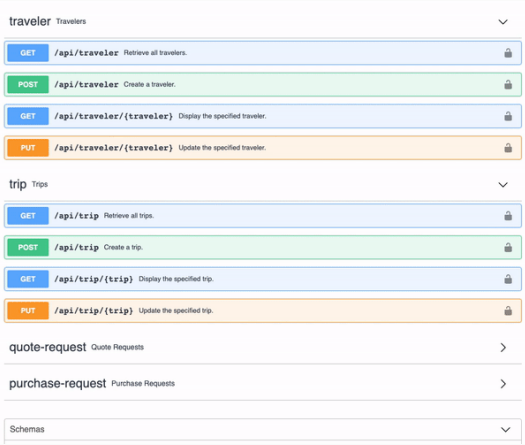
Situation:

- Still only rudimentary grasp on Codebase
- Especially on available routes, endpoints, internal dependencies

Expectations:

- Automatic, Complete, Up-to-date Overview
- Creation of a frequently used reference
- Provides a documentation pipeline and default

Automatically Generate Documentation



(Success from different Project)

Failed, because:

- Autogeneration usually used to generate documentation for *JSON endpoints*
- 'Primitive' Views lacking important information for automatic generation

Was worth a try.

Project Results

- Proper Logging
- Automatically Generate Documentation
- **What Research is the Storage used for?**
- Diff CSV to DFG Schema in Database
- Attempt to properly modularize PersonForm
- Extension Requests

What Research is the Storage used for?

Engineering Sciences (131 Members)		
RB-Nr.	Review Board / Subject Area	Subject Areas
401	Production Technology	⊞
402	Mechanics and Constructive Mechanical Engineering	⊞
403	Process Engineering, Technical Chemistry	⊞
404	Fluid Mechanics, Technical Thermodynamics and Thermal Energy Engineering	⊞
405	Materials Engineering	⊞
406	Materials Science	⊞
407	Systems Engineering	⊞
408	Electrical Engineering and Information Technology	⊞
409	Computer Science 409-01 Theoretical Computer Science 409-02 Software Engineering and Programming Languages 409-03 Security and Dependability 409-04 Operating, Communication, Database and Distributed Systems 409-05 Interactive and Intelligent Systems, Image and Language Processing, Computer Graphics and Visualisation 409-06 Information Systems, Process and Knowledge Management 409-07 Computer Architecture and Embedded Systems 409-08 Massively Parallel and Data-Intensive Systems	⊟
410	Construction Engineering and Architecture	⊞

- For reporting and analysis Purposes
- We want to know more about our users
- This includes affiliation of research projects
- Good existing classification from 'Deutsche Forschungsgesellschaft'

Selection of DFG Subject Area upon Project Creation

Please Specify where your project is mainly located according to the DFG Fachsystematik:

DFG Discipline*

DFG Review Board*

DFG Subject Area*

End of the project*

06.02.2023

How long you need the storage. You will be able to request an extension

Capacity*

Expected storage capacity in TB (1 TB = 1000 GB)

Directory name*

Selection of DFG Discipline

Please Specify where your project is mainly located according to the DFG Fachsystematik:

Category*

Humanities and Social Sciences (1)
Life Sciences (2)
Natural Sciences (3)
Engineering Sciences (4)

Field*

End of the project*

15.02.2023

How long you need the storage. You will be able to request an extension

Capacity*

12

Expected storage capacity in TB (1 TB = 1000 GB)

Selection of DFG Subject Area without Board

Selection of DFG Review Board

Please Specify where your project is mainly located according to the DFG Fachsystematik:

Category*

Engineering Sciences (4)

Board*

Field*

End of the project*

15.02.2023

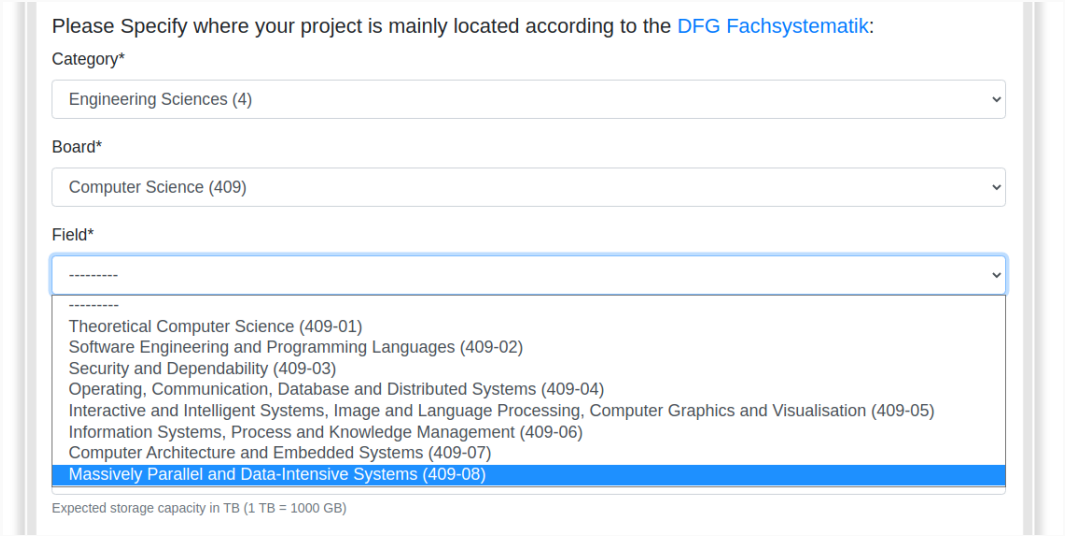
How long you need the storage. You will be able to request an extension

Capacity*

12

Expected storage capacity in TB (1 TB = 1000 GB)

Selection of DFG Subject Area



Website: Requesting Fields for Boards

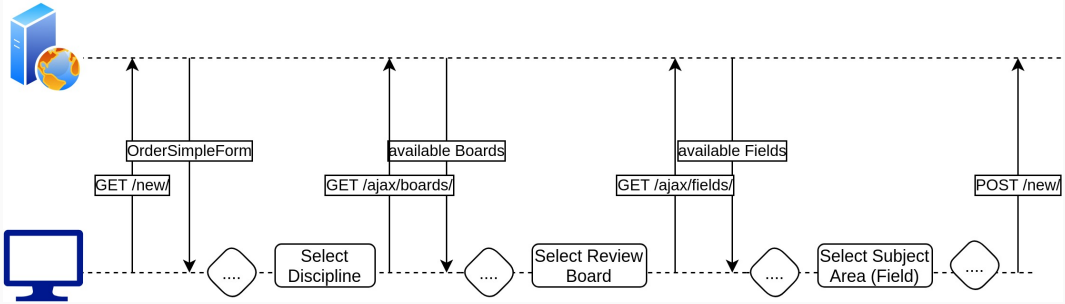
```
1  $("#id_board").change(function () {
2    /* on change of `review board`, request new data for `subject area` field */
3    var url = $("#form").attr("data-fields-url");  // <host>/ajax/fields/
4    var board = $(this).val();    // Which board got selected
5
6    $.ajax({  // request available fields based on selected board
7      url: url,   // resolved to <host>/ajax/fields/
8      data: { 'board': board },   // send board as part of request
9      success: function (data, textStatus, jqXHR) {
10       $("#id_field").html(data);    // fill in available fields
11     }
12   });
13 });
```

Backend: answering with available Fields

```
## urls.py
path('ajax/fields/', views.view_science_fields, name="ajax_load_fields"),

## views.py
def view_science_fields(request):
    b_pk = request.GET.get('board')  # we get the pk of the selected board
    if b_pk:  # select available Fields from this Board
        fields = Science_Field.objects.filter(board__pk=b_pk)
        return render(request, 'dropdown_list_options.html',
                      {'options': fields}
                     )
    return render(request, 'dropdown_list_options.html',
                  {'options': Science_Field.objects.none()}
                 )
```

Visualisation



Project Results

- Proper Logging
- Automatically Generate Documentation
- What Research is the Storage used for?
- **Diff CSV to DFG Schema in Database**
- Attempt to properly modularize PersonForm
- Extension Requests

The DFG schema changes frequently

- The DFG schema changes every four years
- It was last changed in 2020
- So it'll change again in two years
- Not clear how much (probably not a whole lot)

So I implemented a command to compare any csv to what is currently in the database: `manage.py dfg_schema_diff`

Usage of dfg_schema_diff

```
usage: manage.py dfg_schema_diff [-h] [--locale LOCALE] [--columns COLUMNS]
                                 ...
                                 FILE

Show difference from given file schema to DFG schema in database. By default,
ignores the now deprecated hierarchy level 1. ...

positional arguments:
  FILE                  Path to dfg_systematic.csv

options:
  -h, --help            show this help message and exit
  --locale LOCALE       Set the language of the name column to select. Can correctly select
                        both '<locale>' and 'prefLabel@<locale>' columns. (Default: 'en')
  --columns COLUMNS     Dictionary mapping columns (numbers) to expected values "level" (in
                        the hierarchy, category: 0, deprecated/ignored: 1, board: 2, field:
                        3), "notation" (e.g. 101-27), and locale translations, e.g. "en"
                        (double quotes are important!). Defaults to auto.
  ...                   ...
```

Project Results

- Proper Logging
- Automatically Generate Documentation
- What Research is the Storage used for?
- Diff CSV to DFG Schema in Database
- **Attempt to properly modularize PersonForm**
- Extension Requests

```
class OrderSimpleForm(forms.ModelForm):
    p1_email = forms.EmailField(max_length=100, label='Email')
    p1_institute = forms.CharField(max_length=300, required=False, label='Institute')
    p1_organization = forms.CharField(max_length=300, required=False, label='Organization')
    p1_firstname = forms.CharField(max_length=100, required=True, label='Firstname')
    p1_lastname = forms.CharField(max_length=100, required=True, label='Lastname')
    p1_roles = forms.MultipleChoiceField(widget=CheckboxSelectMultipleWithDisabledOption, label='Role
                                         choices=(
                                             ("ROLE_HEAD", "Head of the project"),
                                             ("ROLE_TECH", "Technical contact"),
                                         ))

    p2_email = forms.EmailField(max_length=100, required=False, label='Email')
    p2_institute = forms.CharField(max_length=300, required=False, label='Institute')
    p2_organization = forms.CharField(max_length=300, required=False, label='Organization')
    p2_firstname = forms.CharField(max_length=100, required=False, label='Firstname')
    p2_lastname = forms.CharField(max_length=100, required=False, label='Lastname')
    p2_roles = forms.MultipleChoiceField(widget=forms.CheckboxSelectMultiple, required=False, label=
                                         choices=(
                                             ("ROLE_HEAD", "Head of the project"),
                                             ("ROLE_TECH", "Technical contact"),
                                         ))
    ...                                  ...
```

Intermediate Results

```
1   class PersonForm(forms.ModelForm):
2       roles = forms.MultipleChoiceField(
3           label="Roles", required=False,
4           choices=(
5               ("ROLE_HEAD", "Head of the project"),
6               ("ROLE_TECH", "Technical contact"),
7           ))
8       class Meta:
9           model = Person
10          fields = [ "first_name", "last_name", "email",
11              "institute", "roles", "organization",
12              ]
13
14  PersonFormSet = formset_factory(PersonForm)
15
16  class OrderSimpleForm(forms.ModelForm):
17      """ Form for Project requests. """
18      owner = OwnerForm()  # main contact responsible
19      additional_contacts = PersonFormSet({
20          'form-TOTAL_FORMS': '0',
21          'form-INITIAL_FORMS': '0',
22      })
```

- Abstraction to one dedicated PersonForm
- PersonFormSet can have arbitrarily many Persons (was only five)
- Simplifies implementation in View
- Difficult to propagate ValidationErrors properly
- Would have deleted about 300 lines of boilerplate

Reasons for Failure

- Nesting of `Forms` is **not** supported (`owner` within `OrderSimpleForm`)
- Particularly with the many assumptions and automated parts of `ModelForms`
- Manually register fields for `ValidationErrors`
- Manually put in values for Validation (from `PersonFormSet`)
- Manually take values out to save new Models

Nothing impossible, but requires a lot of intricate details to get right. We decided to instead implement something else.

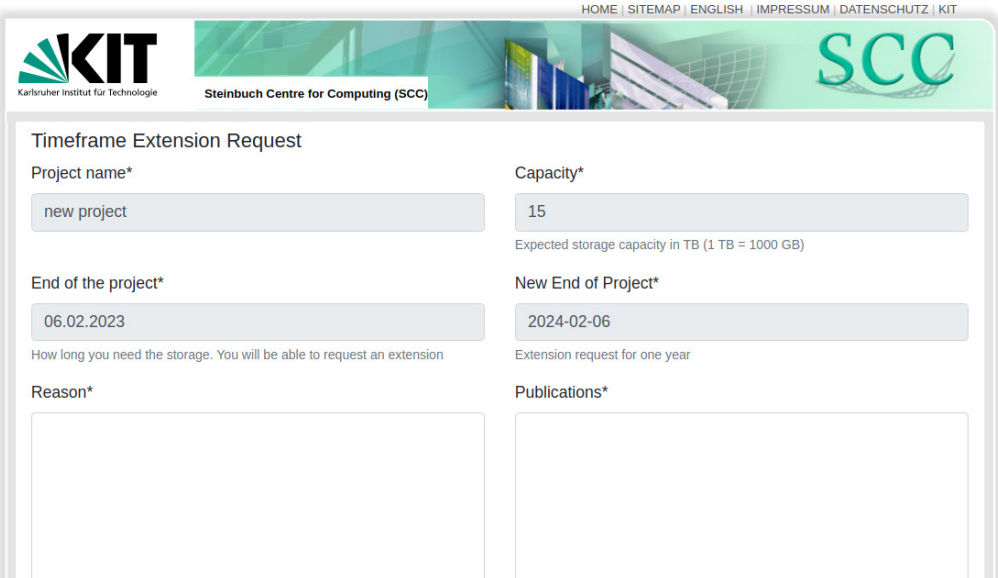
Project Results

- Proper Logging
- Automatically Generate Documentation
- What Research is the Storage used for?
- Diff CSV to DFG Schema in Database
- Attempt to properly modularize PersonForm
- **Extension Requests**

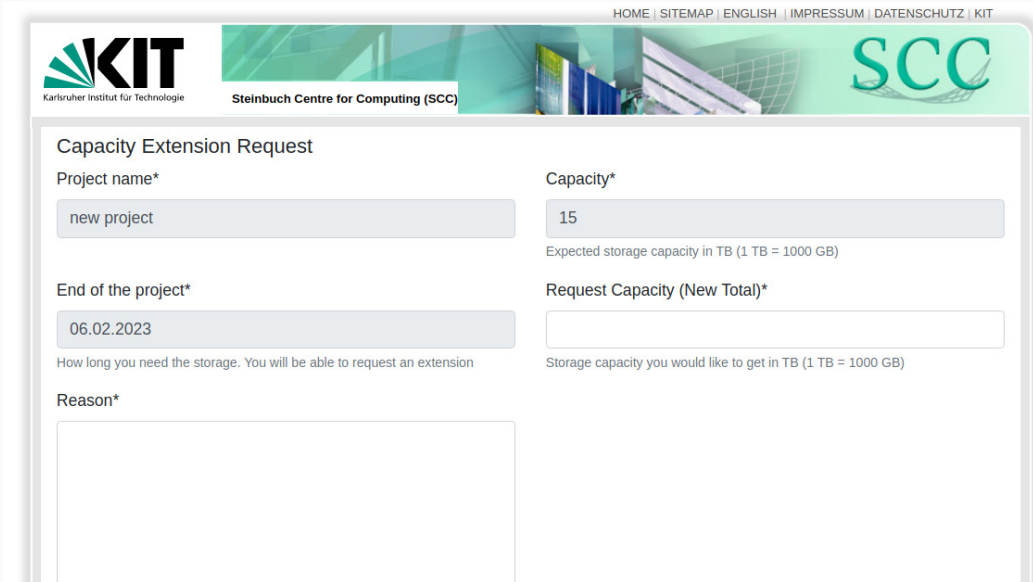
Motivation

It happens frequently that a project needs more space than initially requested, or isn't done by the time their timeframe runs out (initial timeframe was 4 Years). In these cases, timeframe and capacity can be extended manually by an admin. To automate/improve this process, we implemented it directly.

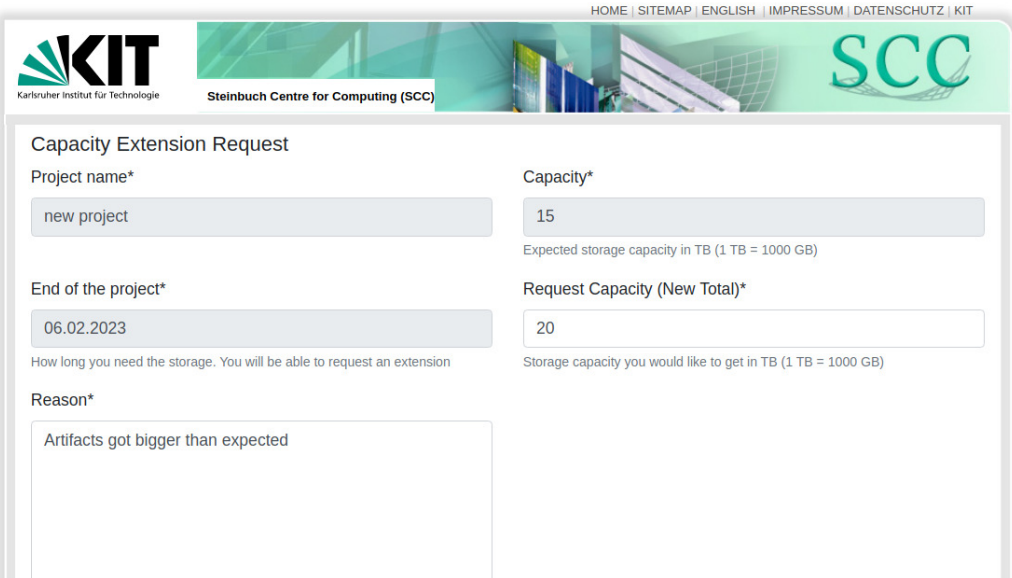
Timeframe Extension Request View



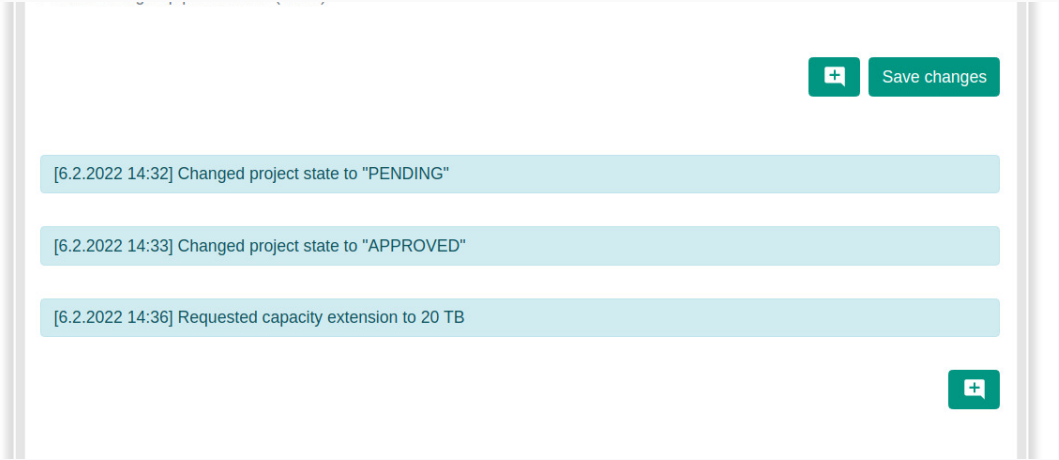
Capacity Extension Request View



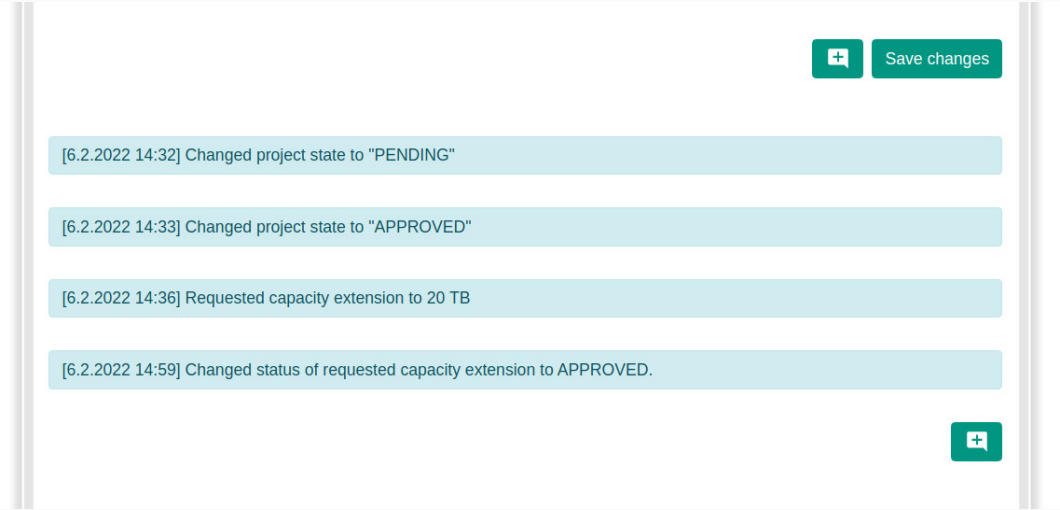
Filled out Capacity Extension



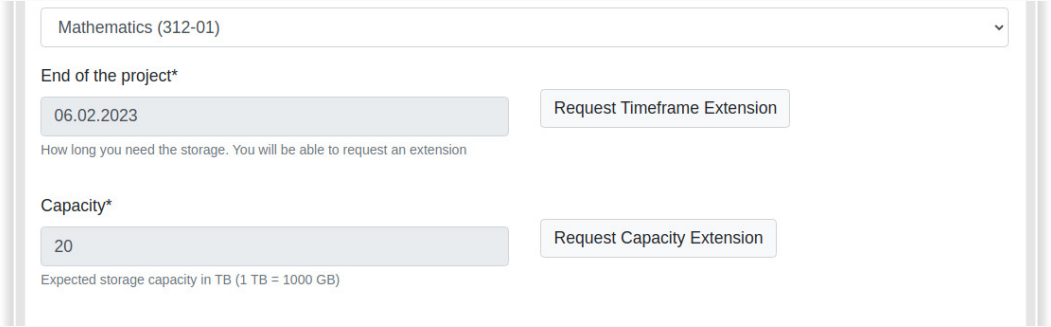
Extension System Message



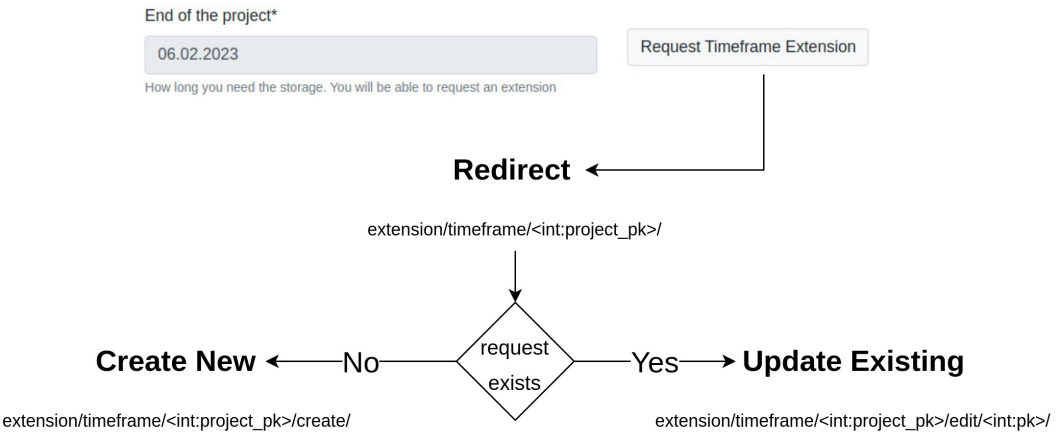
System Message showing Extension Approval



Increased Capacity from User View



Routing for Timeframe Requests



Review

We made substantial Progress!

Next Steps

There's a bunch of stuff still left to do:

- Deployment of the newly implemented features
- Adapt design to current KIT corporate design
- Finish dynamic `PersonForms` implementation
- Login with non-unique identifiers (EPPN)
- Advanced User Administration of Accounts
- Realizing Projects through Storage API-Access
- ...

Sources i

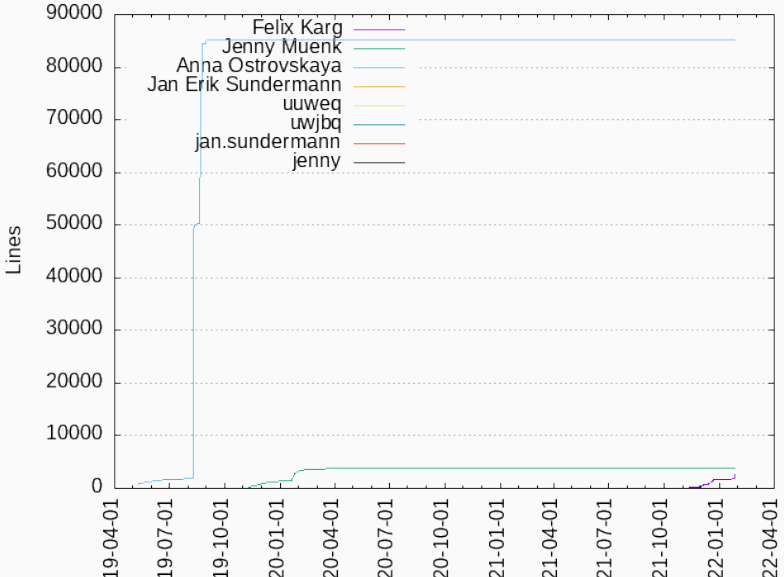
- "Most popular backend frameworks – 2012/2019." https://statisticsanddata.org/data/most-popular-backend-frameworks-2012-2021/, 2019. **accessed 2022-02-07.**
- "Devopedia, "django"." https://devopedia.org/django, 2020. **accessed 2022-02-07.**

End

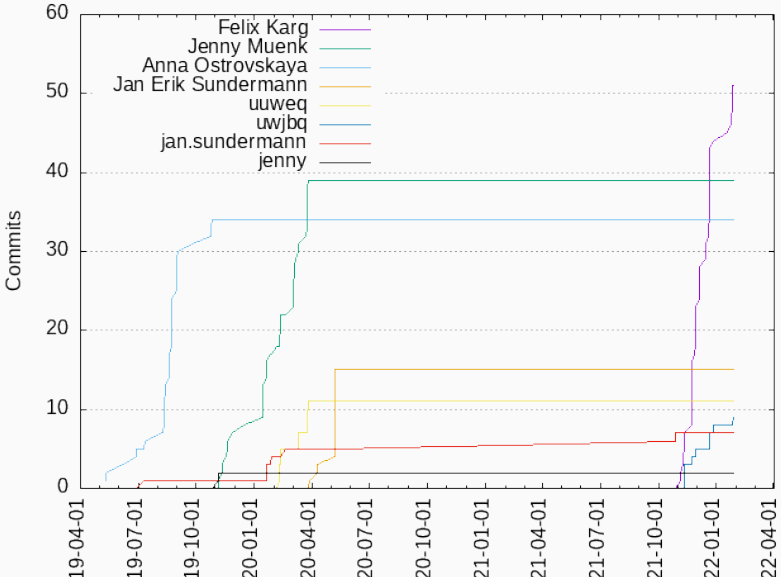
Statistics

More Screenshots

LOC by Author



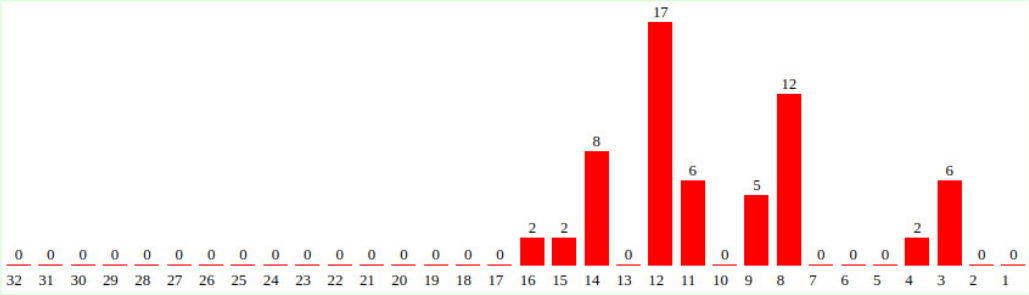
Commits by Author



Recent Activity (Commits by Week)

Weekly activity

Last 32 weeks



LOC stats: Before

Language	Files	Lines	Code	Comments	Blanks
CSS	28	19144	17235	284	1625
JavaScript	27	63499	43096	13725	6678
Python	34	1812	1510	56	246
SVG	14	14	14	0	0
HTML	10	938	838	30	70
—- CSS	1	312	282	28	2
—- JavaScript	5	155	126	13	16
Total	122	86008	62693	14627	8688

(The numbers don't quite add up, because I omitted e.g. Configs, Markdown, ...)

LOC stats: After

Language	Files	Lines	Code	Comments	Blanks
CSS	28	19162	17248	284	1630
JavaScript	27	63499	43096	13725	6678
Python	48	3078	2485	160	434
SVG	14	14	14	0	0
HTML	15	1231	1107	39	85
—- CSS	1	312	282	28	2
—- JavaScript	7	244	202	21	21
Total	134	87055	63960	14245	8850

(The numbers don't quite add up, because I omitted e.g. Configs, Markdown, ...)

LOC stats: Diff

Language	Files	Lines	Code	Comments	Blanks
CSS	0	+18	+13	0	+5
JavaScript	0	0	0	0	0
Python	+14	+1266	+975	+104	+188
SVG	0	0	0	0	0
HTML	+5	+293	+269	+9	+15
—- CSS	0	0	0	0	0
—- JavaScript	+2	+89	+76	+8	+5
Total	+12	+1047	+1267	+382	+162

(The numbers don't quite add up, because I omitted e.g. Configs, Markdown, ...)

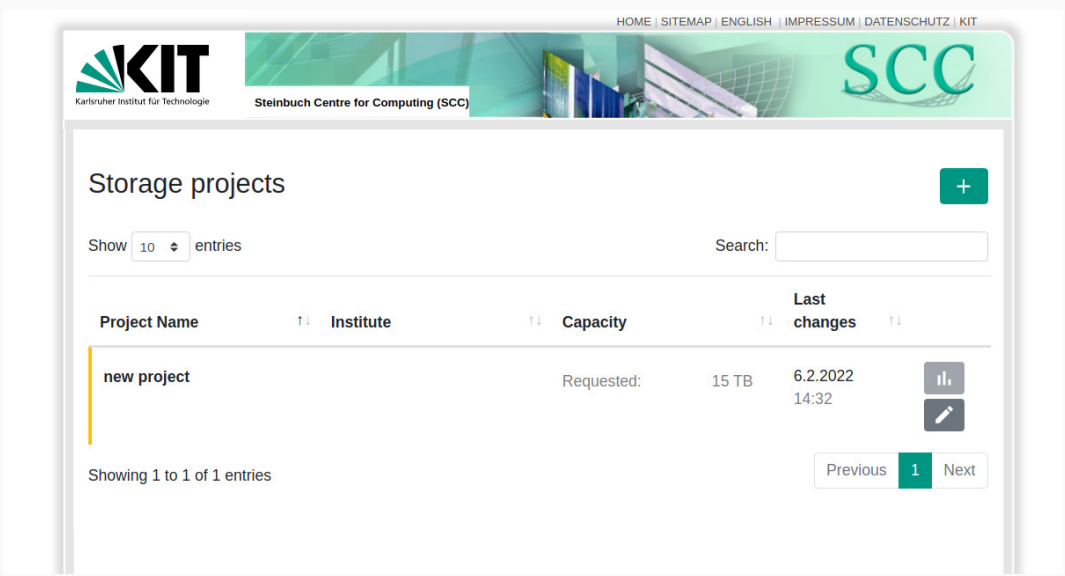
More Screenshots

More Screenshots

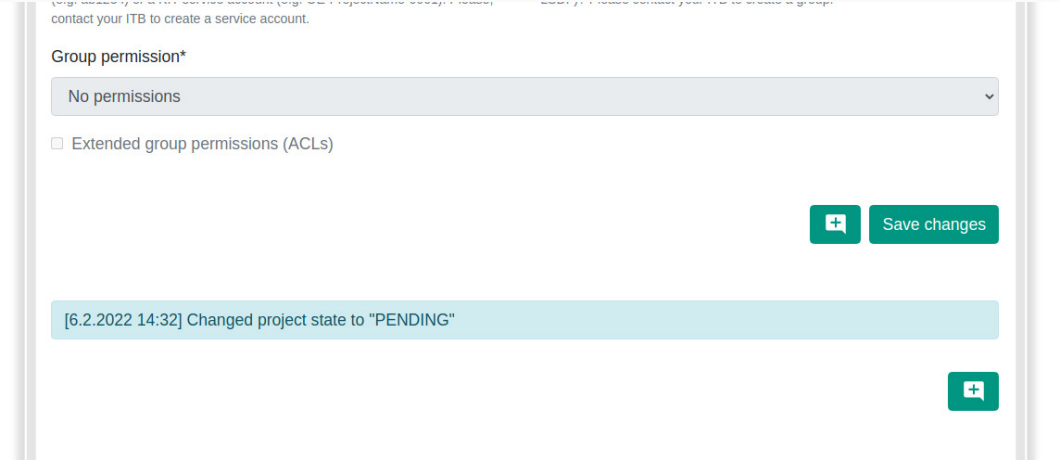
Project Approval

Requesting More Storage

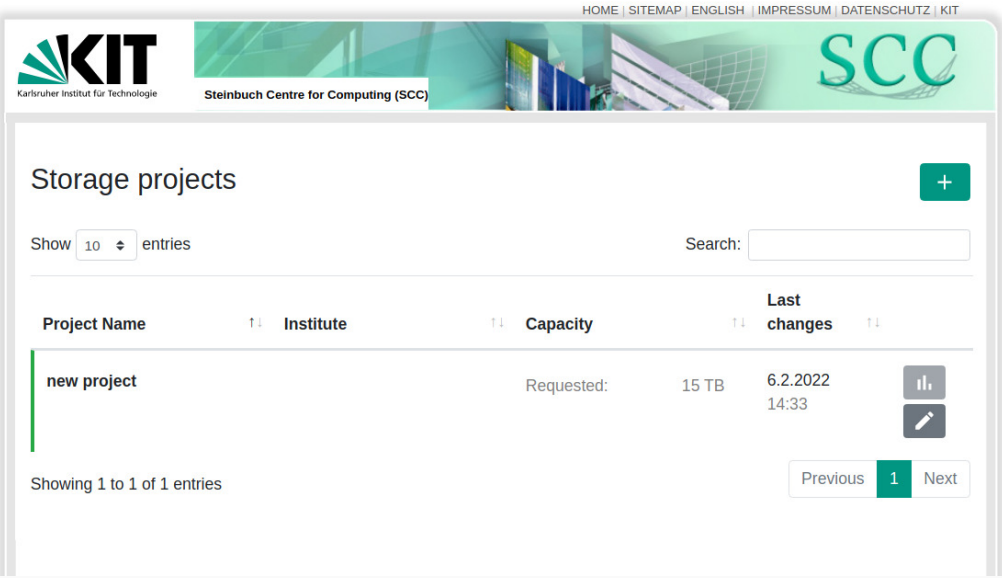
Project Pending



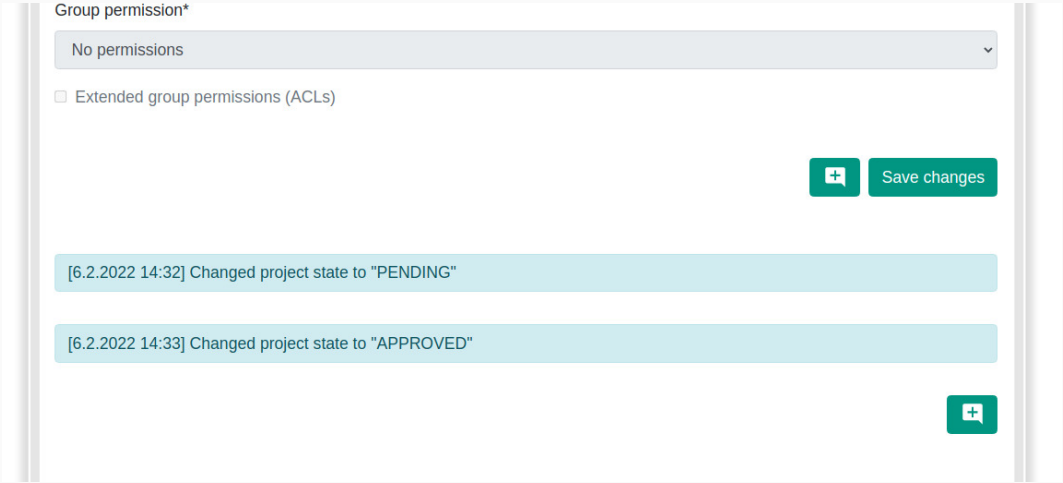
State Change Message



Project Approved



State Change Message

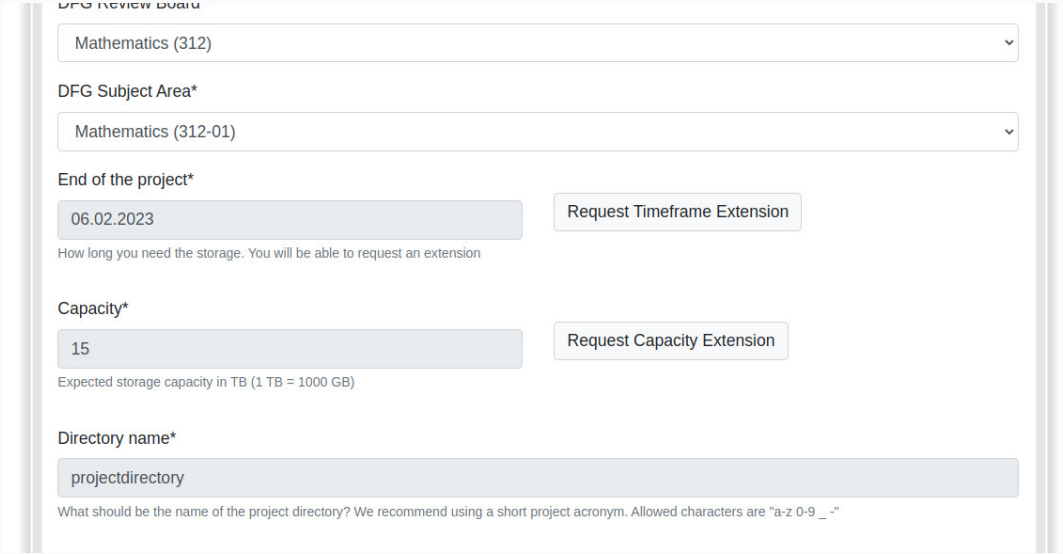


More Screenshots

Project Approval

Requesting More Storage

Extension Requests: The Project Needs to be Approved



Timeframe Extension Request View

Karlsruher Institut für Technologie

Steinbuch Centre for Computing (SCC)

SCC

Timeframe Extension Request

Project name*

new project

Capacity*

15

Expected storage capacity in TB (1 TB = 1000 GB)

End of the project*

06.02.2023

How long you need the storage. You will be able to request an extension

New End of Project*

2024-02-06

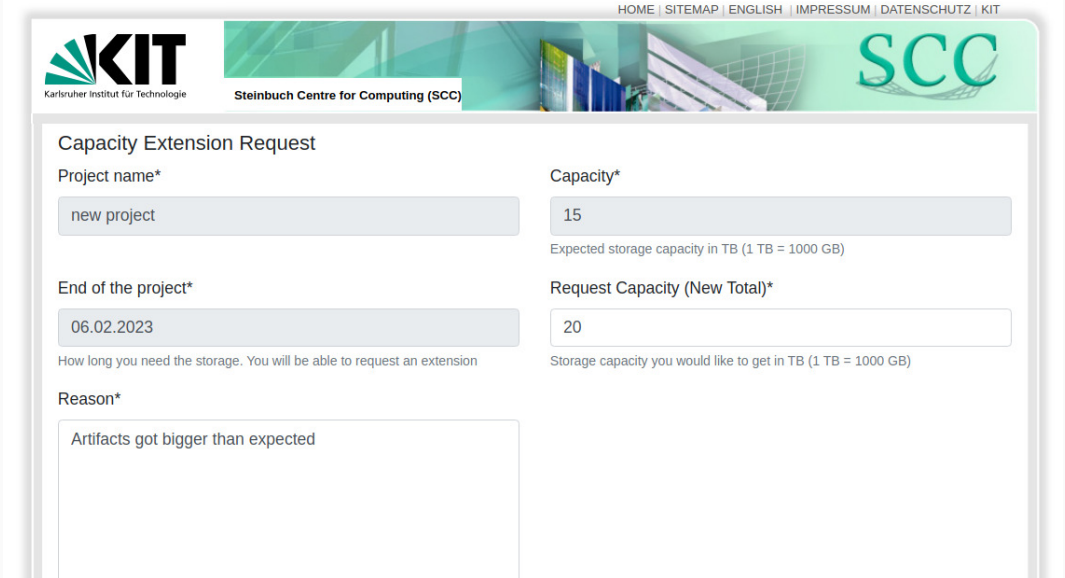
Extension request for one year

Reason*

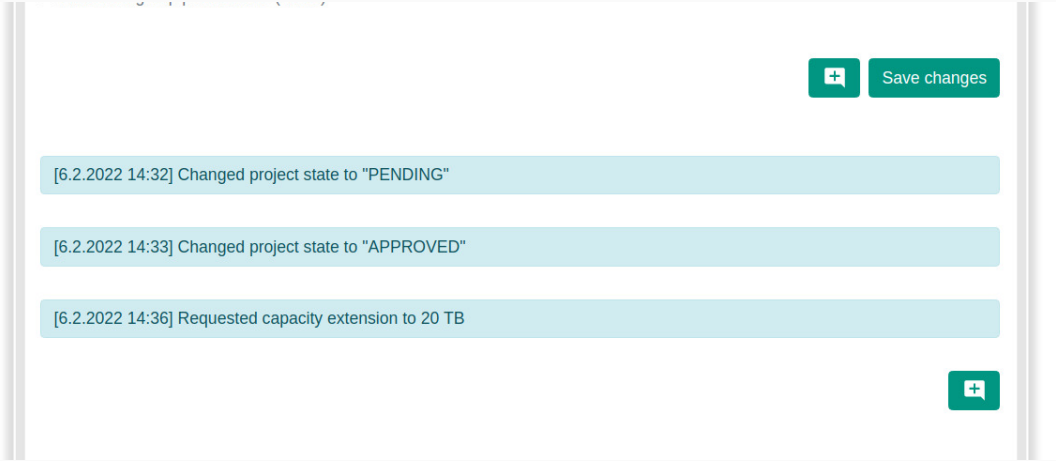
Publications*

Capacity Extension Request View

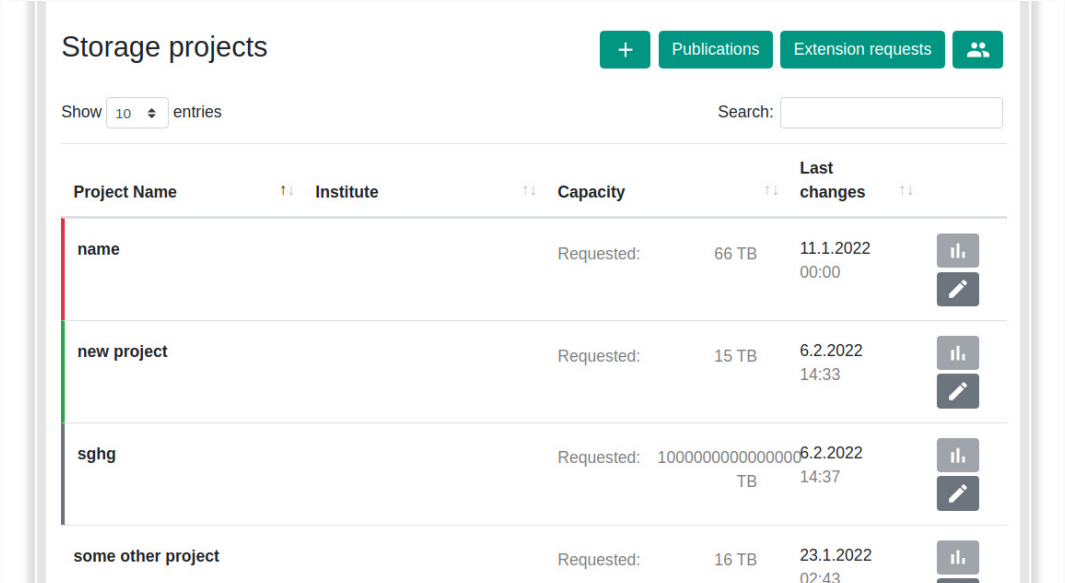
Filled out Capacity Extension



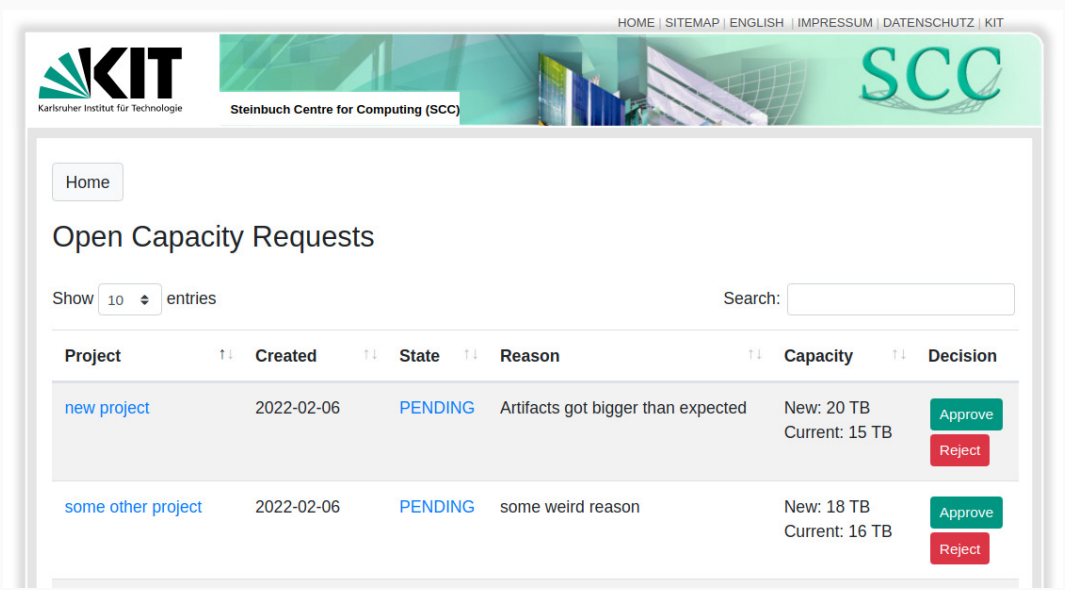
Extension System Message



Projects Admin View

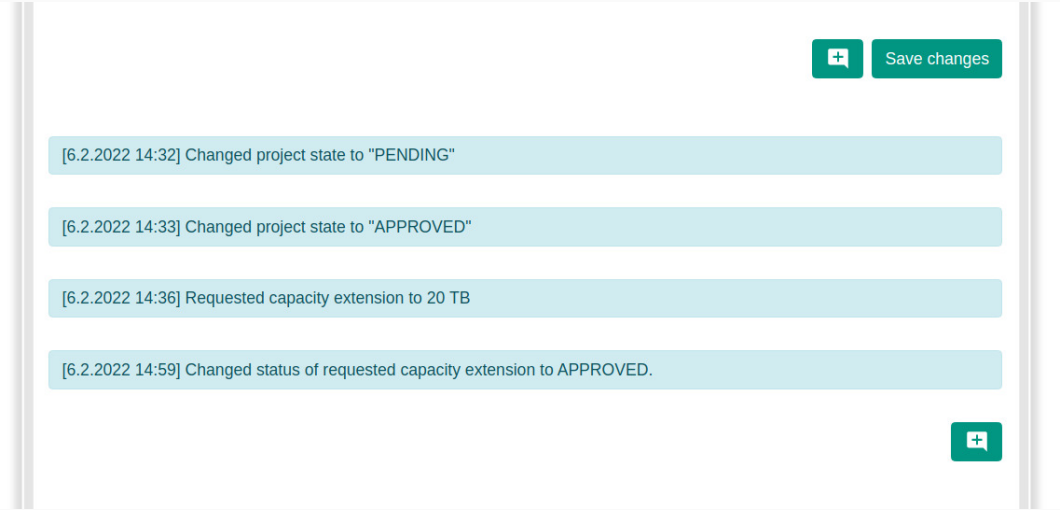


Extension Request Overview



Approving Request

System Message showing Extension Approval



Increased Capacity from User View