# DSpace

13.6.2024

+ dataquest
simply data

# Content

- Part 1 – Architecture (09:00 - 10:30)
  - Definitions, Demo
  - Architecture changes, FE – BE communication
  - New features, Angular in DSpace
  - DevOps, insights, collaboration
- Part 2 – Hands on (11:00 - 12:30)
  - Error case study
  - Import
  - Coding example (tool to show version info)
  - Deployment

# Part I – Architecture

Demo, FE – BE communication, New features, Angular, DevOps, insights, collaboration

All very technical – these slides will be available after

# A few definitions

- Important links in the last slide
- clarin-dspace v5 – DSpace currently used by CLARIN centres
- **clarin-dspace v7 – our implementation**
- Vanilla DSpace v7 – DSpace without modifications
  - init. est. 2017/2018, 7.0 released 2021, stable **2023**
- clarin-dspace v7 wiki – GitHub Wiki on dataquest DSpace repository
- DSpace Documentation – Lyrasis Vanilla DSpace documentation

# Demo – the new DSpace is familiar

- Home, Search, Item, Full-item – similar to v5
- Login, create community, collection, Item
- Some admin features – Health, License administration
- Authorization pages, bulk access edit
- What is new?
  - S3 + checksums
  - Live config reload
  - OpenAIRE complex input field
  - Item versions
  - DOI can be shown

# DSpace architecture changes

- XMLUI/JSPUI to Angular
- Java REST API
- Docker,
- Entity model
- Solr, Postgres – the same

+ dataquest
simply data

# New features

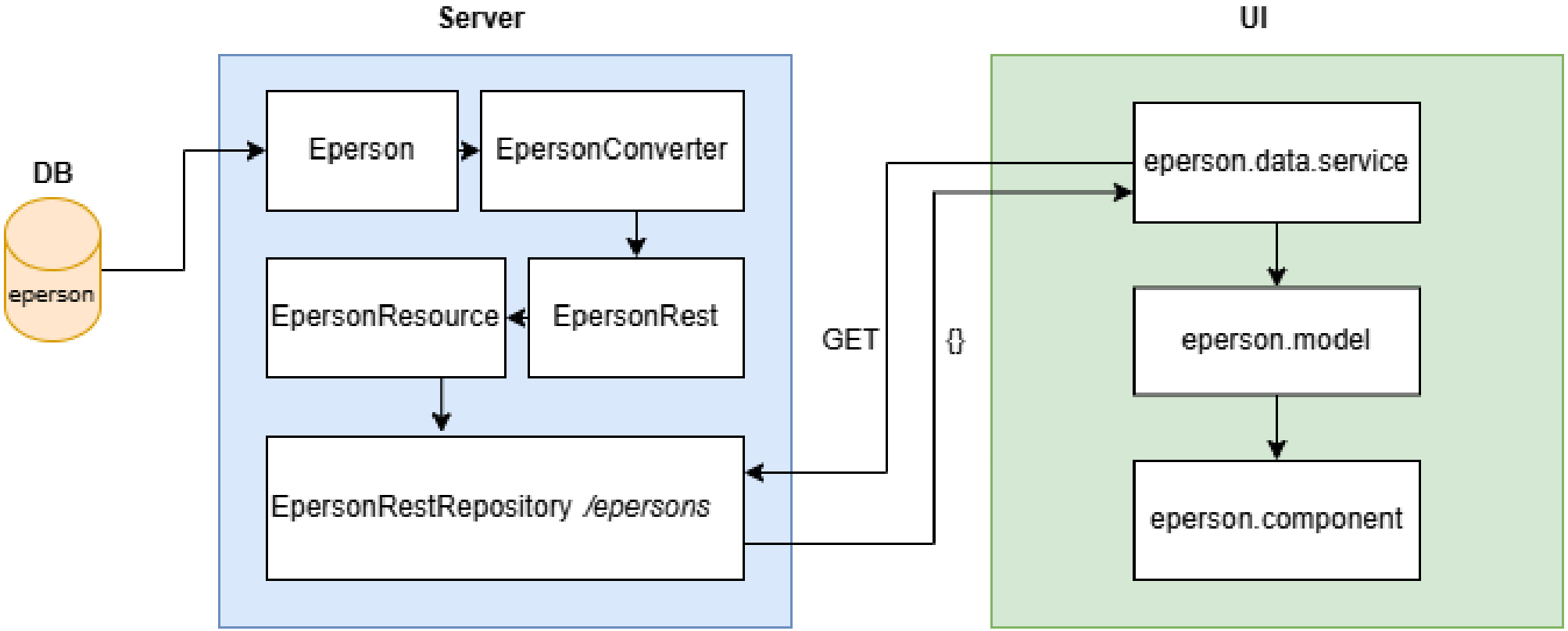
- S3
- Configurable entities
- ORCID
- Docker
- DOI
- Admin "Health" menu
- OpenAIRE API
- OpenAIREv4

- Better support for IIIF
- The Submission Form configuration has changed
- Video/Image Content Streaming
- A completely new, fully featured REST API
- Support for Google Captcha

# Angular UI

- Theming (dspace/custom)
- Translations
- Static pages
- Accessibility - WCAG level AA
- Config

# Angular – BE communication

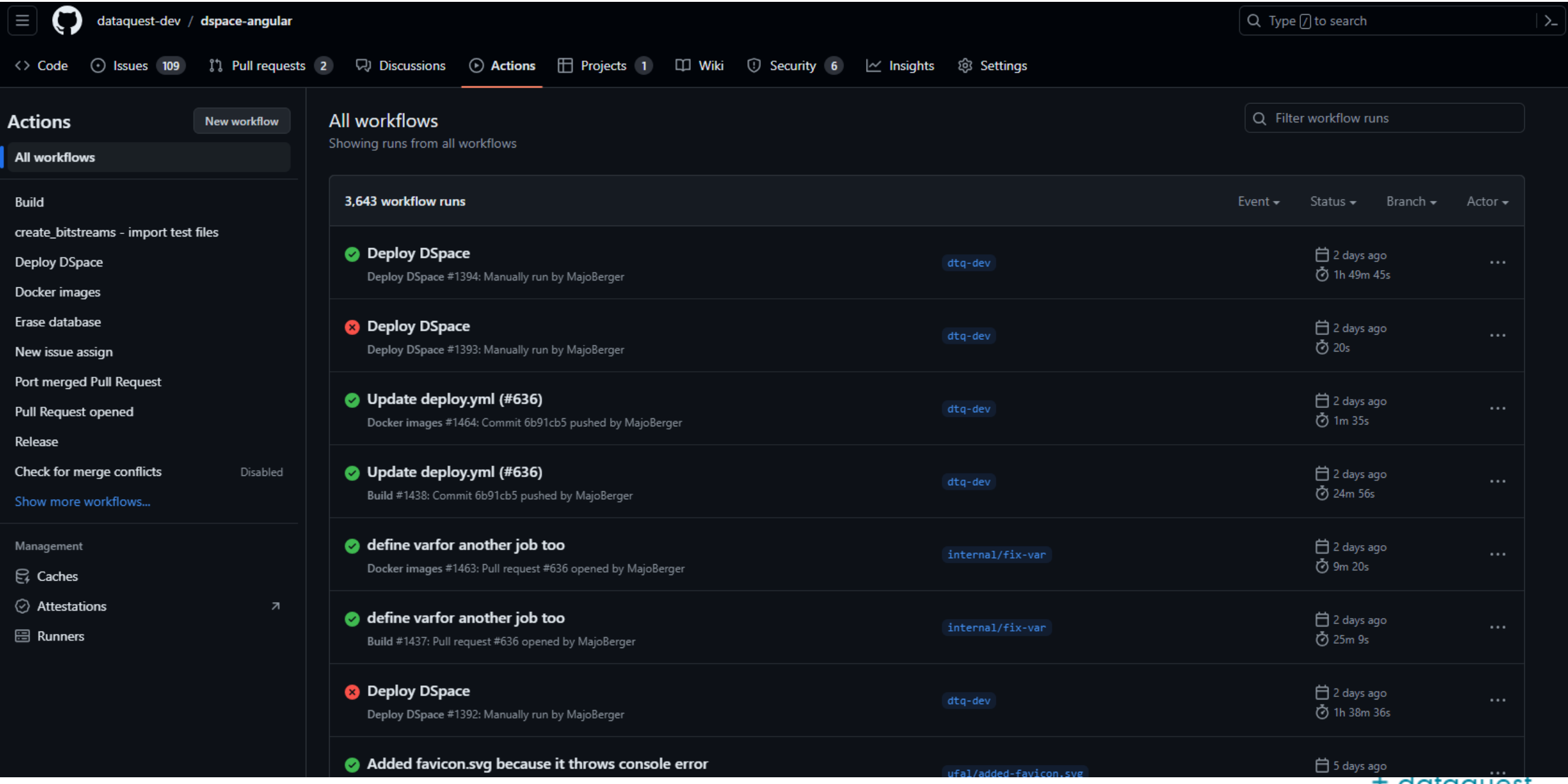


# Rendered configurable submission forms

- `submission-forms.xml`
- Configuration
- Dynamic components to render the configuration

# DevOps – our support for developers

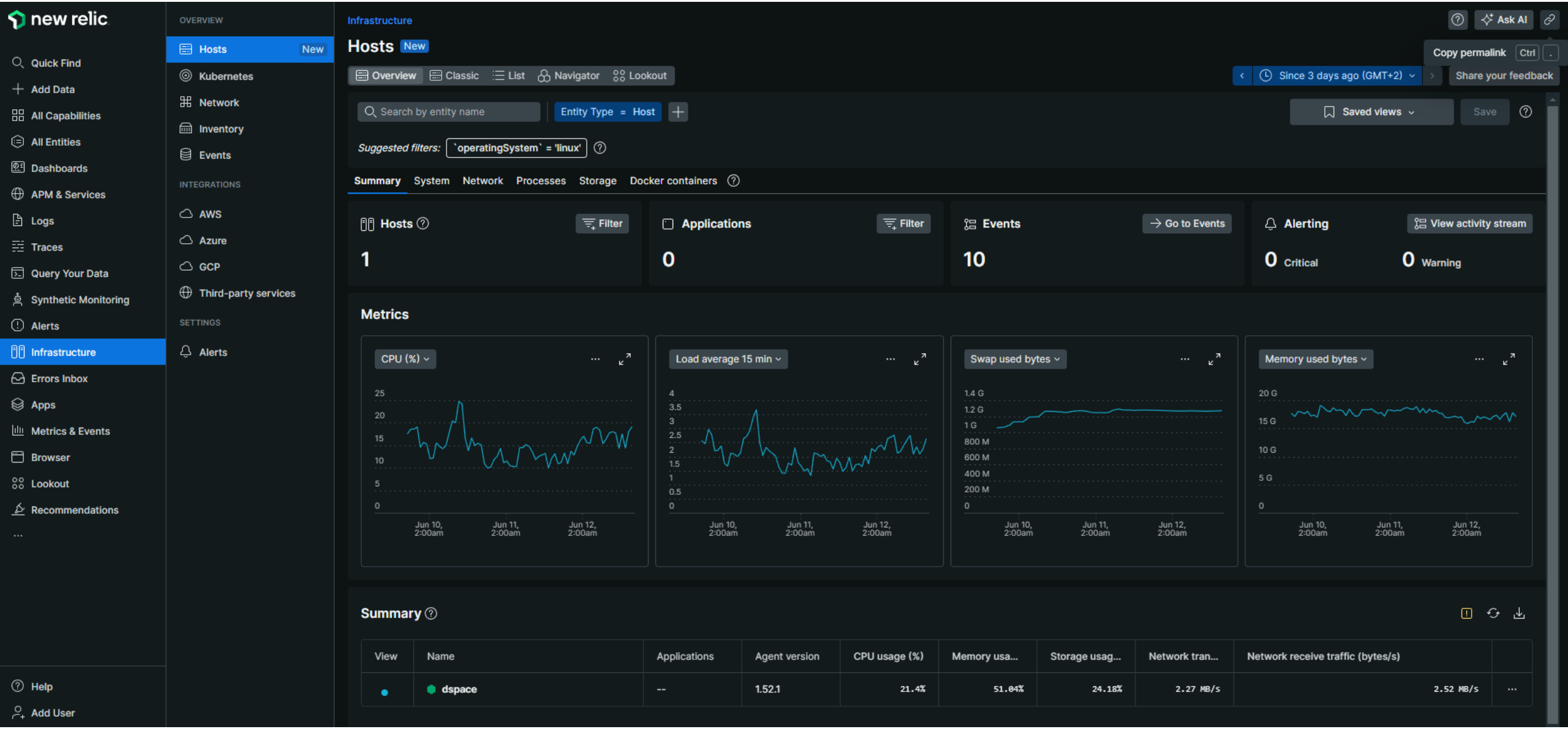
- GitHub actions
  - Deploy to our local dev machines
  - Several Testing (automatic) workflows – unit, integration, UI, BE
  - Build and push
  - Import
  - Tagging for release
- Several instances – e.g., different flavours or setups
- Remote debug
- Use remote database




+ dataquest
simply data

# Production insights

- Logs
  - FE `pm2 monit` or `pm2 logs`
  - BE `/dspace/log/` (`dspace.log` and `warn.log`)
  - Solr `docker dspacesolr0 logs`
  - Nginx – on host, usually `/var/logs/nginx/`
- Monitoring tool – custom warnings and checks
- Configs – most in code, some mounted – enables hot reload
- Errors – in HANDS-ON section – how to find them
- Performance
  - FE needs a lot of resources – CPUs (cores) and RAM (mainly due to SSR)




+ dataquest
simply data

YOUR LOGS

- All logs
- Attributes
- Patterns

MANAGE DATA

- Drop filters
- Parsing
- Data partitions
- Obfuscation
- Lookup tables

Logs

# Attributes

Ask AI | Since 3 days ago (GMT+2) | live tail

Partition (1) | Search for Logs using Lucene | Query logs | NRQL | Saved views

## Attributes

Search 46 attributes

| RECENT | VALUES |
|---|---|
| filePath | 9 |
| hostname | 1 |
| httpversion | 3 |
| logtype | 6 |
| referrer | 23.6k |

APPLICATION ATTRIBUTES

| | |
|---|---|
| hostname | 1 |

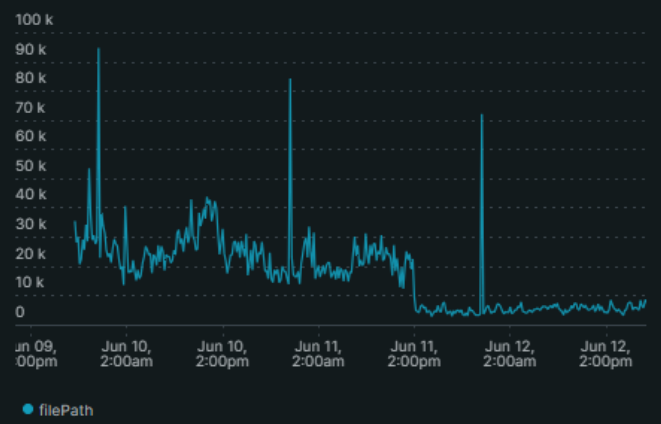
ENTITY

| | |
|---|---|
| guid.INFRA | 1 |
| guids | 1 |

## Values

Search 9 values

| FILEPATH | LOGS |
|---|---|
| /var/log/nginx/access.log | 4.5m |
| /var/lib/docker/volumes/tul-dspace_dspace_logs/_data/warn.log | 1.7m |
| /var/log/syslog | 23.1k |
| /var/lib/docker/containers/66a7c50b524c7e17910f7f5e0647f7c2b950...921 | |
| /var/log/auth.log | 529 |
| /var/log/nginx/error.log | 336 |
| /var/log/dpkg.log | 63 |
| /var/lib/docker/containers/b818d1e8015f7bf4f5b4f478f7c40ac751a8dfc... 6 | |
| /var/log/alternatives.log | 1 |




Expand logs | Expand table | Add column | Add to dashboard | Export | Manage parsing rules

| message | filePath |
|---|---|
| paceApiExceptionControllerAdvice @ Authentication is required (status:401 exception: Access is denied at: org.springframework.security.access.vote.AffirmativeBased.decide(AffirmativeBased.java:73)) | |
| 2024-06-12 19:21:51,722 WARN 351a4a2f-db6a-482e-be02-d01f5c3a7e76 212caa22-337f-4d2b-853f-97cb0e58120e org.dspace.app.rest.exception.DSpaceApiExceptionControllerAdvice @ Authentication is required (status:401 exception: Access is denied at: org.springframework.security.access.vote.AffirmativeBased.decide(AffirmativeBased.java:73)) | /var/lib/docker/volumes/tul-dspace_dspace_logs/_data/warn.log |
| 2024-06-12 19:21:51,730 WARN a7788000-2dbc-45eb-b3b7-f2637b3f360f 738f43b6-7fd7-40de-8f55-90b896c9b73f org.dspace.app.rest.exception.DSpaceApiExceptionControllerAdvice @ Authentication is required (status:401 exception: Access is denied at: org.springframework.security.access.vote.AffirmativeBased.decide(AffirmativeBased.java:73)) | /var/lib/docker/volumes/tul-dspace_dspace_logs/_data/warn.log |
| 2024-06-12 19:21:51,730 WARN a7788000-2dbc-45eb-b3b7-f2637b3f360f f0283559-e64d-4f25-9e50-5f9ba87d1832 org.dspace.app.rest.exception.DSpaceApiExceptionControllerAdvice @ Authentication is required (status:401 exception: Access is denied at: org.springframework.security.access.vote.AffirmativeBased.decide(AffirmativeBased.java:73)) | /var/lib/docker/volumes/tul-dspace_dspace_logs/_data/warn.log |
| 2024-06-12 19:21:51,740 WARN 351a4a2f-db6a-482e-be02-d01f5c3a7e76 09b86771-6167-43d8-93e3-0c9ceb681107 org.dspace.app.rest.utils.Utils @ Tried fetching resource: version for DSpaceObject with ID: 84a2742b-5290-41c9-8d09-fec819898193 | /var/lib/docker/volumes/tul-dspace_dspace_logs/_data/warn.log |
| 2024-06-12 19:21:51,858 WARN 351a4a2f-db6a-482e-be02-d01f5c3a7e76 a95ab665-5bd7-4725-92e0-0af32bc48869 org.dspace.app.rest.exception.DSpaceApiExceptionControllerAdvice @ Authentication is required (status:401 exception: Access is denied at: org.springframework.security.access.vote.AffirmativeBased.decide(AffirmativeBased.java:73)) | /var/lib/docker/volumes/tul-dspace_dspace_logs/_data/warn.log |

simply data

# Experiences from development and deploy

- FE high load, caching, debugging, a lot of logs (nginx)
- Tests from upstream
- pm2 setup, monitoring, config, debug
- Logging too much
- Many requests in FE-BE communication
- Solr searching slightly modified compared to v5
- Nontrivial shibboleth integration – not DSpace, but can cause some issues
- Hardware requirements – much more than v5 or v6 (xmlui)

# Flow of development and collaboration

- Main repository is ufal – will be synchronized shortly after this presentation
- Contributing is easy
  - Issue
  - PR (to ufal repository)
  - We will do review and merge
- Possibly release – notes, tags, images
- Behind the scenes  - for robustness and safety
  - Automatic testing, docker build and possibly deploy
  - Tests
    - Unit tests (BE and FE)
    - Integration tests – from Vanilla
    - Integration tests – REST API, playwright
    - test scenarios

Releases Tags

Draft a new release

Find a release

Mar 21

milanmajchrak

drruk-2024.03....

08956f1

Compare

## drruk-2024.03.8360397154 Latest

### What's Changed

- Fix mistakes in the /home page
- Fixed mistakes in the Item View page
- Fixed mistakes in the Search page
- Use `dc.contributor.other` metadata field for mapping the authors
- Added `edm.xml` registry
- Updated shibboleth cfg properties (firstname, lastname, netid, email)

▾ Assets 2

| | |
|---|---|
| Source code (zip) | Mar 20 |
| Source code (tar.gz) | Mar 20 |

Mar 21

milanmajchrak

none-2024.03....

f08b0eb

## CLARIN-DSpace 7.6.1 - 2024.03

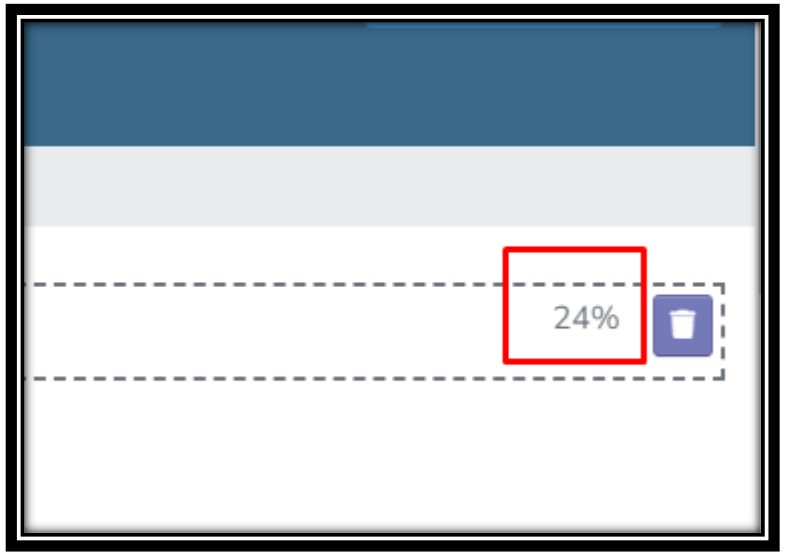
### What's Changed

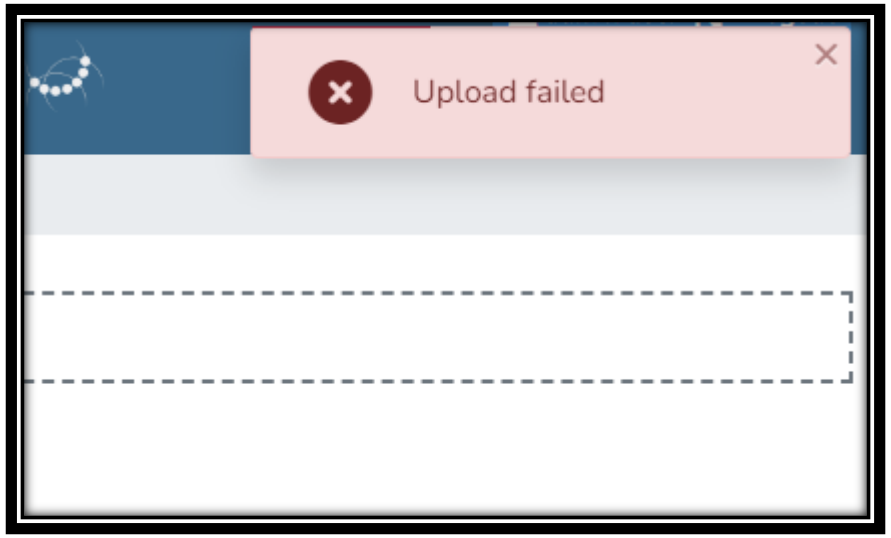
# Part II – Hands on

Error tracing, Import, BE + FE coding, deployment

+ dataquest
simply data

# Error case study – big file upload




+ dataquest
simply data

# Error case study – big file upload

- FE console – nothing
- FE logs in container – nothing
  - `docker exec -it dspace-angular8 sh`
  - `pm2 logs` (add `-n 500` to show more (500) lines)
- BE logs in container
  - `docker exec -it dspace8 bash`
  - `tail /dspace/log/warn.log` (or …/dspace.log)
  - (…) DSpaceApiExceptionControllerAdvice @ Maximum upload size exceeded (status:413 exception: Maximum upload size exceeded (…) the request was rejected because its size (5148576194) exceeds the configured maximum (4294967296) at: (…) .StandardMultipartHttpServletRequest (…)

# Error case study – big file upload

- Config `clarin-dspace.cfg`
  - spring.servlet.multipart.max-file-size = 6GB
  - spring.servlet.multipart.max-request-size = 4GB
- Change the both to 6GB
  - No need to restart dspace due to hot reload?
  - Everything works?
  - IF it didn’t help, try the restart the container
  - `docker restart dspace8`
  - Hot reload does NOT always work. Some settings require restart.
  - Most errors you are expected to encounter will be misconfigurations

# Import - preparation

- Guide also in the links at the end checkout repository
- Checkout repository
  - https://github.com/ufal/dspace-migrate
- Checkout submodule: Python library – client that uses REST API
  - `git submodule update --init libs/dspace-rest-python/`
- Obtain database dump
- TURN OFF MAIL SERVER
- Prepare new empty DSpace installation
  - (`dspace database migrate force` must be run, but it is autorun in docker)
- Create administrator
- Install requirements
  - `pip install –r requirements.txt`

+ dataquest
simply data

# Import - execution

- Run DBs
  - `scripts/start.local.dspace.db.bat`
- Convert dbs to JSONs
  - `python db_to_json.py --database=clarin-dspace`
  - `python db_to_json.py --database=clarin-utilities`
- Update `project_settings.py` (passwords, URLs etc.)
- Start import
  - `python repo_import.py`
- Show logs (prepared, from another run)
  - Describe what is checked

+ dataquest
simply data

# Import - result

- Logs in next slides

```
2024-06-10 16:16:35,143:INFO: Updating [backend.endpoint]->[http://dev-5.pc:88/server/api]
2024-06-10 16:16:35,143:INFO: Updating [db_dspace_7.port]->[5438]
2024-06-10 16:16:35,143:INFO: Updating [db_dspace_5.port]->[15432]
2024-06-10 16:16:35,143:INFO: Updating [db_utilities_5.port]->[15432]
2024-06-10 16:16:35,143:INFO: Updating [input.datadir]->[/opt/dspace-data/clarin-dspace//data/]
2024-06-10 16:16:35,143:INFO: Updating [input.icondir]->[/opt/dspace-data/clarin-dspace//icon/]
2024-06-10 16:16:35,143:INFO: Initialise connection to DSpace REST backend [http://dev-5.pc:88/server/api]
2024-06-10 16:16:35,582:INFO: Authenticated successfully as test@test.edu
2024-06-10 16:16:35,583:INFO: DSpace REST backend is available at [http://dev-5.pc:88/server/api]
2024-06-10 16:16:35,583:INFO: Loading repo objects
2024-06-10 16:16:35,695:INFO: Loaded groups [{'1': ['503c9518-543a-448c-b074-3a4a492e1ad5'], '0': ['27d2b6ea-5daa-4efd-9622-4d4370a6c666']}], other groups:[0]
2024-06-10 16:16:37,038:WARNING: Ignoring metadata fields [[176]], len:[454870->123878]
2024-06-10 16:16:38,468:INFO: New instance database status:
2024-06-10 16:16:38,547:INFO:
                bitstreamformatregistry:       93
                      checksum_results:       10
                          dspaceobject:        5
                           entity_type:        1
                               eperson:        2
                          epersongroup:        2
                    epersongroup2eperson:        1
                           fileextension:      111
                                  handle:        1
                     metadatafieldregistry:      304
                    metadataschemaregistry:       20
                            metadatavalue:        6
                           resourcepolicy:        1
                           schema_version:       67
                                     site:        1
                        user_registration:        2
```

+ dataquest
simply data

```
2024-06-10 16:17:03,829:INFO: =====
2024-06-10 16:17:03,829:INFO: Importing   [   8] communities
2024-06-10 16:17:04,954:INFO: OK Imported [   8] communities
2024-06-10 16:17:04,954:INFO: ========== Validating community ==========
2024-06-10 16:17:04,956:INFO: Table [           community]  OK  compared by len only v5:[8], v7:[8]
2024-06-10 16:17:04,956:INFO: ========== Validating community2community ==========
2024-06-10 16:17:04,957:INFO: Table [ community2community]  OK  compared by len only v5:[5], v7:[5]
2024-06-10 16:17:04,957:INFO:
****************************************

2024-06-10 16:17:04,957:INFO: =====
2024-06-10 16:17:04,957:INFO: Importing   [  13] collections

  0%|          | 0/13 [00:00<?, ?it/s]
100%|██████████| 13/13 [00:04<00:00,  3.22it/s]
2024-06-10 16:17:08,995:INFO: OK Imported [  13] collections
2024-06-10 16:17:08,996:INFO: ========== Validating collection ==========
2024-06-10 16:17:08,998:INFO: Table [          collection]  OK  compared by len only v5:[13], v7:[13]
2024-06-10 16:17:08,998:INFO: Table [          collection]  !!! WARN !!!   NON NULL [logo_bitstream_id] v5:[  3], v7:[  0]
2024-06-10 16:17:08,998:INFO: ========== Validating community2collection ==========
2024-06-10 16:17:08,999:INFO: Table [community2collection]  OK  compared by len only v5:[13], v7:[13]
2024-06-10 16:17:08,999:INFO:
****************************************
```

+ dataquest
simply data

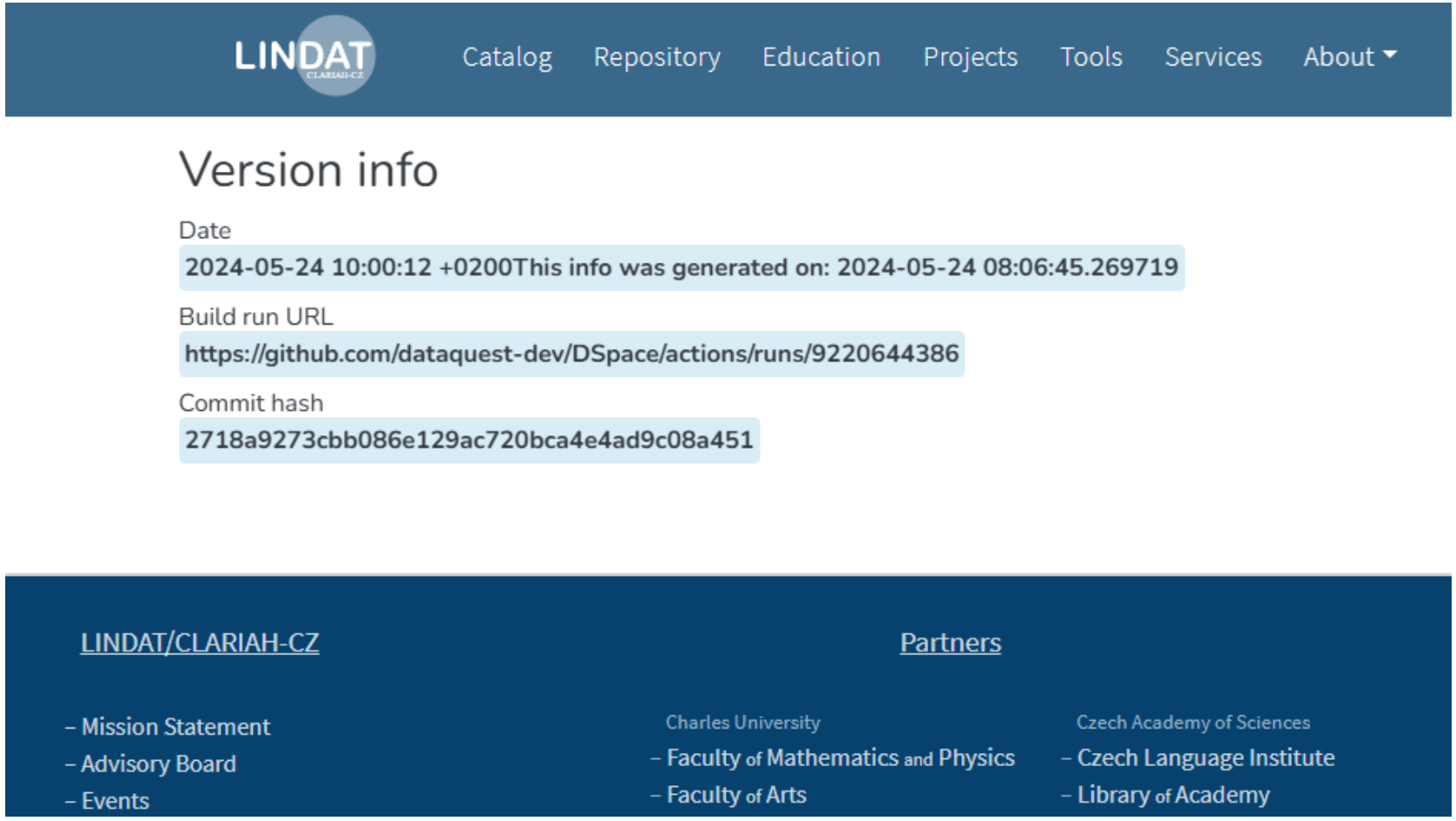
```
2024-06-10 17:39:37,706:INFO: Took [4982.56] seconds to import all data
2024-06-10 17:39:37,706:INFO: Made [293] GET requests, [57647] POST requests.
2024-06-10 17:39:37,706:INFO: New instance database status:
2024-06-10 17:39:37,800:INFO:
                          bitstream:     8175
               bitstreamformatregistry:      103
                             bundle:     3501
                    bundle2bitstream:     8098
                    checksum_results:       10
                          collection:       13
                     collection2item:     2831
                           community:        8
                  community2collection:       13
                   community2community:        5
                    cwf_collectionrole:        4
                          cwf_pooltask:        2
                      cwf_workflowitem:        2
                          dspaceobject:    16373
                           entity_type:        1
                               eperson:     1524
                          epersongroup:       42
                  epersongroup2eperson:       38
                         fileextension:      111
                           group2group:       56
                      group2groupcache:       64
                                handle:     8002
                                  item:     3109
```

+ dataquest
simply data

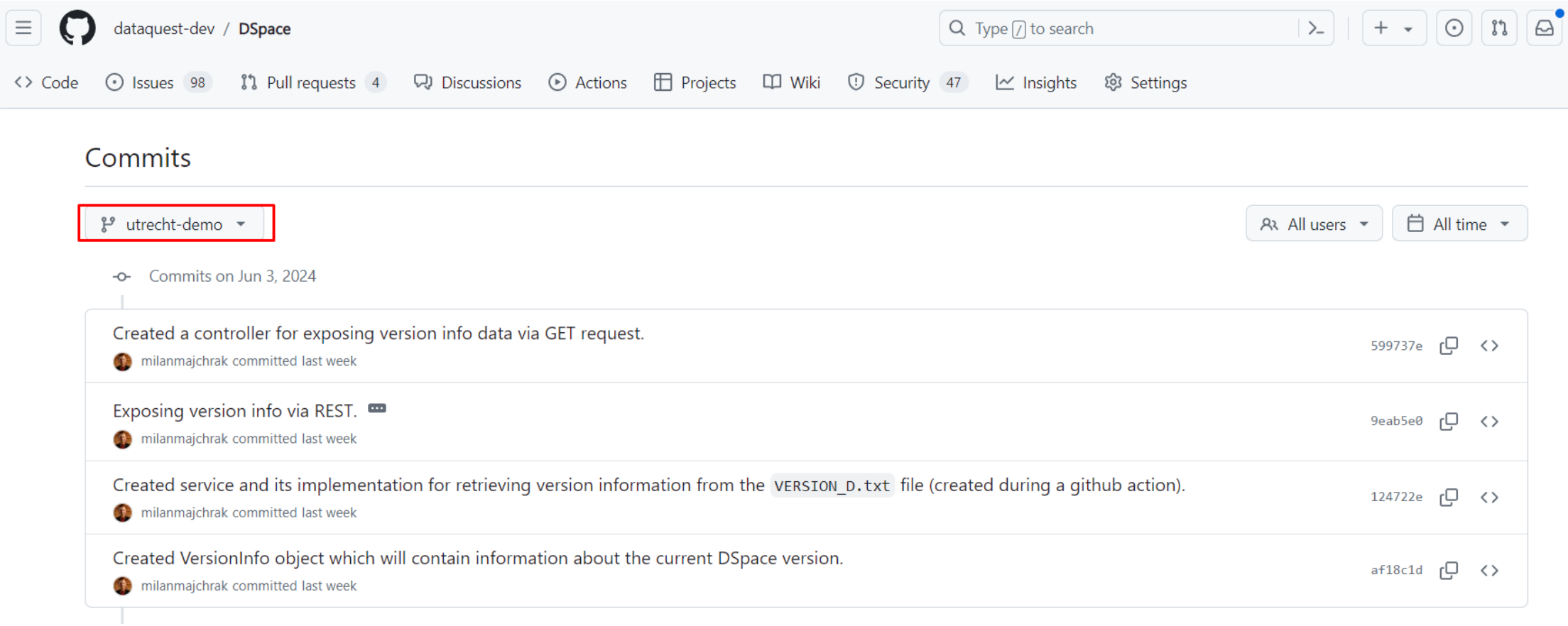
```
2024-06-10 17:39:38,063:INFO: ========== Validating bundle ==========
2024-06-10 17:39:38,067:INFO: Table [              bundle]  OK  compared by len only v5:[3501], v7:[3501]
2024-06-10 17:39:38,067:INFO: ========== Validating collection ==========
2024-06-10 17:39:38,068:INFO: Table [          collection]  OK  compared by len only v5:[13], v7:[13]
2024-06-10 17:39:38,068:INFO: Table [          collection]  OK   NON NULL [logo_bitstream_id] v5:[  3], v7:[  3]
2024-06-10 17:39:38,069:INFO: ========== Validating community2collection ==========
2024-06-10 17:39:38,069:INFO: Table [community2collection]  OK  compared by len only v5:[13], v7:[13]
2024-06-10 17:39:38,069:INFO: ========== Validating community ==========
2024-06-10 17:39:38,070:INFO: Table [           community]  OK  compared by len only v5:[8], v7:[8]
2024-06-10 17:39:38,070:INFO: ========== Validating community2community ==========
2024-06-10 17:39:38,072:INFO: Table [ community2community]  OK  compared by len only v5:[5], v7:[5]
2024-06-10 17:39:38,072:INFO: ========== Validating eperson ==========
2024-06-10 17:39:38,082:INFO: Table [eperson]: v5:[1523], v7:[1524]
  only in v5:[['None|None']]
  only in v7:[['.....', '.....']]
2024-06-10 17:39:38,083:INFO: ========== Validating epersongroup2eperson ==========
2024-06-10 17:39:38,085:INFO: Table [epersongroup2eperson]: v5:[30], v7:[38]
  only in v5:[[]]
  only in v7:[['.....']]
2024-06-10 17:39:38,085:INFO: ========== Validating epersongroup ==========
2024-06-10 17:39:38,086:INFO: Table [        epersongroup]  !!! WARN !!!  compared by len only v5:[34], v7:[42]
2024-06-10 17:39:38,086:INFO: Table [        epersongroup]  !!! WARN !!!   NON NULL [eperson_group_id] v5:[ 34], v7:[  0]
```

+ dataquest
simply data

# Demo example result

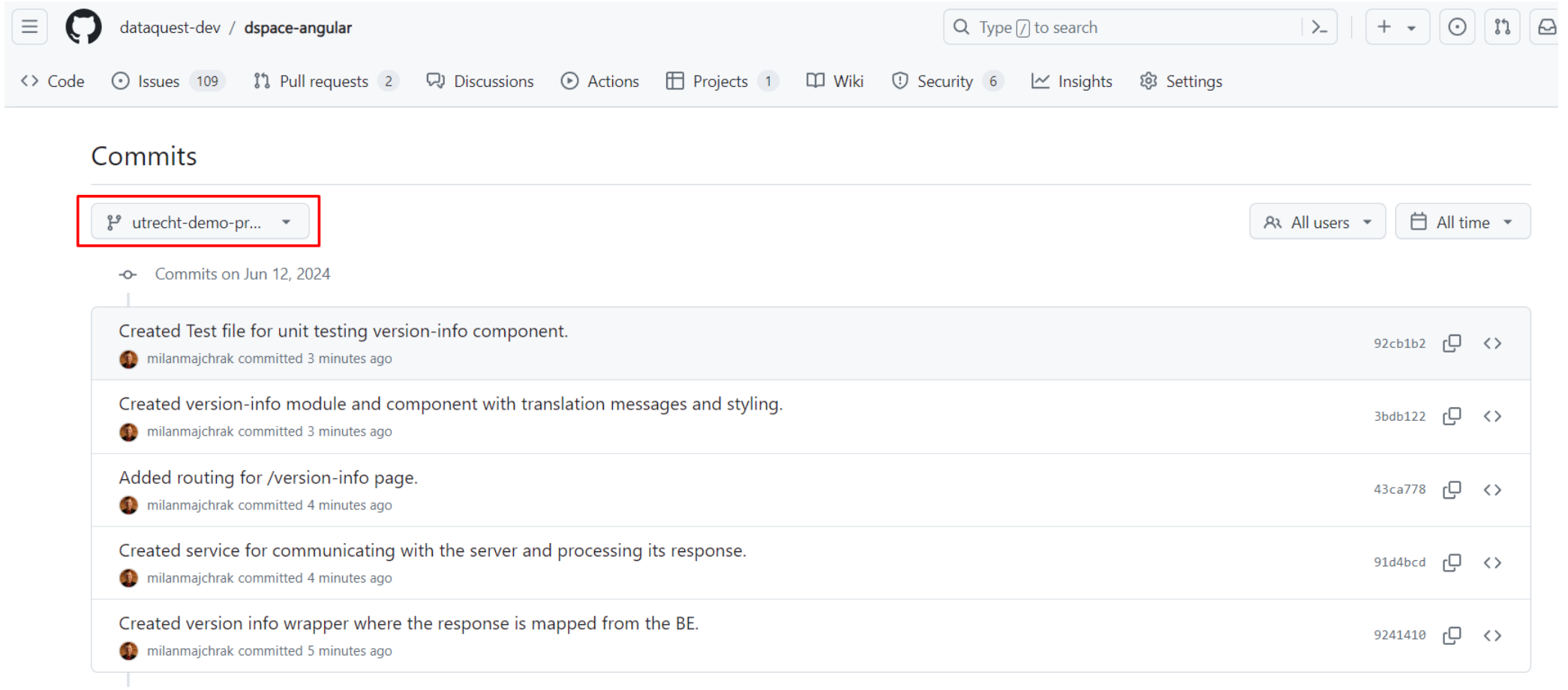


# Commits - BE

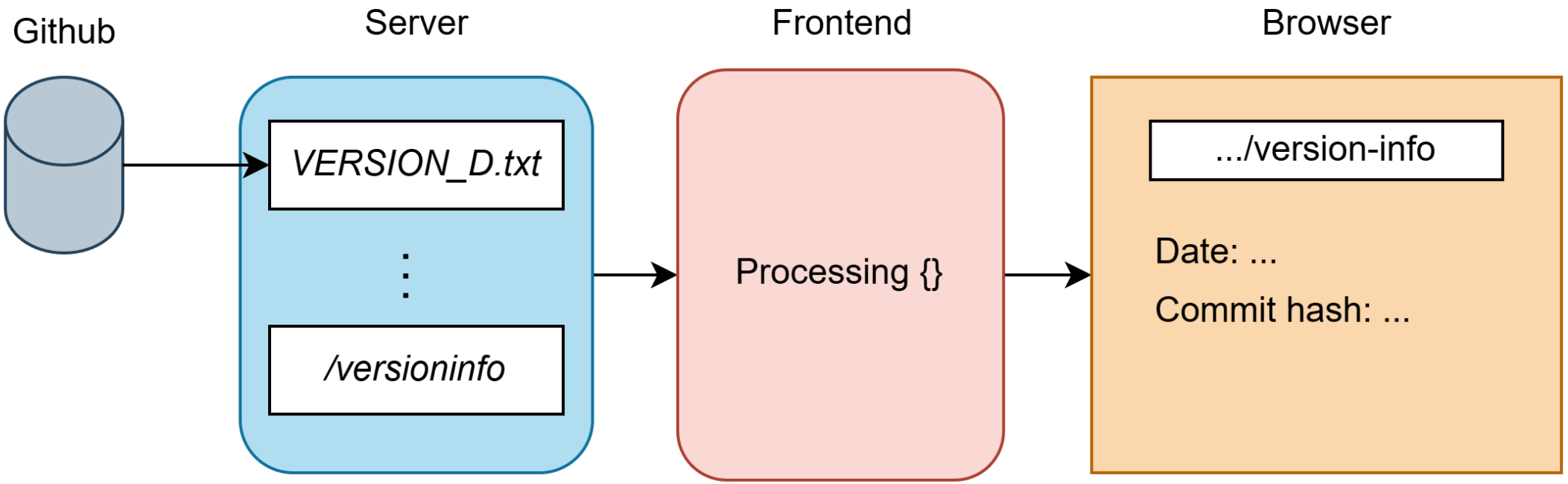



+ dataquest
simply data

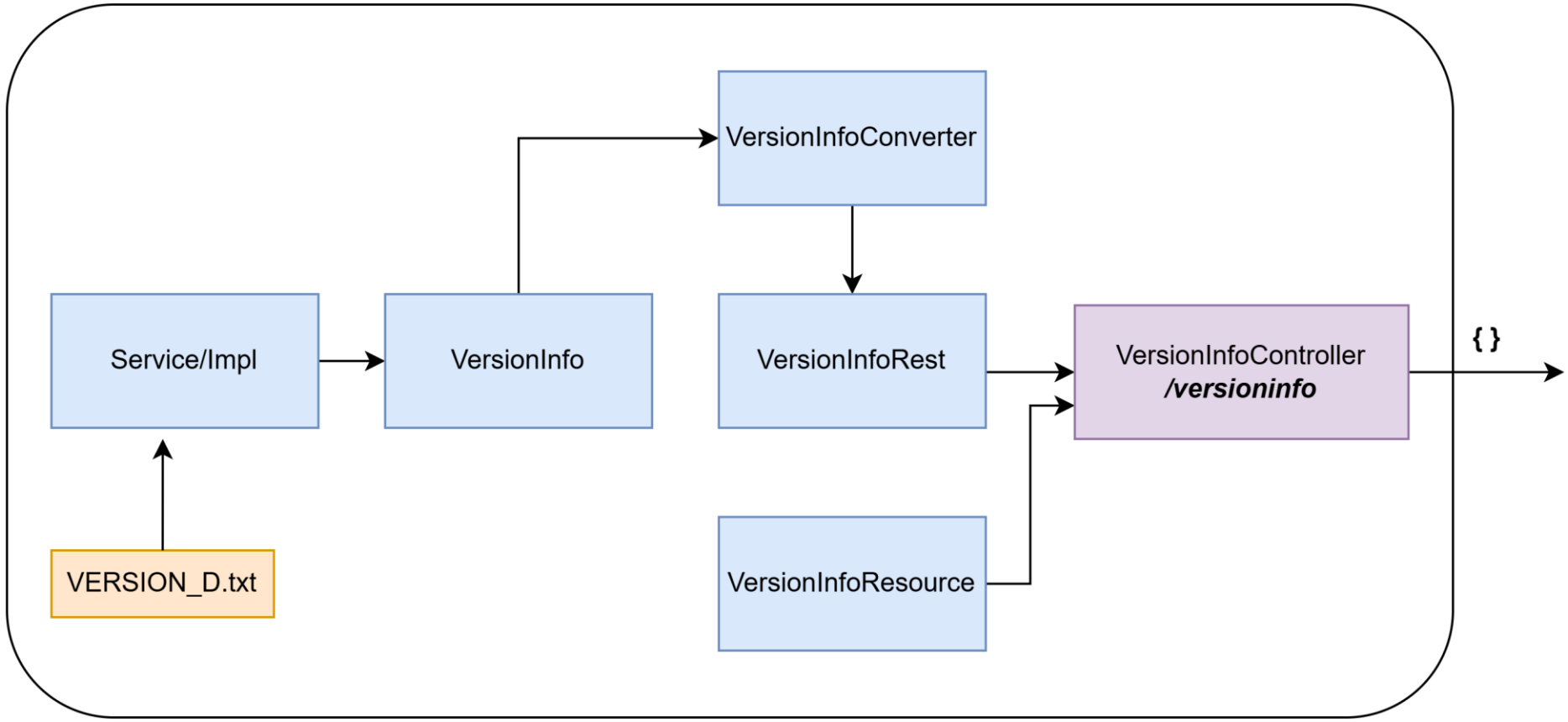
# Commits - FE


dataquest-dev / dspace-angular
Type / to search
Code
Issues 109
Pull requests 2
Discussions
Actions
Projects 1
Wiki
Security 6
Insights
Settings
Commits
utrecht-demo-pr...
All users
All time
Commits on Jun 12, 2024
Created Test file for unit testing version-info component.
milanmajchrak committed 3 minutes ago
92cb1b2
Created version-info module and component with translation messages and styling.
milanmajchrak committed 3 minutes ago
3bdb122
Added routing for /version-info page.
milanmajchrak committed 4 minutes ago
43ca778
Created service for communicating with the server and processing its response.
milanmajchrak committed 4 minutes ago
91d4bcd
Created version info wrapper where the response is mapped from the BE.
milanmajchrak committed 5 minutes ago
9241410

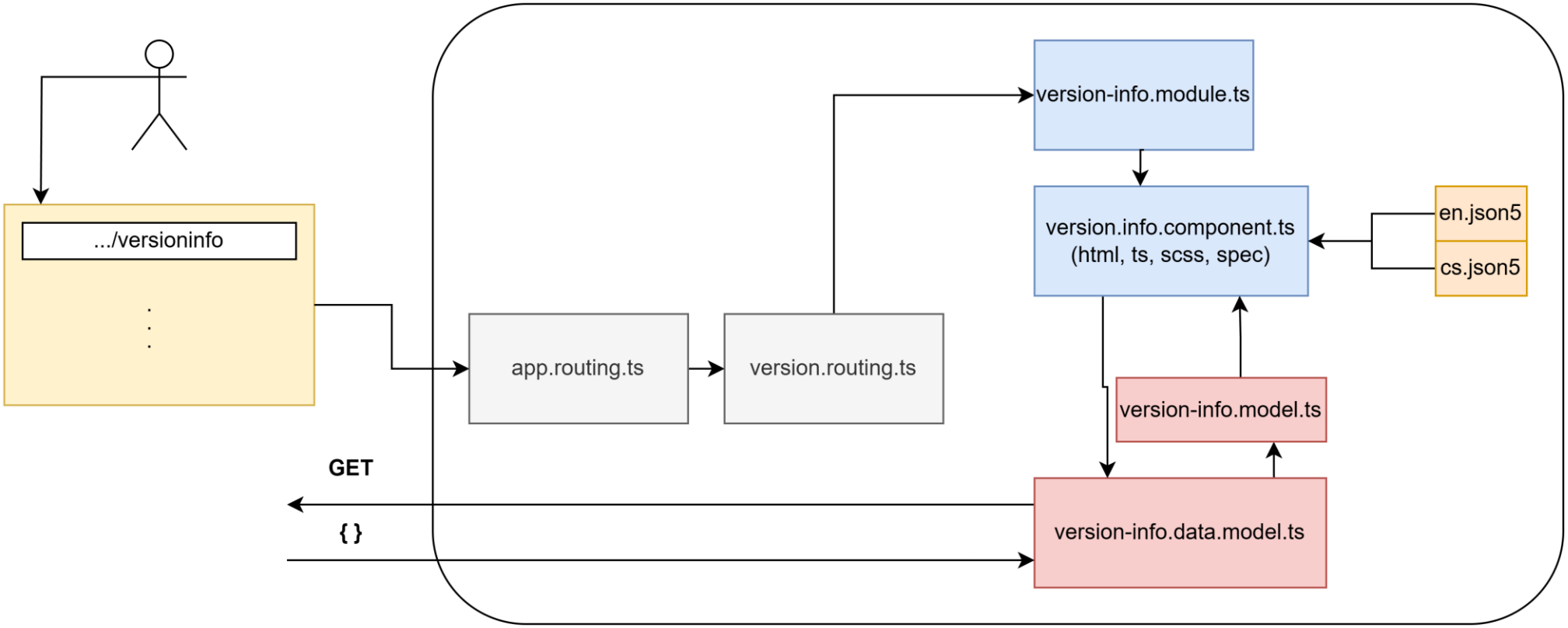
# Demo example light analysis



# BE analysis


VersionInfoConverter
Service/Impl
VersionInfo
VersionInfoRest
VersionInfoController
/versioninfo
{ }
VERSION_D.txt
VersionInfoResource

# FE analysis


.../versioninfo
app.routing.ts
version.routing.ts
version-info.module.ts
version.info.component.ts
(html, ts, scss, spec)
en.json5
cs.json5
version-info.model.ts
version-info.data.model.ts
GET
{ }

# Deployment example

- Git clone FE – locally
- Edit .env
- Compose up (explain when compose pull might be necessary)
- Show logs
- Go to page at end
- Configs, mounts, volumes
- Pm2 monit

# Links

- Main repository - ufal:
  - BE: https://github.com/ufal/clarin-dspace/tree/clarin-v7
  - FE: https://github.com/ufal/dspace-angular/tree/clarin-v7
  - Import: https://github.com/ufal/dspace-migrate
- Our DSpace wiki
  - https://github.com/ufal/clarin-dspace/wiki
- Vanilla Dspace documentation
  - https://wiki.lyrasis.org/display/DSDOC7x
- Import guide
  - https://github.com/ufal/dspace-migrate/blob/main/README.md
- Demo example commits
  - https://github.com/dataquest-dev/DSpace/commits/utrecht-demo/
  - https://github.com/dataquest-dev/dspace-angular/commits/utrecht-demo-presentation

# Thank you for your attention

Milan Majchrák
Marián Berger

+ dataquest
simply data