Handout XML2RDF-Workshop

DH2019 Utrecht, 08.07.2018

# Querying and Aggregating with XPath

* XPath queries nodes within the XML-tree based on **paths** and **axis**. The XPath subsequently returns found nodes, attribute values, and textual contents.
  + A path query’s **arbitrary starting** point is notated with **//**
  + Sibling-nodes are separated by **/** thus notated with //node1/node2
  + **Conditions** within the path query are encapsulated by **[ ]**
  + **Attribute values** are queried with **@Attribute**
  + Attributes as parts of a condition within a path query are notated with //node[@Attribute=“Value“]
  + Namespaces are mostly notated as a prefix: **tei:**placeName
* **Note: Namespaces may be specified additionally!**

# Important XPath Functions

* starts-with(@key, 'W.')
* matches(@source, '/gnd/')
* replace(@source,'http://d-nb.info/gnd/(.\*?)', '$1', '')
* not(matches(@source, '/gnd/'))
* tokenize(tei:location/tei:geo, ' ')
* string-join(//persName, ' – ')

# XPath Practice Sheet

Download: <http://www.benedictus.mgh.de/openmgh/bsb00000824.zip>

|  |  |
| --- | --- |
| Ausdruck | Selektion |
| //Q{http://www.tei-c.org/ns/1.0}lb | Declaration of the TEI namespace for lb |
| //text/body/p | Find all paragraphs (p) in path /TEI/text/body |
| //text/body/p[1] | Find first paragraph (p) in path /TEI/text/body |
| //text/body/p[@type='first'] | Find first paragraph (p) with attribute (type='first') in path /TEI/text/body |
| //body//p/child::\* | Find all child nodes of node p |
| //body//p/child::text() | Find all textual child nodes of node p |