“Jats to RDF”- transformation. Adaptation of Peroni’s stylesheet to PsychOpen, differences to Biotéa and Done log

- Peroni’s “Jats to RDF”-xsl stylesheet available at http://purl.org/spar/jats2rdf. Corresponding document see: http://purl.org/spar/jats2rdf

Changes that have been applied to “Jats to RDF” to meet the requirements of describing PsychOpen articles

- Adapted stylesheet to receive unique URIs.
    - In the original stylesheet, URIs are not unique:
      For example, a URI that describes an article is
      `rdf:about="textual-entity">`.
      Changed this to:
      `rdf:about="zpid:doi/10.5964/ejop.v8i3.308/textual-entity">`
      I.e URIs whose scope is a specific article have an Infix with the article’s doi to disambiguate them from URIs belonging to another article (-> similar to biotéa URI handling).
      URIs whose scope is smaller than the article, e.g. where the scope is a specific article reference, have an additional URI infix. For example:

      rdf:about=""zpid:doi/10.5964/ejop.v8i3.308/reference-46/textual-entity , where the “reference”-Uri infix serves to disambiguate.

      The original stylesheet used random numbers here, but the resulting URIs are not unique across different rdf-files and are not human-friendly.

      Care has been taken, where possible, to adapt URIs with a larger scope than one specific article in such a way that they share the same URI across all rdf documents:

      `rdf:resource="zpid:journal/Europe%27sJournalofPsychology"/>`

      `rdf:about="zpid:/book/Handbookofclinical%0Aneurology">`

      This was done for publications by using the publication name as part of the URI (same URI-style as in Biotéa), but not for authors (as in biotéa), since disambiguation of persons is not possible on the basis of the given information.

      Of course, the same disambiguation problem exists for publications, since there might exist publications of the same type (e.g. journal, book or thesis) that share the same title, so this is not the best solution.

Added information which exist in Biotéa but not in Peroni’s stylesheet

- External links:
    - Owl:sameAs to doi-organization’s URL that links to the article’s web location
    - rdfs:seeAlso to pubmed-ncbi-URL that links to the article’s web location
    - owl:sameAs to pubmed-identifier-URL that links to the article’s web location
    - rdfs:seeAlso to nlm-catalog to link a journal to its NLM-catalog entry.
    - Rdfs:seeAlso to link to the publisher (http://www.psychopen.eu/)
- The real publisher (used “<foaf:name>PsychOpen”) and “<prov:wasAttributedTo> [Agent of type Organization with label “PsychOpen”] “

In the original sheet, on encountering the attribute "@pub-id-type[. = 'publisher-id'], the placeholder „A publisher“ is used.

- Added template that matches @publication-type=”web” to rdfize web references. Used Fabio:WebPage for conceptualization, because most links occur in citations and refer to a particular web link the citation has been taken from. Alternative concepts would be: Fabio:WebManifestation or Fabio:WebSite
- Corrected: article-type “disertation” -> “dissertation”
- Removed function “getLabelForId” which was used by the original to identify publishers from the content of the journal-id-type. This used to generate multiple publishers per article
- Added @publication-type[. = 'web'

Added information which does not exist in Biotéa

- Added 2 templates that match attribute values: `"@contrib-type=”editor”` and `"@contrib-type=”translator”.` The templates use spar concepts to describe editors and translators (in orig. stylesheet only authors are matched as “contributors”, editors and translators as “collaborators” only, however, in PsychOpen the “contributor”-attribute is used. In biotéa, bibo:editorList is used.
- Correspondents’ email address (foaf:mbox)

Other changes / bug fixes to the original Jats2Spar stylesheet

- Removed some templates that process JATS-tags not currently used in PsychOpen. If such a tag is encountered during processing, an error message is generated in the log-file, starting with the string “TRANS:”. Thus, should the JATS tag occur in future, the template can be added and adopted to PsychOpen Uri-style. As an example, see the following adapted template:

```
<xsl:template match="addr-line">
  <xsl:message>TRANS: unprocessed address-line tag encountered</xsl:message>
</xsl:template>
```

- Transformation with the original sheet crashed during processing of JATs-tag “<abbrev-journal-title>. Solution which works for PsychOpen: removed the restriction “`not(@abbrev-type)]">` from the matching condition: `<!--xsl:template match="abbrev-journal-title[not(@abbrev-type)]">`
- No triples to state the “ending Page” were generated, although “lpage”-handling exists in the stylesheet. Solution: adapted the entry-point for the template. Used to be:
  `<xsl:if test="preceding-sibling::lpage">`
  Now:
  `<xsl:if test="self::lpage">`
- Template “affiliation”: Previously, all text within the affiliation node had been included into the affiliation name, including the text of child node “label”. Label in PsychOpen is something like an alias or footnote, e.g. “[a]”. Added a function which excludes the label text from the affiliation name if it is present.
- Publisher: currently hard-coded as “PsychOpen” + link to PsychOpen uri
- Removed “datacite identifier” (data was not generated correctly, a commonplace literal was assigned depending of value of journal.id-type attribute. In PsychOpen, this attribute was used to merely provide alternative abbreviated journal titles (e.g. a different one for pmc than for "nlm-ta", and not to specify the publisher (as intended in jats2spar mapping))
- Removed template that matches <product> and relate product-type attribute matching templates.

    - Explanation: In the original sheet, some triples that do not make sense are generated due to the generic nature of the product template: For example, if the product is a book, the publisher is used as the product name, e.g. “foaf:name "Oxford University Press"”: In the xml, this info was provided within <publisher-name> tag and was supposed to relate to the described product, i.e. the book instance. The <publisher-location> was rdfized to product ..vCard - vCard:locality. 4) product partOf source. source dcterms:title [booktitle].-> The product should be a a materialization of the ConceptualWork "book". The <source> value is merely the <book>-title. -> Might be hard to generalize rdfization of products, because product types have to be marked up differently. Maybe too many degrees of freedom in JATS specification for product description.
    - Motivation for removal -> Rarely used tag, too much work to implement, uncertain outcome, maybe little value (?)
- Other removals: (if tags or attributes removed are encountered in the text, this is logged. Items are Always tags unless marked otherwise
    - <@publication-format>
    - <address>
    - <addr-line>
    - <anonymous>
    - Author-comment
    - <award-id>
    - Conference
    - Conf-acronym
    - Conf-date
    - Conf-name
    - Conf-num
    - Conf-log
    - Conf-sponsor
    - Conf-theme
    - Contrib-id
    - Copyright-holder
    - Copyright-statement
    - Date-in-citation
    - Degree
    - Fax
    - Funding-source
    - Gov
    - Institution
    - Issue-id
    - Issue-title
    - Issue-sponsor
    - Issue-part
    - Part-title
    - Phone
    - Prefix
    - Principle-award-recipient
    - Principal-investigator
    - Related-article
    - Related-object

    - Season
    - Self-uri
    - Std
    - Std-organization
    - Supplementary-material
    - Uri
    - Volume-id
    - Volume-series
    - @calendar
    - @collab-type
    - @contrib-id-type
    - @date-type[. = 'pub']
    - @equal-contrib
    - @mimetype
    - @publication-format
    - "@publication-type[. = 'standard']
    - @publication-type[. = 'working-paper']
    - @pub-id-type[. = 'doaj']
    - @pub-id-type[. = 'isbn']
    - @pub-id-type[. = 'manuscript']
    - @pub-id-type[. = 'sici']
    - @pub-type[. = 'epub-ppub']
    - @pub-type[. = 'epreprint']
    - @pub-type[. = 'ppreprint']
    - @pub-type[. = 'ecorrected']
    - @pub-type[. = 'pcorrected']
    - @pub-type[. = 'eretracted']
    - @pub-type[. = 'pretracted']
    -
- Added email of <corresp> (correspondent) (foaf:mbox)
- Publisher: at the moment we add hard-coded
- Different textual entities were generated to describe dates, instead of using the article's textual entity. Motivation for using the article's textual entity:
    - Although different versions of the text might be interpreted as separate textual entities, the dates such as "date accepted", "date corrected" refer imho to the same entity: Once a version is changed, it becomes a different version. Thus, if the separate entities refer to text versions, a version can never have a correction date.
- Refined check in template "create-structure". The check was intended by Peroni to check language information attached to object literals. But it also captured email-addresses. Fixed this.

Testing

- Created a "reference article" that covers all tags used by PsychOpen I could find. Stored in java project jats2spar as: /src/main/resources//jats2spar/src/main/resources/psychOpen_testArticle.xml

TODO

- See also http://chrismaloney.org/notes/_diff/JATS2RDF?to=fb7b9307e31ec7e0120de29c1742c90e8f15421
- TODO authors' biographies are not linked to the author in Peroni's sheet and zpid sheet.
- TODO volume and issue are not linked to the journal, only to the paper (but journal can be determined by a SPARQL query involving the paper, since journal:paper is a 1:1 relation)
- TODO article authors are not stored in ordered lists in the rdf (as opposed to authors of references). Thus it is not possible to query for e.g. the first author of an article.
- Publisher as user-configurable passed in value?