RSS Harvests

The Crawler RSS (https://github.com/Landsbokasafn/crawlrss) module developed by Kristinn Sigurðsson at the National Library of Iceland was introduced in NAS 5.2 but has been non-functional for some time. From version 7.4 it is now functioning again. 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

```xml
<list>
  <bean class="is.landsbokasafn.crawler.rss.RssFeed">
    <property name="uri" value="https://www.dr.dk/nyheder/service/feeds/indland"/>
    <property name="impliedPages">
      <list>
        <value>https://www.dr.dk/nyheder/</value>
        <value>http://www.dr.dk/nyheder/allenyheder/indland</value>
      </list>
    </property>
  </bean>
  <bean class="is.landsbokasafn.crawler.rss.RssFeed">
    <property name="uri" value="https://www.dr.dk/nyheder/service/feeds/udland" />
    <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="https://www.dr.dk/nyheder/service/feeds/penge" />
    <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. Note that the module does not follow redirects on rss uris, so these must be the actual uris for the resource.

To use the rss-template one needs to define, for any domain, a configuration using this template. In NetarchiveSuite every configuration must have at least one seed, but for these rss templates the seed(s) will be ignored and replaced with the non-existent url "http://foo.invalid". Then simply define a harvest configuration using the crawlrss template and use it in a harvest.