

# Summary of Software development for the Rucio Scientific Data Management system (as IRIS-HEP Fellows project)



Lev Pambuk,

Odesa National University of Technology

Mentors: Martin Barisits (CERN EP),  
Mario Lassnig (CERN EP)

Period: Jun – Sep, 2023

SCIENTIFIC DATA MANAGEMENT



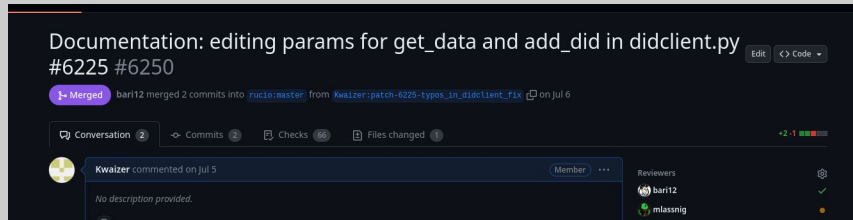
**RUCIO**

# Outline

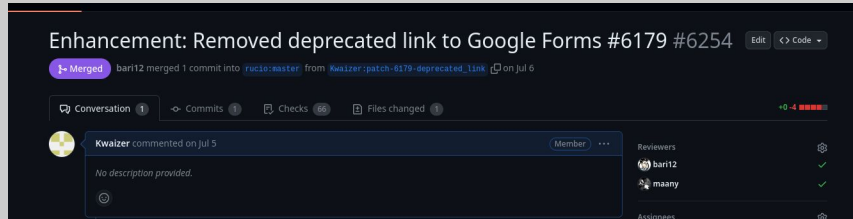
- Project & Objectives
  - Introductory work
  - Client developments
  - Operations developments
  - Core developments
  - Transfer developments
- Introduced:
  - Enhancements
  - Bug fixes
  - New features
  - Deletions

---

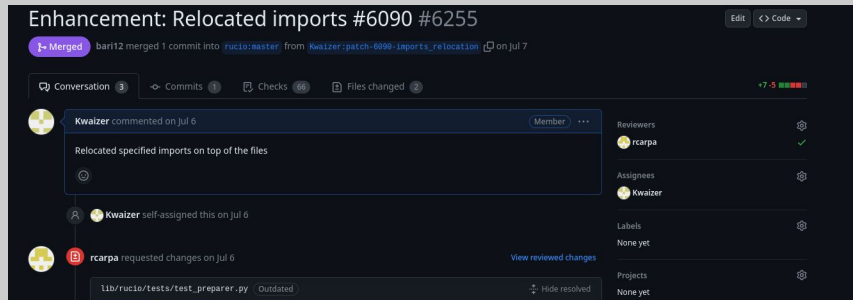
# 1. Correcting typos in docstrings for Python API of DIDClient



# 2. Removing link to deprecated Google Form in Rucio WebUI



# 3. Rearranging of imported statements in files and improving test



## 4. Improving error reporting during upload when reusing a deleted DID



**dchristidis** commented on Feb 20 Member ...

**Description**

What the operators see:

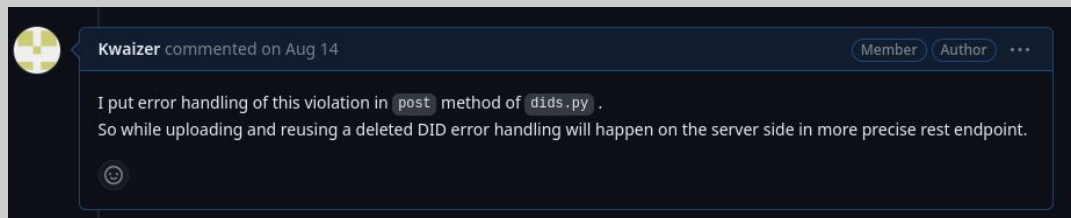
```
Database exception.
Details: (raised as a result of Query-invoked autoflush; consider using a session.no_autoflush block if t... /1
(cx_Oracle.DatabaseError) ORA-20101: Primary key constraint DELETED_DIDS_PK violated
ORA-06512: at "ATLAS_RUCIO.CHECK_DID_UNIQUENESS", line 12
ORA-04088: error during execution of trigger 'ATLAS_RUCIO.CHECK_DID_UNIQUENESS'
[SQL: INSERT INTO atlas_rucio.dids (scope, name, account, did_type, is_open, complete, availability, bytes, len
[parameters: {'scope': 'XXX', 'name': 'XXX', 'account': 'XXX', 'did_type': 'F', 'is_open': None, 'complete': No
(Background on this error at: https://sqlalche.me/e/14/4xp6)
```

What the user sees:

```
2023-02-18 08:10:23,784 INFO Preparing upload for file XXX
2023-02-18 08:10:23,888 ERROR Database exception.
Details: An unknown Database Exception has occurred.
```

**Motivation**

The user is unable to deduce what is going wrong.




**Kwaizer** commented on Aug 14 Member Author ...

I put error handling of this violation in `post` method of `dids.py` .

So while uploading and reusing a deleted DID error handling will happen on the server side in more precise rest endpoint.

# 5. Webdav Protocol stat didn't return data as specified

 **ThePhisch** commented on Nov 15, 2022 Member ...


### Motivation

The docstring for `rse/protocols/webdav.py::Default::stat` states that a dict with two keys (`filesize` and `adler32`) are returned. This matches the implementations of `stat` in the SSH and xrootd protocols (although these protocols do not explicitly say that it is `adler32` but any one of `GLOBALLY_SUPPORTED_CHECKSUMS`). At the moment, the implementation of webdav only returns the `filesize`.






### Modification


Add an entry to the dictionary containing the checksum.


## Protocols: Adding an entry to the dictionary #5977 #6257

 **Merged** bari12 merged 1 commit into `rucio:master` from `Kwaizer:patch-5977-stat_dict` 2 weeks ago

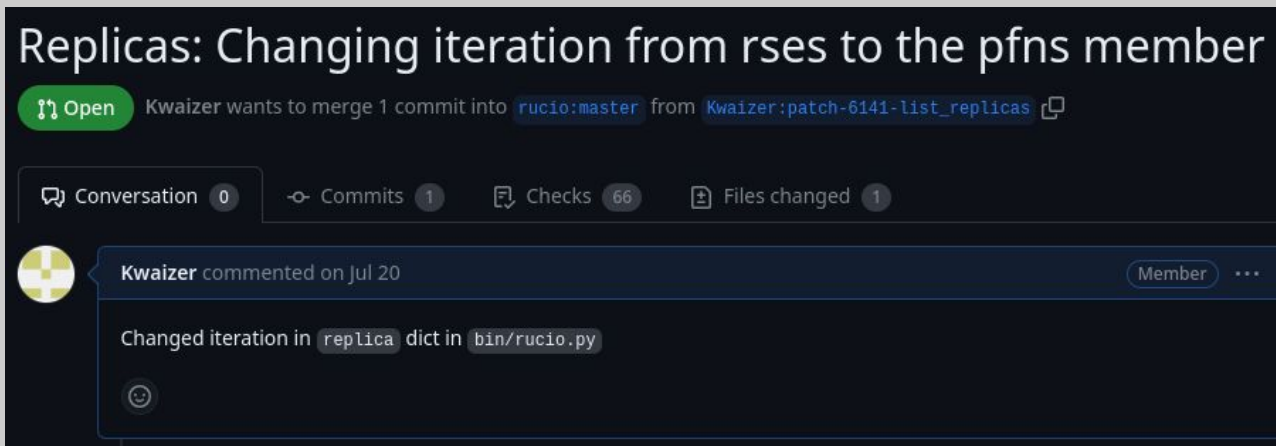
 Conversation **3**  Commits **1**  Checks **42**  Files changed **1**

Changes from all commits  File filter  Conversations  Jump to  

 **Protocols: Removing mentioning of `adler32` in `webdav.py` #5977**

 **Kwaizer** committed on Aug 14

## 6. Rucio client lists replicas in wrong order



Replicas: Changing iteration from rses to the pfns member

[Open](#) Kwaizer wants to merge 1 commit into `rucio:master` from `Kwaizer:patch-6141-list_replicas`

Conversation 0 Commits 1 Checks 66 Files changed 1

Kwaizer commented on Jul 20 Member

Changed iteration in `replica` dict in `bin/rucio.py`

The screenshot shows a GitHub pull request interface. At the top, the title is "Replicas: Changing iteration from rses to the pfns member". Below the title, there is a green "Open" button and a line of text indicating a merge: "Kwaizer wants to merge 1 commit into rucio:master from Kwaizer:patch-6141-list\_replicas". A navigation bar shows "Conversation 0", "Commits 1", "Checks 66", and "Files changed 1". A comment from user "Kwaizer" is visible, dated "Jul 20", with a "Member" badge. The comment text reads: "Changed iteration in replica dict in bin/rucio.py".

# 7. register-after-upload doesn't transition existing replicas to Available

Clients: Transition replicas to Available #6278 #6315

Open Kwaizer wants to merge 1 commit into rucio:master from Kwaizer:patch-6278-register\_after\_upload\_transition

Conversation 0 Commits 1 Checks 42 Files changed 1

Changes from all commits File filter Conversations Jump to

```
lib/rucio/client/uploadclient.py
404 404
405 405 # add file to rse if it is not registered yet
406 406 replicastate = list(self.client.list_replicas([file_did], all_states=True))
407 + for replica in replicastate:
408 +     if 'states' in replica and rse in replica['states'] and replica['states'].get(rse) != 'AVAILABLE':
409 +         replica['states'].get(rse).replace(
410 +             replica['states'].get(rse), 'AVAILABLE')
```

register-after-upload doesn't transition existing replicas to

Open rcarpa opened this issue on Jul 21 · 0 comments

rcarpa commented on Jul 21 · edited Member

When this option is set to `rucio_upload`, this code is executed:

```
rucio/lib/rucio/client/uploadclient.py
Line 290 in 11cecaf
290     if register_after_upload:
```


but, `_register_file` doesn't do anything if the replica exists:


```
rucio/lib/rucio/client/uploadclient.py
Line 408 in 11cecaf
408     if rse not in replicastate[0]['rses']:
```

Even if it's not in the available state. We should transition the replica to available if it's not already in this state.

## 8. Allow to declare suspicious replicas by RSE and LFN

Allow to declare suspicious replicas by RSE and LFN #5906

 Open nsmith- opened this issue on Oct 6, 2022 · 0 comments

 nsmith- commented on Oct 6, 2022 Member ...



**Motivation**

In #5392 `declare_bad_file_replicas` was modified to allow either a list of PFNs (string) or a list of replicas (`'scope': , 'name': , 'rse_id': <rse_id> or "rse": <rse_name>`) as inputs. The `declare_suspicious_file_replicas` API does not have the same feature.

**Modification**


`declare_suspicious_file_replicas` should be modified to have the same feature.

Enhancement: Modified `declare_suspicious_file_replicas` #5906

 Open Kwaizer wants to merge 1 commit into `rucio:master` from `Kwaizer:patch-5906-declare_rse_lfn` 



# 9. Subscription: Adding check for subscription #6233

 bjwhite-fnal commented on Jun 23 Contributor ...

### Description

Currently when you try to ask Rucio if a user has Subscriptions and they have none, a `SubscriptionNotFound` exception is intentionally raised.

This has been very confusing for myself and other users as we learn to use Subscriptions. It would be much clearer if this just indicated that there were no subscriptions for this user, without the scary looking traceback.

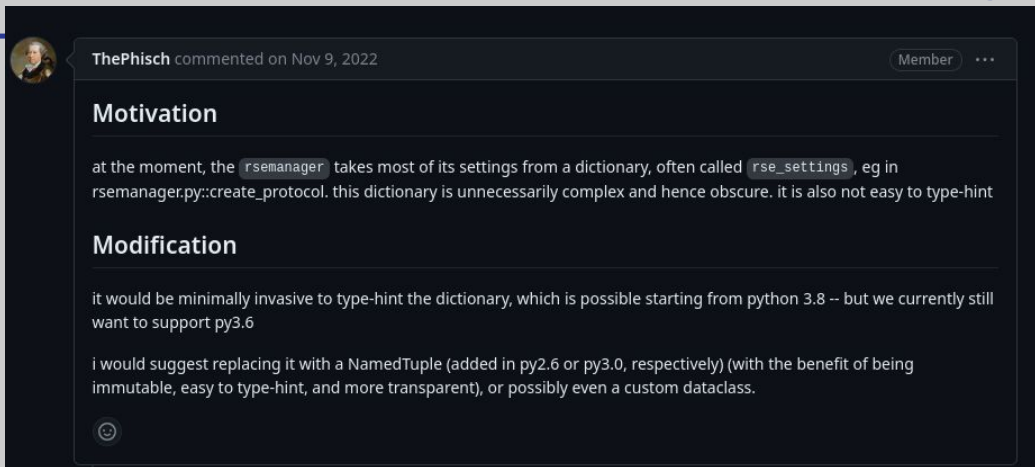
```
(rucio-clients) [bjwhite@GALACTICA:~/rucio-clients]$ rucio-admin subscription list --account root
Subscription not found.
Details: Subscription for account 'root' named 'None' not found
Rucio exited with an unexpected/unknown error, please provide the traceback below to the developers.
Traceback (most recent call last):
```

```
923 +     if 'No subscriptions for this account' in subs:
924 +         print("There are no subscriptions for this account")
925 +     else:
926 +         for sub in subs:
927 +             if args.long:
928 +                 print('\n'.join('%s: %s' % (str(k), str(v)) for (k, v) in list(sub.items())))
929 +                 print()
930 +             else:
931 +                 print("%s: %s %s\n priority: %s\n filter: %s\n rules: %s\n comments: %s" % (sub['account'], sub['name'], sub['state'], sub['policyid'], sub['filter'],
sub['replication_rules'], sub.get('comments', '')))
929 932         return SUCCESS
930 933
931 934
```

lib/rucio/client/subscriptionclient.py

```
.. @ -96,6 +96,8 @@ def list_subscriptions(self, name=None, account=None):
96 96         return self._load_json_data(result)
97 97     else:
98 98         exc_cls, exc_msg = self._get_exception(headers=result.headers, status_code=result.status_code, data=result.content)
99 +         if f'Subscription for account \"{account}\" named \"None\" not found' in exc_msg:
100 +             return str('No subscriptions for this account')
```

# 10. Core & Internals: rse\_settings dictionary



ThePhisch commented on Nov 9, 2022

Member ...

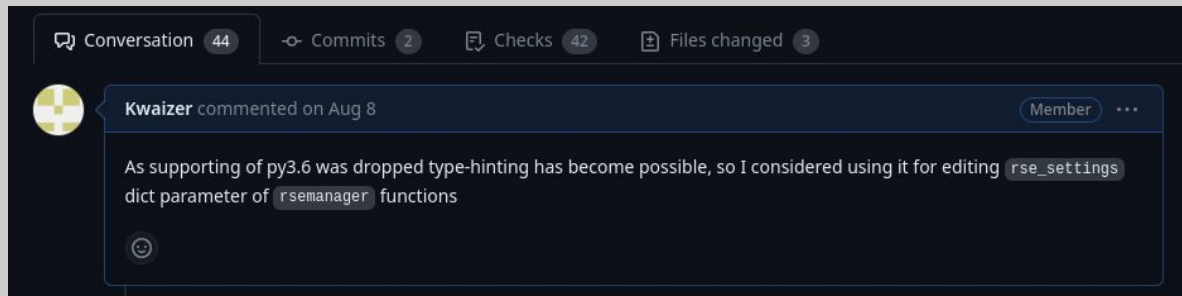
## Motivation

at the moment, the `rsemanger` takes most of its settings from a dictionary, often called `rse_settings`, eg in `rsemanger.py::create_protocol`. this dictionary is unnecessarily complex and hence obscure. it is also not easy to type-hint

## Modification

it would be minimally invasive to type-hint the dictionary, which is possible starting from python 3.8 -- but we currently still want to support py3.6

i would suggest replacing it with a NamedTuple (added in py2.6 or py3.0, respectively) (with the benefit of being immutable, easy to type-hint, and more transparent), or possibly even a custom dataclass.



Conversation 44

Commits 2

Checks 42

Files changed 3

Kwaizer commented on Aug 8

Member ...

As supporting of py3.6 was dropped type-hinting has become possible, so I considered using it for editing `rse_settings` dict parameter of `rsemanger` functions

```
106 +
107 + class RSEDomainLANDict(TypedDict):
108 +     read: Optional[int]
109 +     write: Optional[int]
110 +     delete: Optional[int]
111 +
112 +
113 + class RSEDomainWANDict(TypedDict):
114 +     read: Optional[int]
115 +     write: Optional[int]
116 +     delete: Optional[int]
117 +     third_party_copy_read: Optional[int]
118 +     third_party_copy_write: Optional[int]
119 +
120 +
121 + class RSEDomainsDict(TypedDict):
122 +     lan: RSEDomainLANDict
123 +     wan: RSEDomainWANDict
124 +
125 +
126 + class RSEProtocolDict(TypedDict):
127 +     auth_token: Optional[str] # FIXME: typing.NotRequired
128 +     hostname: str
129 +     scheme: str
130 +     port: int
131 +     prefix: str
132 +     impl: str
133 +     domains: RSEDomainsDict
134 +     extended_attributes: Optional[Union[str, dict[str, Any]]]
135 +
136 +
137 + class RSESettingsDict(TypedDict):
138 +     availability_delete: bool
139 +     availability_read: bool
140 +     availability_write: bool
141 +     credentials: Optional[dict[str, Any]]
142 +     lfn2pfn_algorithm: str
143 +     qos_class: Optional[str]
144 +     staging_area: bool
145 +     rse_type: str
146 +     sign_url: Optional[str]
147 +     read_protocol: int
148 +     write_protocol: int
149 +     delete_protocol: int
150 +     third_party_copy_read_protocol: int
151 +     third_party_copy_write_protocol: int
152 +     id: str
153 +     rse: str
154 +     volatile: bool
155 +     verify_checksum: bool
156 +     deterministic: bool
157 +     domain: list[str]
158 +     protocols: list[RSEProtocolDict]
```

To be frank, I think it's actually better to use `dict[str, Any]`. It creates a placeholder for future work: using `typing.TypedDict` to unequivocally define the structure and expected content of dictionaries. We've discussed it multiple times, but haven't started working on it.



# 11. Adding size information to list\_rules()

Rules: Size information is returned when listing replication rules #5978 #6297

Merged bari12 merged 3 commits into rucio:master from Kwaizer:patch-5978-size\_info on Aug 11

Conversation 17 Commits 3 Checks 42 Files changed 2 +22 -12

Kwaizer commented on Aug 3

Modified `core` to accept the `size` info and return it while listing replication rules

Reviewers

- dchristidis ✓
- bari12 ●
- mlassnig ●

Assignees

No one—assign yourself

Kwaizer requested review from bari12 and mlassnig as code owners 2 months ago

Rules: Size information is returned when listing replication rules #5978 #6297

Merged bari12 merged 3 commits into rucio:master from Kwaizer:patch-5978-size\_info on Aug 11

Conversation 17 Commits 3 Checks 42 Files changed 2 +22 -12


Commits on Aug 8, 2023

- Rules: SQLAlchemy 2.0 migration in list\_rules rucio#5978  
Kwaizer committed on Aug 8 6baa8bc
- Rules: Replacement with to\_dict() in list\_rules() rucio#5978  
Kwaizer committed on Aug 8 6e979be

Commits on Aug 11, 2023

- Rules: Listing replication rules returns size info rucio#5978  
Kwaizer committed on Aug 11 ✓ 856faf7

# 12. Improper use of NoResultFound

 dchristidis commented on Aug 11 • edited

Member ...

### Description

We need to review all uses of `NoResultFound` because there's at least one that is misleading (it will never be raised):


```
rucio/lib/rucio/core/did.py
Lines 1649 to 1671 in 9d43293
```

```
1661     ).with_hint(
1662         models.DataIdentifierAssociation, "INDEX(CONTENTS CONTENTS_PK)", 'oracle'
1663     ).filter_by(
1664         scope=scope,
1665         name=name
1666     )
1667     for tmp_did in session.execute(stmt).yield_per(5).scalars():
1668         yield {'scope': tmp_did.child_scope, 'name': tmp_did.child_name, 'type': tmp_did.child
1669             'bytes': tmp_did.bytes, 'adler32': tmp_did.adler32, 'md5': tmp_did.md5}
1670     except NoResultFound:
1671         raise exception.DataIdentifierNotFound("Data identifier '%(scope)s:%(name)s' not found" %
```

### Steps to reproduce

```
>>> from rucio.common.types import InternalScope
>>> from rucio.core.did import list_content
>>> list(list_content(scope=InternalScope('foo'), name='bar'))
[]
```

Conversation 0 Commits 2 Checks 42 Files changed 3

 Kwaizer commented last month

Member ...

Initial idea of this PR was introducing `yielded` flag (like in `lib/rucio/core/replica.py`) which will indicate if any result was yielded at all and raise exception if not. Such implementation wasn't fully successful for all uses in `lib/rucio/core/did.py` due to failed tests so I considered removing unused constraints here for now.

# Thank you for your attention!



4th of October 2023