keep.

Preserving the future

AIDA

Training Solr use cases

Luís Faria lfaria@keep.pt
Miguel Guimarães mguimaraes@keep.pt

Apache Solr access

Solr can be accessed at: https://scala.meemoo.be/solr/

It is protected with a basic authentication layer

Password were provided by KEEP Solutions

Documentation can be found at: Apache Solr Reference Guide 7.7

Solr

- **Dashboard**
- Logging
- Cloud
- Collections
- Java Properties
- Thread Dump
- Suggestions

Collection Sele... ▼

Core Selector

Instance

Start	about 16 hours ago

Versions

solr-spec	7.7.3
solr-impl	7.7.3 1a0d2a901dfec
lucene-spec	7.7.3
lucene-impl	7.7.3 1a0d2a901dfec

1. Open Collection Sellector

JVM

Runtime	Oracle Corporation O
Processors	4
Args	-DSTOP.KEY=solrrock
	-DSTOP.PORT=7983
	-Djetty.home=/opt/so
	-Djetty.port=8983
	-Dsolr.data.dir=/opt/s
	-Dsolr.data.home=/op
	-Dsolr.default.confdir
	-Dsolr.install.dir=/opt
	-Dsolr.jetty.https.port
	-Dsolr.log.dir=/opt/so
	-Dsolr.solr.home=/op
	-Duser.timezone=UT(
	-DzkClientTimeout=1
	-DzkRun
	-XX:+CMSParallelRer

Solr

Dashboard
Logging
Cloud
Collections
Java Properties
Thread Dump
Suggestions

Collection Sele...

DisposalConfir
mation
DIPFile
AIP
Risk
DIP
Represent
Preservat
nt
TransferredReso
urce

Instance

Start   about 16 hours ago

Versions

solr-spec   7.7.3
solr-impl   7.7.3 1a0d2a901dfec93676b0fe8be425101ceb754b85 -
lucene-spec   7.7.3
lucene-impl   7.7.3 1a0d2a901dfec93676b0fe8be425101ceb754b85 -

JVM

Runtime   Oracle Corporation OpenJDK 64-Bit Server VM 11.0.12 1
Processors   4
Args   -DSTOP.KEY=solrrocks
                   -DSTOP.PORT=7983
                   /solr/server
                   ot/solr/server/solr/mycores
                   e/opt/solr/server/solr/mycores
                   -Dsolr.default.confdir=/opt/solr/server/solr/configsets/_d
                   -Dsolr.install.dir=/opt/solr
                   -Dsolr.jetty.https.port=8983
                   -Dsolr.log.dir=/opt/solr/server/logs
                   -Dsolr.solr.home=/opt/solr/server/solr
                   -Duser.timezone=UTC
                   -DzkClientTimeout=15000
                   -DzkRun
                   -XX:+CMSParallelRemarkEnabled

2. Choose a collection
Example: AIP

Solr

- Dashboard
- Logging
- Cloud
- Collections
- Java Properties
- Thread Dump
- Suggestions

AIP ▼

- Overview
- Analysis
- Dataimport
- Documents
- Files
- Query
- Stream
- Schema

Core Selector ▼

Collection: AIP

Config name:	AIP
Max shards per node:	16
Replication factor:	1
Auto-add replicas:	🚫
Router name:	compositeId

3. Click Query

Solr

- Dashboard
- Logging
- Cloud
- Collections
- Java Properties
- Thread Dump
- Suggestions

AIP
- Overview
- Analysis
- Dataimport
- Documents
- Files
- Query
- Stream
- Schema

Core Selector

Request-Handler

/select

— common —

q

:

4. *:*

fq

sort

start, rows

0   10

fl

df

Raw Query Parameters

key1=val1&key2=val2

wt

☐ indent off
☐ debugQuery

☐ dismax
☐ edismax
☐ hl
☐ facet
☐ spatial
☐ spellcheck

5.Click Execute Query

Execute Query

https://scala.meemoo.be/solr/AIP/select?q=*%3A*

```
{
  "responseHeader":{
    "zkConnected":true,
    "status":0,
    "QTime":7,
    "params":{
      "q":"*:*",
      "_":"1639566314733"}},
  "response":{"numFound":66822,"start":0,"maxScore":1.0,"docs":[
    {
      "uuid":"ae717dcc-4e1a-4c85-972e-488585781676",
      "id":"ae717dcc-4e1a-4c85-972e-488585781676",
      "state":"ACTIVE",
      "permission_users_CREATE":["ingest-OR-w66976m",
        "admin",
        "evankeer"],
      "permission_users_READ":["admin",
        "evankeer"],
      "permission_users_UPDATE":["admin",
        "evankeer"],
      "permission_users_DELETE":["admin",
        "evankeer"],
      "permission_users_GRANT":["admin",
        "evankeer"],
      "permission_groups_CREATE":["Curator-meemoo",
        "administrators"],
      "permission_groups_READ":["Curator-meemoo",
        "administrators",
        "meemoo"],
      "permission_groups_UPDATE":["Curator-meemoo",
        "administrators"],
      "permission_groups_DELETE":["Curator-meemoo",
        "administrators"],
      "permission_groups_GRANT":["administrators"],
      "posalHoldStatus":false,
      "estors":["855a86f1-1463-4eed-bf11-70a47cfb4828"],
      "parentId":"855a86f1-1463-4eed-bf11-70a47cfb4828",
      "type":"OTHER",
      "createdOn":"2021-10-21T18:32:11Z",
```

Simple queries

keep.
Preserving the future

Collection: AIP

1. **EAD metadata**

   a. All text fields (search operator)

   b. All boolean or controlled vocabulary fields (dynamic fields, data types)

2. **Meemoo.xml metadata**

   a. Synchronisation AIP status

   b. Pruned / not pruned

   c. All text fields (search operator)

Collection: AIP

3.   AIP.json metadata

   a.   id

   b.   parentId

   c.   type

   d.   state

   e.   ingestJobId

   f.   hasRepresentation

   g.   createdOn datetime (before, after, on, between)

   h.   updatedOn datetime (before, after, on, between)

keep.
Preserving the future

Collection: File

1. **PREMIS object metadata**

   a. size (bigger than, smaller than, equals, between)

   b. formatDesignation

   c. formatVersion

   d. formatPronom: formatRegistry/formatRegistryKey

   e. formatMimetype: formatRegistry/formatRegistryKey

   f. originalName

Collection: Representation

1. **AIP.json metadata**
   a. id
   b. original (boolean facet)
   c. representationStates
   d. type
   e. createdOn datetime (before, after, on, between)
   f. updatedOn datetime (before, after, on, between)

Collection: PreservationEvent

1. **PREMIS event metadata**

   a. id

   b. eventType

   c. eventDateTime (before, after, on, between)

   d. eventOutcome

   e. linkingAgentIdentifierValue

   f. linkingObjectIdentifier (linkingSourceObjectIdentifier and linkingOutcomeObjectIdentifier)

Advanced queries

Advanced queries: Boolean operators

Operator	Description
AND	Requires both terms to be present for a match.
OR	Requires that either term (or both terms) be present for a match.
+	Requires that the following term be present.
-	Prohibits the following term (similar to what Google does)
NOT	Requires that the following term not be present.

Advanced queries

Demonstration

Exercises

keep.
Preserving the future

Exercise 1

For the collection **AIP**:

1. Search in all AIPs with string **'jkleevens'**

2. Show the facets for the field **'meemoo_sync_aip_status_txt'**

3. How many AIPs are pruned?

4. Search the AIP with id **'80ed7f6f-5e56-4c91-a9bd-5458e13154b3'** and return the value of the field **'title'**

5. How many AIPs with parentId **'de9eaa85-edb2-4bb3-aaca-cb3ec84a6e3f'**?

www.keep.pt

Exercise 1 (continuation)

6.   Show the facets for the field **'type'**

7.   How many AIPs are in the state **'ACTIVE'**?

8.   What is the ingestJobId that created the higher number of AIPs?

9.   Search all AIPs with representations (hasRepresentations)

10.   Search all AIPs created on December

11.   Search all AIPs created between November 10 and now

12.   Search all AIPs updated on **'2021-12-08T15:32:24Z'**

Exercise 2

For the collection **File**:

1. Search all Files bigger than 9000 bytes

2. How many files are between 19968 and 20000 bytes (inclusive)?

3. How many files are between 19968 and 20000 bytes (exclusive)?

4. How many 'Portable Network Graphics 1.0' files are?

5. Facet the files by formatPronom

6. Facet the files by formatMimetype

Exercise 2 (continuation)

7. What is the MIME type with most files?

8. Search the file with the originalName containing the keyword Heylighen. What is the aipId of the resulting query?

keep.

Preserving the future

Exercise 3

For the collection **Representation**:

1. Search all representations with id **'rep2'**

2. Facet the representations by original field

3. Face the representations by representationStates field and list its values

4. How many representations there are with the type **'MIXED'**?

5. Search all representations created on October

6. Search all representations updated today

www.keep.pt

keep.
Preserving the future

Exercise 4

For the collection **PreservationEvent**:

1. Search the preservation event with id
    'urn:roda:premis:event:76efb77d-1a80-4a1a-bc74-1b3906ef4b05'
2. Facet the preservation events by eventType field
3. How many preservation events there are of eventType **'message digest calculation'**?
4. How many preservation events have the eventOutcome **'PARTIAL_SUCCESS'**?

keep.

Preserving the future

Exercise 4 (continuation)

For the collection **PreservationEvent**:

5. Search all preservation events with linkingAgentIdentifier
    'linkingAgentIdentifier:"{\"type\":\"URN\",\"value\":\"urn:roda:premis:agent:ingest-OR-w66976m\",\"roles\":[]}"'

6. Search all preservation events with linkingOutcomeObjectIdentifier

7. **'{\"type\":\"URN\",\"value\":\"urn:roda:aip:00e3894e-ff3f-4cad-8ce9-afed74e50ca8\",\"roles\":[\"outcome\"]}'**

keep.
Preserving the future

Exercise 5

For the collection **AIP**:

1. Search all AIPs that are pruned and the parentId

    'ed3964e1-d8a7-4786-a594-8f87c08a90f5'

2. Search all AIPs that descriptiveMetadataId does not contain the string meemoo.xml

Exercise 6

For the collection **File**:

1. Search all files that contains the string Word in the formatDesignation and the size is bigger than 50000 (inclusive)

2. Search all files that contains the string 'pdf' in fileFormat however requiring to have an X

keep.

Preserving the future

Exercise 7

For the collection **PreservationEvent**:

1. Search all preservation events with the eventOutcome **'SUCCESS'** or **'PARTIAL_SUCCESS'**

2. Search all preservation events with the eventOutcome **'FAILURE'** and the eventType **'format identification'**