Appendix C - Migrate the Heritrix templates to NetarchiveSuite 3.6.0+

Contents

If you are just using the predefined templates with few changes like changed
the email-address and website information, the easiest way to migrate is to modify the
predefined templates found in the binary distribution of NetarchiveSuite in
the harvestdefinitionbasedir/order_templates_dist directory
and change the email-adress and website information again.

If you do this, you also get the more inconsequential updates to the template:

- The removal of obsolete attributes from some elements
- Addition of new attributes to some elements

Then you just update the existing templates in your database with these modified ones using the HarvestTemplateApplication tool mentioned in Appendix B - Managing Heritrix Harvest Templates (order.xml).
Note that some templates are no longer distributed with NetarchiveSuite. If you want to keep using
those, you need to follow the procedure described below.

If you have already put a lot effort in making your own templates, you can update your
existing templates by "only" upgrading the scope element in the templates from either a
DomainScope, HostScope, or a PathScope.

Before we explain how to migrate these scopes to a DecidingScope, you need to know something
about the anatomy of these scopes.

1) Header (includes scope class,and attributes):

```
<newObject name="scope" class="org.archive.crawler.scope.PathScope">
            <boolean name="enabled">true</boolean>
            <string name="seedsfile">seeds.txt</string>
            <boolean name="reread-seeds-on-config">true</boolean>
                   <integer name="max-link-hops">10</integer>
            <integer name="max-trans-hops">5</integer>
```

2) An OrFilter element named "exclude-filter" containing a number of filters as components:
a HopsFilter, a PathDepthFilter, a PathologicalPathFilter, a URIRegExpFilter, a URIListRegExpFilter (filter to avoid common crawlertraps),
and potentially other types of filters:
Each of these filters will have to be converted to a similar DecideRule. Explanation to follow.

```
            <newObject name="exclude-filter" class="org.archive.crawler.filter.OrFilter">
                <boolean name="enabled">true</boolean>
                <boolean name="if-matches-return">true</boolean>
                <map name="filters">
                    <newObject name="hops_filter" class="org.archive.crawler.filter.HopsFilter">
                        <boolean name="enabled">true</boolean>
                    </newObject>
                    <newObject name="pathdepth" class="org.archive.crawler.filter.PathDepthFilter">
                        <boolean name="enabled">true</boolean>
                        <integer name="max-path-depth">20</integer>
                        <boolean name="path-less-or-equal-return">false</boolean>
                    </newObject>
                    <newObject name="pathologicalpath" class="org.archive.crawler.filter.PathologicalPathFilter"
>
                        <boolean name="enabled">true</boolean>
                        <integer name="repetitions">3</integer>
                    </newObject>
                    <newObject name="dr_dk" class="org.archive.crawler.filter.URIRegExpFilter">
                        <boolean name="enabled">true</boolean>
                        <boolean name="if-match-return">true</boolean>
                        <string name="regexp">.*dr\.dk.*epg\.asp.*</string>
                    </newObject>
                    <newObject name="globale_crawlertraps" class="org.archive.crawler.filter.
URIListRegExpFilter">
                        <boolean name="enabled">true</boolean>
                        <boolean name="if-match-return">true</boolean>
                        <string name="list-logic">OR</string>
                        <stringList name="regexp-list">
                            <string>.*core\.UserAdmin.*core\.UserLogin.*</string>
                            <string>.*core\.UserAdmin.*register\.UserSelfRegistration.*</string>
                            <string>.*\/w\/index\.php\?title=Speci[ae]l:Recentchanges.*</string>
                            <string>.*act=calendar&amp;cal_id=.*</string>
                            .....
                            <string>.*calendar\.asp\?qMonth=.*</string>
                            <string>.*calendar\.php\?sid=.*</string>
                            <string>.*worldscinet\.com.*</string>
                            <string>.*www3\.interscience\.wiley\.com.*</string>
                            <string>.*www-gdz\.sub\.uni-goettingen\.de.*</string>
                        </stringList>
                    </newObject>
                </map>
            </newObject>
```

3) Additional filters. Here we have a "Force-accept-filter", an "additionalScopeFocus" filter, and a "transitive Filter", of which only the transitiveFilter element needs to be converted. The two other elements are just deleted.

```
            <newObject name="force-accept-filter" class="org.archive.crawler.filter.OrFilter">
                <boolean name="enabled">true</boolean>
                <boolean name="if-matches-return">true</boolean>
                <map name="filters">
                </map>
            </newObject>
            <newObject name="additionalScopeFocus" class="org.archive.crawler.filter.FilePatternFilter">
                <boolean name="enabled">true</boolean>
                <boolean name="if-match-return">true</boolean>
                <string name="use-default-patterns">All</string>
                <string name="regexp"/>
            </newObject>
            <newObject name="transitiveFilter" class="org.archive.crawler.filter.TransclusionFilter">
                <boolean name="enabled">true</boolean>
                <integer name="max-speculative-hops">1</integer>
                <integer name="max-referral-hops">15</integer>
                <integer name="max-embed-hops">15</integer>
            </newObject>
      </newObject> <!-- end of scope element -->
```

== How to convert from the former scopes to a decidingscope ==

Converting the header is easy.
All headers have the form:

```
 <newObject name="scope" class="org.archive.crawler.deciderules.DecidingScope">
            <boolean name="enabled">true</boolean>
            <string name="seedsfile">seeds.txt</string>
            <boolean name="reread-seeds-on-config">true</boolean>
            <!-- DecideRuleSequence. Multiple DecideRules applied in order with last non-PASS the resulting
decision -->
            <newObject name="decide-rules" class="org.archive.crawler.deciderules.DecideRuleSequence">

                <map name="rules">
                        <newObject name="rejectByDefault"
                                class="org.archive.crawler.deciderules.RejectDecideRule"/>
```

plus a special defining deciderule that emulates the DomainScope, the HostScope, or the PathScope.
1) The defining deciderule for DomainScope is (the only one using a special purpose DecideRule):

```
<newObject name="acceptURIFromSeedDomains" class="dk.netarkivet.harvester.harvesting.OnNSDomainsDecideRule">
                                <string name="decision">ACCEPT</string>
                                <string name="surts-source-file">seeds.txt</string>
                                <boolean name="seeds-as-surt-prefixes">false</boolean>
                                <string name="surts-dump-file"/>
                                <boolean name="also-check-via">false</boolean>
                                <boolean name="rebuild-on-reconfig">true</boolean>
</newObject>
```

2) The defining deciderule for HostScope is:

```
<newObject name="OnHostsRule" class="org.archive.crawler.deciderules.OnHostsDecideRule">
            <string name="decision">ACCEPT</string>
            <string name="surts-dump-file"/>
            <boolean name="also-check-via">false</boolean>
            <boolean name="rebuild-on-reconfig">true</boolean>
          </newObject>
```

3) The defining deciderule for PathScope is:

```
<newObject name="acceptIfSurtPrefixed" class="org.archive.crawler.deciderules.SurtPrefixedDecideRule">
                                <string name="decision">ACCEPT</string>
                                <string name="surts-source-file"></string>
                                <boolean name="seeds-as-surt-prefixes">true</boolean>
                                <string name="surts-dump-file"></string>
                                <boolean name="also-check-via">false</boolean>
                                <boolean name="rebuild-on-reconfig">true</boolean>
</newObject>
```

After the header and the defining deciderule, we a deciderule corresponding to the 'hops_filter'.
Note that the two last attributes 'max-link-hops', and 'max-trans-hops' in the header cease to be general scope attributes. Instead max-trans-hops become an attribute for the "acceptIfTranscluded" mentioned above, and the 'max-link-hops' attribute becomes an attribute for the new 'hops_filter' deciderule.
The following

```
<integer name="max-link-hops">10</integer>
<newObject name="hops_filter" class="org.archive.crawler.filter.HopsFilter">
        <boolean name="enabled">true</boolean>
</newObject>
```

is then translated to the following deciderule

```
<newObject name="rejectIfTooManyHops" class="org.archive.crawler.deciderules.TooManyHopsDecideRule">
        <integer name="max-hops">10</integer>
</newObject>
```

Following this, we need to add a translation of the 'pathdepth' element, and the 'pathologicalpath' element, plus a translation of the 'transitiveFilter' element in the last part of the scope.
The following

```
     <newObject name="pathdepth" class="org.archive.crawler.filter.PathDepthFilter">
                         <boolean name="enabled">true</boolean>
                         <integer name="max-path-depth">20</integer>
                         <boolean name="path-less-or-equal-return">false</boolean>
     </newObject>
     <newObject name="pathologicalpath" class="org.archive.crawler.filter.PathologicalPathFilter">
              <boolean name="enabled">true</boolean>
        <integer name="repetitions">3</integer>
     </newObject>

     <newObject name="transitiveFilter" class="org.archive.crawler.filter.TransclusionFilter">
                <boolean name="enabled">true</boolean>
                <integer name="max-speculative-hops">1</integer>
                <integer name="max-referral-hops">15</integer>
                <integer name="max-embed-hops">15</integer>
     </newObject>
```

is translated to

```
<newObject name="rejectIfPathological" class="org.archive.crawler.deciderules.PathologicalPathDecideRule">
        <integer name="max-repetitions">3</integer>
</newObject>
<newObject name="acceptIfTranscluded" class="org.archive.crawler.deciderules.TransclusionDecideRule">
        <integer name="max-trans-hops">5</integer>
        <integer name="max-speculative-hops">1</integer>
</newObject>
<newObject name="pathdepthfilter" class="org.archive.crawler.deciderules.TooManyPathSegmentsDecideRule">
        <integer name="max-path-depth">20</integer>
</newObject>
```

.

Note that the attributes 'max-referral-hops' and 'max-embed-hops' in the 'transitiveFilter' element have been merged into one single attribute 'max-trans-hops' which is now no longer an attribute of the scope, as it was in the old scopes.

Now you only need to convert all remaining URIRegExpFilter and URIListRegExpFilter elements to a corresponding DecideRule. The deciderule corresponding to URIRegExpFilter is MatchesRegExpDecideRule, and the deciderule corresponding to URIListRegExpFilter is MatchesListRegExpDecideRule.
Converting the dr_dk element (a URIRegExpFilter)

```
<newObject name="dr_dk" class="org.archive.crawler.filter.URIRegExpFilter">
                        <boolean name="enabled">true</boolean>
                        <boolean name="if-match-return">true</boolean>
                        <string name="regexp">.*dr\.dk.*epg\.asp.*</string>
                    </newObject>
```

gives us:

```
<newObject name="dr_dk" class="org.archive.crawler.deciderules.MatchesRegExpDecideRule">
        <string name="decision">REJECT</string>
        <string name="regexp">.*dr\.dk.*epg\.asp.*</string>
 </newObject>
```

Converting the globale_crawlertraps element (URIListRegExpFilter)

```
<newObject name="globale_crawlertraps" class="org.archive.crawler.filter.URIListRegExpFilter">
                        <boolean name="enabled">true</boolean>
                        <boolean name="if-match-return">true</boolean>
                        <string name="list-logic">OR</string>
                        <stringList name="regexp-list">
                            <string>.*core\.UserAdmin.*core\.UserLogin.*</string>
                            <string>.*core\.UserAdmin.*register\.UserSelfRegistration.*</string>
                            <string>.*\/w\/index\.php\?title=Speci[ae]l:Recentchanges.*</string>
                            <string>.*act=calendar&amp;cal_id=.*</string>
                            .....
                            <string>.*calendar\.asp\?qMonth=.*</string>
                            <string>.*calendar\.php\?sid=.*</string>
                            <string>.*worldscinet\.com.*</string>
                            <string>.*www3\.interscience\.wiley\.com.*</string>
                            <string>.*www-gdz\.sub\.uni-goettingen\.de.*</string>
                        </stringList>
                    </newObject>
```

gives us

```
<newObject name="globale_crawlertraps" class="org.archive.crawler.deciderules.MatchesListRegExpDecideRule">
            <string name="decision">REJECT</string>
            <string name="list-logic">OR</string>
            <stringList name="regexp-list">
                <string>.*core\.UserAdmin.*core\.UserLogin.*</string>
                <string>.*core\.UserAdmin.*register\.UserSelfRegistration.*</string>
                <string>.*\/w\/index\.php\?title=Speci[ae]l:Recentchanges.*</string>
                <string>.*act=calendar&amp;cal_id=.*</string>
                .....
                <string>.*calendar\.asp\?qMonth=.*</string>
                <string>.*calendar\.php\?sid=.*</string>
                <string>.*worldscinet\.com.*</string>
                <string>.*www3\.interscience\.wiley\.com.*</string>
            </stringList>
</newObject>
```

Finally we need to wrap up the the sequence of deciderules and the scope itself.
So we add

```
                </map> <!-- end rules -->
            </newObject> <!-- end decide-rules  -->
        </newObject> <!-- End DecidingScope -->
```

.