Arvados - Feature #17944

add vocabulary validation to controller

07/28/2021 03:53 PM - Ward Vandewege

Status:	Resolved	**Start date:**	11/02/2021
Priority:	Normal	**Due date:**	
Assigned To:	Lucas Di Pentima	**% Done:**	100%
Category:		**Estimated time:**	0.00 hour
Target version:	2021-11-24 sprint		

Description

Cf. https://doc.arvados.org/admin/workbench2-vocabulary.html

- When an Arvados object that has properties (collections, container_requests, groups, links) is created or updated, the API server will validate the properties contents. Properties are key-value pairs (property definitions are called “tags” in the vocabulary file)
- Property keys are checked against the standardized key identifiers defined in the vocabulary file. The key is also checked against the aliases (labels) for each tag. If a property key matches one of the aliases, the API server returns an error indicating that the client is required to use the standardized identifier for the key.
- The property value is checked that it is in the range of values for the tag as defined in the vocabulary file.
  When “strict” is true, the value must be one of the standardized value identifiers listed for that tag. If it is not a standardized value identifier, the API server returns an error. It does not accept aliases, but if the provided value matches an alias, the error message should indicate as such.
- When “strict” is false or undefined, the value must either be one of the standardized values listed for that tag, or it must be a value that is not listed in aliases. If the value is listed in aliases, it should return an error that the client is required to use the standardized identifier.
- When a value is rejected due to use of an alias and not the standardized value identifier, the error message should include what standardized value identifier was expected.
- Use case insensitive match to check if a key or value matches an alias
- Respect the value of "strict_tags" in the vocabulary file for unknown property keys, can specify either:
  - strict_tags: false -- Property keys which are not defined in the vocabulary are not checked
  - strict_tags: true -- Property keys which are not defined in the vocabulary are rejected
- Property validation is applied to all users, including admins
- The configuration file will be stored somewhere on the filesystem of the host that runs Arvados controller. The controller will have an API endpoint that Workbench 2 or other applications can use to fetch the vocabulary file.
- If a vocabulary file is configured but cannot be read at startup, Arvados controller will fail with an error.
- If the same alias is associated with more than one standardized identifier, fail with an error.
- The config-check subcommand will detect and report configuration and vocabulary file errors.
- To ease migration, if a record is updated but the update does not change the properties, it should not reject the update of unrelated fields even if the current properties are invalid
- When strict_tags is enabled, need to recognize and special case properties already in use by Arvados tools. Some properties (list is likely incomplete?)
  - type
  - template_uuid
  - groups
  - username
  - image_timestamp
  - docker-image-repo-tag
  - filters
  - container_request
  - Other configured managed properties

Also: arvados-cwl-runner has a 'cache http download' feature that notes the provenance by setting the source URL as the key that maps to an object containing the cache headers. This usage is incompatible with "strict_tag".

Implementation:

- Validation happens in controller for create and update calls
- Add config parameter to API/VocabularyPath, expected to be local to the machine the controller runs on.
- The vocabulary file will be loaded and cached by controller; file timestamp will be checked on any request. If the vocabulary file can't be read (e.g. permissions, invalid json, etc), the existing cached version will be used and a health warning/prometheus alert should be raised.
- If the file can't be read on startup, that's an error. config-check should also check this, and will need to take into account that this

is only an error if the context is the controller.
- Apply validation before all save/update requests. Admin users do not get special treatment.
- The validation code should handle existing data gracefully: if a record has existing properties are invalid, but the update does not include properties, updates to other fields in the collection should still be permitted.
- Update wb2 to get the file from controller (it will need to export the cache copy as a valid URL to the JSON)

Subtasks:

Task # 18268: Review 17944-backend-vocabulary-validation-rebased (arvados repo) & 17944...	**Resolved**

Related issues:

Blocks Arvados Epics - Story #17454: Vocabulary checking of properties by API...	**Resolved**	**10/01/2021**	**03/31/2022**

Associated revisions

Revision a7876235 - 11/11/2021 05:51 PM - Lucas Di Pentima

Merge branch '17944-backend-vocabulary-validation-rebased' into main.

Refs #17944

Arvados-DCO-1.1-Signed-off-by: Lucas Di Pentima <lucas.dipentima@curii.com>

Revision eabdb7bd - 11/11/2021 05:53 PM - Lucas Di Pentima

Merge branch '17944-vocabulary-endpoint-retrieval' into main. Closes #17944

Arvados-DCO-1.1-Signed-off-by: Lucas Di Pentima <lucas.dipentima@curii.com>

Revision 5c431672 - 11/11/2021 07:52 PM - Lucas Di Pentima

Merge branch '17944-backend-vocabulary-validation-rebased' into main.

Refs #17944

Arvados-DCO-1.1-Signed-off-by: Lucas Di Pentima <lucas.dipentima@curii.com>

Revision bb1b66c4 - 11/11/2021 08:03 PM - Lucas Di Pentima

Merge branch '17944-vocabulary-endpoint-retrieval' into main. Closes #17944

Arvados-DCO-1.1-Signed-off-by: Lucas Di Pentima <lucas.dipentima@curii.com>

History

#1 - 07/28/2021 03:53 PM - Ward Vandewege

- Blocks Story #17454: Vocabulary checking of properties by API server/controller added

#2 - 07/28/2021 03:59 PM - Ward Vandewege

- Description updated

#3 - 07/28/2021 04:26 PM - Ward Vandewege

- Description updated

#4 - 07/28/2021 04:28 PM - Ward Vandewege

- Description updated

#5 - 07/28/2021 04:34 PM - Ward Vandewege

- Description updated

#6 - 07/28/2021 04:36 PM - Ward Vandewege

- Story points set to 3.0

#7 - 07/28/2021 04:51 PM - Ward Vandewege

- Target version set to 2021-08-18 sprint

#8 - 08/03/2021 02:56 PM - Peter Amstutz

- Target version deleted (2021-08-18 sprint)

#9 - 08/03/2021 02:57 PM - Peter Amstutz

- Target version set to 2021-09-01 sprint

- Project changed from Arvados Epics to Arvados

#10 - 08/10/2021 05:36 PM - Peter Amstutz

- Target version changed from 2021-09-01 sprint to 2021-09-15 sprint

#11 - 08/31/2021 07:48 PM - Peter Amstutz

- Target version changed from 2021-09-15 sprint to 2021-09-29 sprint

#12 - 09/14/2021 07:14 PM - Peter Amstutz

- Target version changed from 2021-09-29 sprint to 2021-10-13 sprint

#13 - 09/14/2021 07:14 PM - Peter Amstutz

- Target version changed from 2021-10-13 sprint to 2021-10-27 sprint

#14 - 10/12/2021 07:38 PM - Peter Amstutz

- Release set to 45

#15 - 10/12/2021 08:34 PM - Peter Amstutz

- Description updated

#16 - 10/12/2021 09:02 PM - Peter Amstutz

- Description updated

#17 - 10/13/2021 03:51 PM - Peter Amstutz

- Assigned To set to Lucas Di Pentima

#18 - 10/20/2021 07:44 PM - Lucas Di Pentima

- Status changed from New to In Progress

#19 - 10/25/2021 03:16 PM - Lucas Di Pentima

- Description updated

#20 - 10/27/2021 02:50 PM - Lucas Di Pentima

- Target version changed from 2021-10-27 sprint to 2021-11-10 sprint

#21 - 11/02/2021 12:48 PM - Lucas Di Pentima

Updates at 3849d76f2 - branch 17944-backend-vocabulary-validation
Test run: developer-run-tests: #2766 icon?job=developer-run-tests&amp;build=2766

Updates

- Adds API.VocabularyPath config knob used by controller.
- Updates the documentation.
- Adds arvados.Vocabulary type with loading validation property check func & tests.
- Vocabulary checking supports exceptional cases, including system & managed properties.
- Adds /arvados/v1/vocabulary endpoint to allow clients to request the vocabulary.
- Adds vocabulary checks to create/update calls of collections, groups & container requests.
- Sets up a file watcher to reload the vocabulary file on demand, only replacing the previously cached vocabulary on success.
- Adds a health check on boot so that controller won't start with a non-valid vocabulary.

Pending

- Expose the new endpoint as a Workbench.VocabularyURL exported config so that older workbenches automatically migrate to the correct vocabulary.
- Update wb2 to request the vocabulary from the new endpoint.
- Add vocabulary file reloading test -- I tried making it a unit test (part of lib/controller/handler_test.go), but the file watcher didn't seem to work, probably will need making it an integration test.
- Add arvados#link resources to do vocabulary checking.
- Notify of invalid vocabulary via Prometheus alert.

Notes

The commit history is messy because I changed the approach to vocabulary loading and didn't know how to merge it with the rest properly. Commit 149285d changes the way controller loads and cache the vocabulary.

#22 - 11/02/2021 05:12 PM - Peter Amstutz

- Release changed from 45 to 46

#23 - 11/02/2021 05:31 PM - Peter Amstutz

- Release changed from 46 to 45

#24 - 11/02/2021 11:13 PM - Lucas Di Pentima

Arvados repo

Rebased & squashed into latest main (new branch): 1cd689f - branch 17944-backend-vocabulary-validation-rebased
Test run: developer-run-tests: #2770 icon?job=developer-run-tests&amp;build=2770

Workbench2 repo

Updates at arvados-workbench2|ed768b6 - branch 17944-vocabulary-endpoint-retrieval
Test run: developer-tests-workbench2: #506 icon?job=developer-tests-workbench2&amp;build=506

- Retrieves the vocabulary from the new endpoint.
- Fixes the Cypress tests by configuring the testing Arvados instance with an example vocabulary.

#25 - 11/04/2021 02:08 AM - Lucas Di Pentima

Updates at 3f32ceb - branch 17944-backend-vocabulary-validation-rebased
Test run: developer-run-tests: #2773 icon?job=developer-run-tests&amp;build=2773

- Adds support for property checking to links.

#26 - 11/05/2021 08:17 PM - Lucas Di Pentima

Updates at 7b7de0b - branch 17944-backend-vocabulary-validation-rebased
Test run: developer-run-tests: #2775 icon?job=developer-run-tests&amp;build=2775

- Adds /_health/vocabulary health endpoint
- Improves vocabulary reloading by replacing the fsnotify method with a periodic modify time on the file.

#27 - 11/08/2021 09:53 PM - Peter Amstutz

reviewing 17944-backend-vocabulary-validation-rebased @ 7b7de0ba345c02103bbaa9fb981424c59d440d55

I found a bug:

```
$ arv group update -u x2z00-j7d0g-3lryy66m7p8s9dn --group '{"properties":{"IDTAGSIZES": "blah blah"}}'
Error: tag value "blah blah" for key "IDTAGSIZES" is not listed as valid
peter@curiipeter:[pts/2]:~
$ arv group update -u x2z00-j7d0g-3lryy66m7p8s9dn
 --group '{"properties":{"IDTAGCATEGORIES": "IDTAGCAT3", "IDTAGSIZES": "blah blah"}}'
{
 "created_at":"2021-11-05T20:10:25.544636000Z",
 "delete_at":null,
 "description":null,
 "etag":"bejp1nrti83yrbylhe5a9dqf3",
 "group_class":"project",
 "href":"/groups/x2z00-j7d0g-3lryy66m7p8s9dn",
 "is_trashed":false,
 "kind":"arvados#group",
 "modified_at":"2021-11-08T18:53:51.031010000Z",
 "modified_by_client_uuid":"x2z00-ozdt8-8mg1eis92gnjrk6",
 "modified_by_user_uuid":"x2z00-tpzed-9r15h8tfsbqjc8l",
 "name":"humans",
 "owner_uuid":"x2z00-tpzed-9r15h8tfsbqjc8l",
 "properties":{
  "IDTAGCATEGORIES":"IDTAGCAT3",
  "IDTAGSIZES":"blah blah"
 },
 "trash_at":null,
 "uuid":"x2z00-j7d0g-3lryy66m7p8s9dn",
 "writable_by":[
  "x2z00-tpzed-9r15h8tfsbqjc8l",
  "x2z00-tpzed-9r15h8tfsbqjc8l"
 ]
}
```

It seems that when one property is valid it doesn't error out on the other one.

error messages

> tag value "IDVALANIMALS2" for key "IDTAGCATEGORIES" is not listed as valid

Can we rephrase that:

> tag value "IDVALANIMALS2" is not valid for key "IDTAGCATEGORIES"

Also:

> tag value "XS" for key "IDTAGSIZES" is not defined but is an alias for "IDVALSIZES1"

This is pretty good but it doesn't quite tell the user what they are supposed to do, maybe something like this?

> tag value "XS" for key "IDTAGSIZES" is an alias, must be provided as "IDVALSIZES1"

Another suggestion:

> tag key "IDTAGSIZESz" is not defined

This is a little vague (not defined where?), let's make the messages a little more specific to say "defined in the vocabulary":

> tag key "IDTAGSIZESz" is not defined in the vocabulary

Also there seems to be a formatting error here

```
$ arv link update -u x2z00-o0j2j-4s1hh8z52bbunh3  --link '{"properties":{"IDTAGSIZES": null}}'
Error: tag value %!(MISSING)q(<nil>) for key "IDTAGSIZES" is not a valid type (<nil>)
peter@curiipeter:[pts/2]:~
$ arv link update -u x2z00-o0j2j-4s1hh8z52bbunh3  --link '{"properties":{"IDTAGSIZES": 12}}'
Error: tag value %!(MISSING)q(float64=12) for key "IDTAGSIZES" is not a valid type (float64)
```

Validation

Is it checking that you don't have aliases that conflict with vocabulary values? I see tests for duplicated aliases, but not conflicts between terms an aliases.

It looks like you do case-insensitive checks for conflicts with aliases, but the actual property checking is case sensitive. Could we also consider the ToLower() version of identifier an alias, so this would give an '"idvalsizes1" is an alias of "IDVALSIZES1"' error, instead of an invalid error:

```
$ arv link update -u x2z00-o0j2j-4s1hh8z52bbunh3  --link '{"properties":{"IDTAGSIZES": "idvalsizes1" }}'
Error: tag value "idvalsizes1" for key "IDTAGSIZES" is not listed as valid
```

hot reloading

It should log an info message when successfully hot reloading (unless it does, and I missed it). I couldn't tell if it reloaded the vocabulary file or not.

#28 - 11/09/2021 10:06 PM - Lucas Di Pentima

Updates at e7aec8c18 addressing Peter's feedback.
Test run: developer-run-tests: #2781 icon?job=developer-run-tests&amp;build=2781

- Fixes premature vocabulary check success.
- Improves error messages.
- Improves keys & value collision validation against aliases.
- Adds a case-insensitive check for key/value against labels.
- Fixes logging when reloading vocabulary
- Adds & improves tests.

Regarding the logging message on vocabulary reload: you should see it when making a request that needs the vocabulary (voc export endpoint, or create/update request)

#29 - 11/09/2021 10:36 PM - Lucas Di Pentima

Updates at c51e85a53
Test run: developer-run-tests: #2782 icon?job=developer-run-tests&amp;build=2782

- Test fixing.

#30 - 11/10/2021 02:44 PM - Peter Amstutz

I noticed this:

```
                     "HIGH": {
-                        "labels": [{"label": "High"}]
+                        "labels": [{"label": "High priority"}]
                     },
```

Is this because now "HIGH" and "High" are in conflict? It should probably allow aliases that only differ by case from the identifier they are associated with. There's no confusion in this case.

This error message is not as messed up as before, but it still expresses something is wrong without saying what to do about it:

> $ arv group update -u x2z00-j7d0g-3lryy66m7p8s9dn --group '{"properties":{"IDTAGCATEGORIES": 12}}'
> Error: tag value of type float64 for key "IDTAGCATEGORIES" is not a valid

How about:

> Error: value type for tag key "IDTAGCATEGORIES" was float64, but expected a string or list of strings

#31 - 11/10/2021 04:02 PM - Lucas Di Pentima

- Target version changed from 2021-11-10 sprint to 2021-11-24 sprint

#32 - 11/10/2021 05:51 PM - Peter Amstutz

From sprint review:

- in Workbench 2, if you type in a name that is an alias for one of the vocabulary tags and then hit "tab" without selecting it in the dropdown, it'll use the raw text as the key instead of mapping it to the vocabulary. This will give the user a "can't use alias" error. When the user clicks/tabs off the text box, it should still convert it to the vocabulary id.

#33 - 11/10/2021 07:43 PM - Lucas Di Pentima

Updates at 617d78398
Test run: developer-run-tests: #2784 icon?job=developer-run-tests&amp;build=2784

Peter Amstutz wrote:

> I noticed this:
> [...]
> Is this because now "HIGH" and "High" are in conflict? It should probably allow aliases that only differ by case from the identifier they are associated with. There's no confusion in this case.

Good point. I've fixed que validation code to only consider value aliases collisions from different value ids.

> This error message is not as messed up as before, but it still expresses something is wrong without saying what to do about it:
> [...]
> How about:
> Error: value type for tag key "IDTAGCATEGORIES" was float64, but expected a string or list of strings

Thanks. Fixed & added some more tests.

#34 - 11/10/2021 08:49 PM - Lucas Di Pentima

Updates at 54d36a634 adds upgrade notes and a reference to the vocabulary endpoint.

#35 - 11/10/2021 11:31 PM - Lucas Di Pentima

Updates at arvados-workbench2|c4cc8cb0
Test run: developer-tests-workbench2: #508 icon?job=developer-tests-workbench2&amp;build=508

- Adds property key and value auto-selection with case-insensitive matching.

- Expands test.

#36 - 11/11/2021 05:04 PM - Peter Amstutz

Lucas Di Pentima wrote:

> Updates at 54d36a634 adds upgrade notes and a reference to the vocabulary endpoint.

One note from gitter, these specific test changes can be reverted, since "High" is no longer considered a conflict in this case:

```
-                         "labels": [{"label": "High"}]
+                         "labels": [{"label": "High priority"}]
```

The rest LGTM!

Lucas Di Pentima wrote:

> Updates at arvados-workbench2|c4cc8cb0
> Test run: developer-tests-workbench2: #508 icon?job=developer-tests-workbench2&amp;build=508
>
> - Adds property key and value auto-selection with case-insensitive matching.
> - Expands test.

This LGTM.

#37 - 11/11/2021 05:07 PM - Lucas Di Pentima

Updates at e60cae2f8
Test run: developer-run-tests: #2785 icon?job=developer-run-tests&amp;build=2785

- Reverted changes in tests at c51e85a5

#38 - 11/11/2021 05:41 PM - Lucas Di Pentima

f1b121ccb - Removed a link to the 2.2 doc site on the upgrading notes so that tests pass.
Test run: developer-run-tests-doc-and-sdk-R: #927 icon?job=developer-run-tests-doc-and-sdk-R&amp;build=927

#39 - 11/11/2021 05:56 PM - Lucas Di Pentima

- % Done changed from 0 to 100

- Status changed from In Progress to Resolved

Applied in changeset arvados-workbench-2:arvados-workbench2|eabdb7bdd468b09c633ddb8b33fd8095ad27bb60.