RSS Harvests

From NAS 5.2 onwards, it is possible to harvest RSS feeds using the Crawler RSS (https://github.com/Landsbokasafn/crawlrss) module developed by Kristinn Sigurðsson at the National Library of Iceland. In order to use the module it is necessary to configure the feeds to be harvested in a special crawler-bean template. At present it is not possible to define the seeds of an RSS harvest directly through the NAS GUI. A sample template suitable for use with NAS can be downloaded from https://raw.githubusercontent.com/netarchivesuite/crawlrss/master/src/main/conf/jobs/CrawlRSS-Sample-Profile/netarkivet-crawlrss.dr.dk.cxml.

The template can be customised by replacing this section

```
<list>
  <bean class="is.landsbokasafn.crawler.rss.RssFeed">
    <property name="uri" value="http://www.dr.dk/nyheder/service/feeds/indland" /> <!--RSS url -->
    <property name="impliedPages">
      <list>
        <value>https://www.dr.dk/nyheder/</value>
        <value>http://www.dr.dk/nyheder/allenyheder/indland</value> <!-- Landing Page -->
      </list>
    </property>
  </bean>
  <bean class="is.landsbokasafn.crawler.rss.RssFeed">
    <property name="uri" value="http://www.dr.dk/nyheder/service/feeds/udland" /> <!--RSS url -->
    <property name="impliedPages">
      <list>
        <value>http://www.dr.dk/nyheder/allenyheder/udland</value>
      </list>
    </property>
  </bean>
  <bean class="is.landsbokasafn.crawler.rss.RssFeed">
    <property name="uri" value="http://www.dr.dk/nyheder/service/feeds/penge" /> <!--RSS url -->
    <property name="impliedPages">
      <list>
        <value>http://www.dr.dk/nyheder/allenyheder/penge</value>
      </list>
    </property>
  </bean>
</list>
```

with your own list of feeds to be harvested. Associated with each rss-feed uri is a list of implied pages. These can be ordinary html landing pages associated with the feed. By harvesting these together with the rss-feed one can ensure a consistent browsing experience in the harvested data.

To use the rss-template one needs to define, for any domain, a configuration with an empty seed list. Strictly speaking, seed lists cannot be completely empty, but a seed list can consist solely of a single comment character “#”. Then simple define a harvest configuration using the crawlrss template together with the empty seed list.