Digital Muqtabas Integration in CLARIN using Canonical Text Service

CLARIN Annual Conference 2017

Till Grallert, Jochen Tiepmar, Thomas Eckart
Dirk Goldhahn, Christoph Kuras

Agenda

- (Brief) Overview about Canonical Text Service (CTS)
- CTS CLARIN Interface
- Digital Muqtabas
- Review Feedback

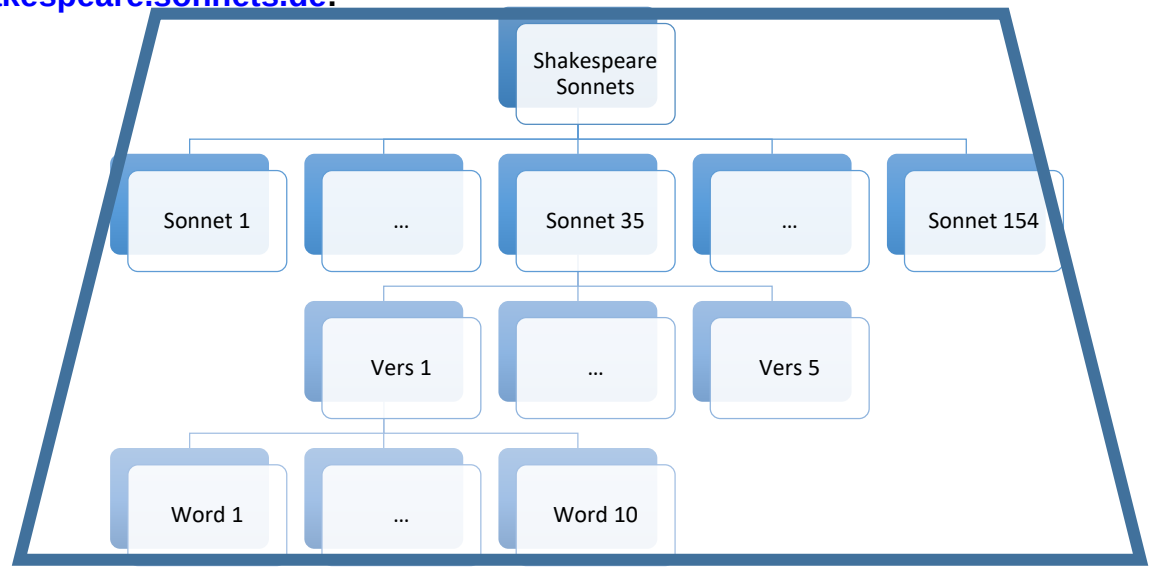
Overview CTS

Canonical Text Services (CTS)

- protocol for a webbased citable text service
- Unique Identifiers(**U**nique **R**esource **N**ame, **URN**) refer to text passages and text parts
- Developed in Homer Multitext Project(www.homermultitext.org), Smith et.al.2009
  http://www.homermultitext.org/hmt-docs/specifications/ctsurn/
  http://www.homermultitext.org/hmt-docs/specifications/cts/
- This implementation was done in Billion Words Project (ESF)

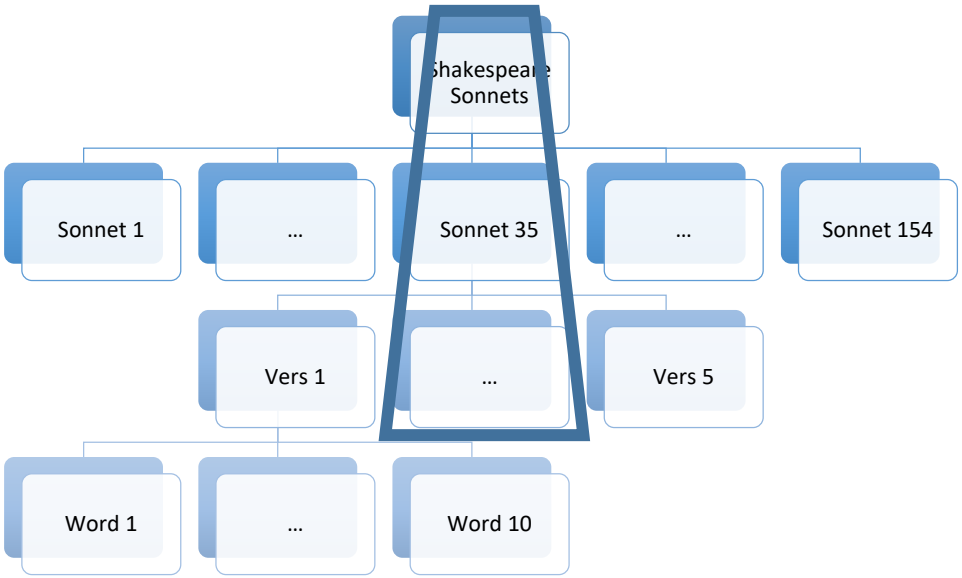
CTS URNs

urn:cts:demo:shakespeare.sonnets:
urn:cts:demo:shakespeare.sonnets.de:



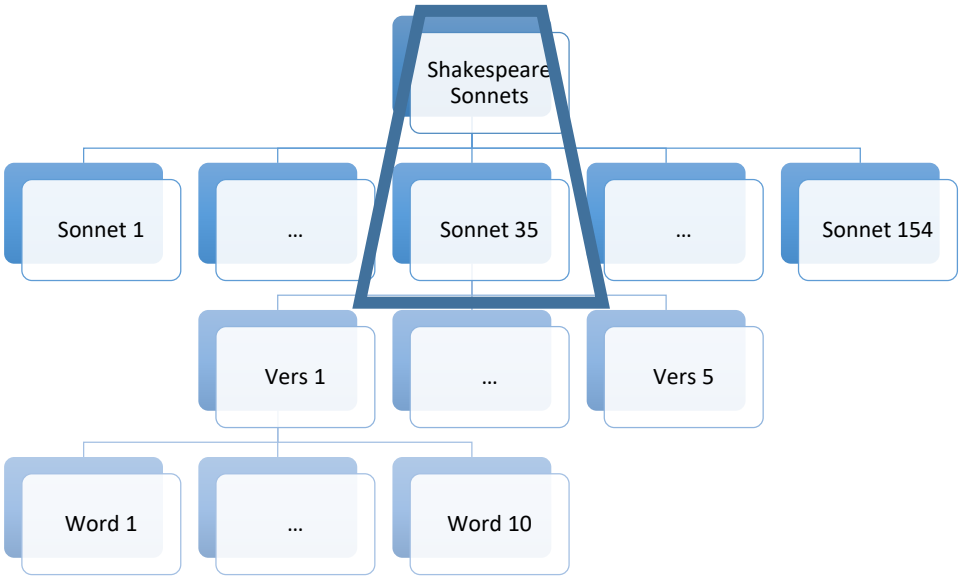
CTS URNs

urn:cts:demo:shakespeare.sonnets:35.4



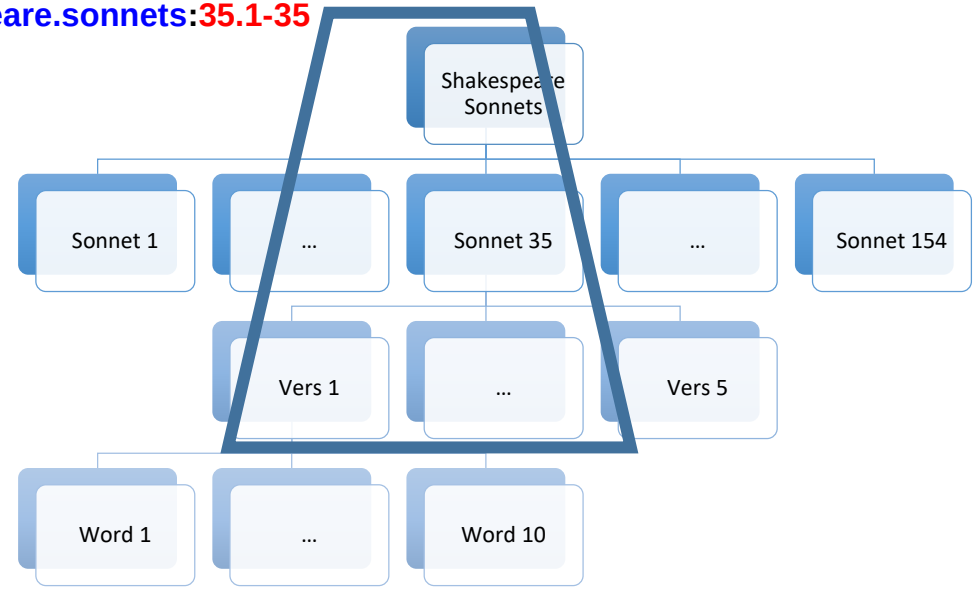
CTS URNs

urn:cts:demo:shakespeare.sonnets:35



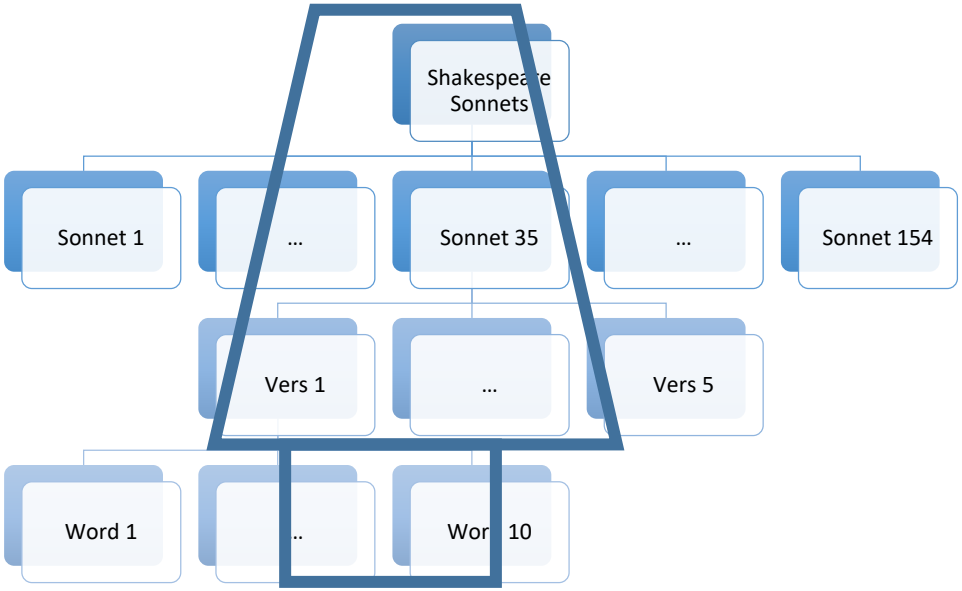
CTS URNs

urn:cts:demo:shakespeare.sonnets:35.1-35.5
urn:cts:demo:shakespeare.sonnets:35.1-35



CTS URNs

urn:cts:demo:shakespeare.sonnets:35.1@grieved-35.5@faults[1]



CTS Request

http://cts.informatik.uni-leipzig.de/pbc/cts/?request=GetPassage&urn=urn:cts:pbc:bible.parallel.eng.kingjames:1.3.5-1.3.8@the[2]


```
<GetPassage>
 - <request>
     <requestName>GetPassage</requestName>
   - <requestUrn>
       urn:cts:pbc:bible.parallel.eng.kingjames:1.3.5-1.3.8@the[2]
     </requestUrn>
   </request>
 - <reply>
   - <urn>
       urn:cts:pbc:bible.parallel.eng.kingjames:1.3.5-1.3.8@the[2]
     </urn>
   - <passage>
       For God doth know that in the day ye eat thereof , then your eyes shall be opened , and ye shall be as gods , knowing good and evil . And when
       the woman saw that the tree was good for food , and that it was pleasant to the eyes , and a tree to be desired to make one wise , she took of the
       fruit thereof , and did eat , and gave also unto her husband with her ; and he did eat . And the eyes of them both were opened , and they knew
       that they were naked ; and they sewed fig leaves together , and made themselves aprons . And they heard the
     </passage>
     <license>Public Domain</license>
   - <source>
       Retrieved via Canonical Text Service http://cts.informatik.uni-leipzig.de/pbc/cts/ with CTS URN
       urn:cts:pbc:bible.parallel.eng.kingjames:1.3.5-1.3.8@the[2]
     </source>
   </reply>
</GetPassage>
```

CTS CLARIN Interface

- Presentation in VLO:

urn:cts:pbc:bible

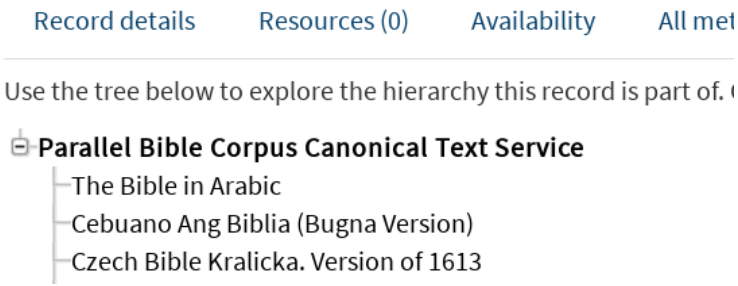
urn:cts:pbc:bible.parallel.arb.norm:

urn:cts:pbc:bible.parallel.ceb.bugna:

urn:cts:pbc:bible.parallel.ces.kralicka:

...

CMDI 1.2 compliant metadata



CTS CLARIN Interface

VLO / Faceted search / Record: مجلة المقتبس Majallat al-Muqtabas TEI edition

مجلة المقتبس Majallat al-Muqtabas TEI edition

Show the original provider's page for this record

Record details | Resources (15) | Availability | All metadata | Technical details | Hierarchy

Name	Type
cts	other
cts	other
cts	other
cts	other
cts	other
cts	other
cts	other
cts	other
cts	other

CTS CLARIN Interface

VLO / Faceted search / Record: مجلة المقتبس Majallat al-Muqtabas TEI edition

مجلة المقتبس Majallat al-Muqtabas TEI edition

Show the original provider's page for this record

Record details | Resources (15) | Availability | All metadata | Technical details | Hierarchy

Name		Type
cts	urn:cts:muqtabas:oclc.4770057679_i.22.TEIP5:	other
cts	urn:cts:muqtabas:oclc.4770057679_i.22.TEIP5:1	other
cts	urn:cts:muqtabas:oclc.4770057679_i.22.TEIP5:2	other
cts	urn:cts:muqtabas:oclc.4770057679_i.22.TEIP5:3	other
cts	urn:cts:muqtabas:oclc.4770057679_i.22.TEIP5:4	other
cts	urn:cts:muqtabas:oclc.4770057679_i.22.TEIP5:5	other
cts	urn:cts:muqtabas:oclc.4770057679_i.22.TEIP5:6	other
cts	urn:cts:muqtabas:oclc.4770057679_i.22.TEIP5:7	other
cts	urn:cts:muqtabas:oclc.4770057679_i.22.TEIP5:8	other

Digital Muqtabas

al-Muqtabas / المقتبس
“monthly” journal published by Muḥammad Kurd ʿAlī between 1906 and 1918/19 in Cairo and, from 1908 onwards, in Damascus.

- 9 volumes, 96 issues , c. 7000 pages
- 3.851.614 tokens (words), 5042 articles, 136 named authors

•available at c. 30 libraries (North America, Europe, Middle East):

- original prints (mostly incomplete)
- some copies of a “gray” reprint
- a number of microfiche copies from a single source

Importance of mundane texts / periodicals

They are at the core of various discourses

- Modernity / -ism at the end of empire
- Arabic renaissance
- Arab nationalism
- Islamic reform movement

They form large corpora with an equal distribution along a temporal axis (*al-Muqtabas*: 12yrs, *al-Manār*: 43 yrs, *al-Muqtaṭaf*: 76 yrs)

- linguistic analysis
- historical semantics
- rich data sets for social history

Digital *Muqtabas* (2015-): unite transcription and facsimiles

aims

- **validate** the transcription against the facsimiles
- **improve** the transcription with the help of the “crowd”
- make everything **citable** for scholars, **linkable** for machines
- provide the new edition with the broadest possible licence to facilitate access and re-use

principles

- re-purpose **available** and **established** tools, technologies, and material
- preference for **open** and **simple** formats and tools

Digital *Muqtabas* (2015-): deliverables

open scholarly digital editions of *[Majallat] al-Muqtabas* providing

- TEI XML files (transcription and links to facsimiles)
- plain text files (markdown)
- MODS and BibTeX files for every article
- customised version of TEI Boilerplate (XSLT and CSS) with stable URLs for every element (example)
- access to bibliographic metadata through a public Zotero group (OpenArabicPE/digital-muqtabas)

within a framework (git, GitHub, Zenodo) that allows for / provides

- collaborative, open, version-controlled improvements of the edition
- re-use through open licences:CC0 (text, metadata), CC BY-SA 4.0 (edition), and MIT licence (tools)
- long-term preservation and DOIs (10.5281/zenodo.597319)

Digital *Muqtabas* (2015-): current state

Project evolved into “Open Arabic Periodical Editions” (OpenArabicPE)
Editorial decisions: modelling / TEI schema design
Editorial work:

- mark-up of page breaks (1-2 h per issue)
- add parts missing from transcription (all foreign words and footnotes, entire pages)
- correct transcription
- correct publication dates for all issues.

Web-display:

- needs some polishing
- search functions beyond the Zotero group and individual issues

Review Feedback

The references to important points (like that the edition is available on GitHub or how to actually use the services) are difficult to find

Information about CTS
http://cts.informatik.uni-leipzig.de/

Information about Digital Muqtabas
https://openarabicpe.github.io/slides/2017-dig-eg-gaz/
https://github.com/tillgrallert/digital-muqtabas

Review Feedback

An issue not mentioned at all is how the project will deal with copyright as some of 136 identified contributors to the journal will have died less than 70 years ago. Kurd Ali himself died 64 years ago.
All the shamela resources are notorious for IPR problems as it is unclear of large amounts of data therein where they came from.


```
− <edition urn="urn:cts:muqtabas:oclc.4770057679_i.1.TEIP5:">
  − <title>
      مجلة المقتبس Majallat al-Muqtabas, Vol. 1, no.1 TEI edition
    </title>
    <author>محمد كرد علي Muḥammad Kurd ʿAlī</author>
  − <license>
      Distributed under a Creative Commons Attribution-ShareAlike 4.0
    </license>
  − <source>
      https://github.com/tillgrallert/digital-muqtabas/blob/master/xml/oc
      Canonical Text Service http://cts.informatik.uni-leipzig.de/muqtab
    </source>
    <publicationDate>2015</publicationDate>
    <language>arb</language>
    <contentType>xml</contentType>
  </edition>
```

```
<GetPassage>
 − <request>
    <requestName>GetPassage</requestName>
    <requestUrn>urn:cts:muqtabas:oclc.4770057679_i.44.TEIP5:2.1</requestUrn>
  </request>
 − <reply>
    <urn>urn:cts:muqtabas:oclc.4770057679_i.44.TEIP5:2.1</urn>
  − <passage>
      الرسالة العذراءمنقولة من مجموع في موازين البلاغة وأدوات الكتابة كتب بها أبو اليسر ابراهيم بن محمد بن المدبر
    </passage>
  − <license>
      Distributed under a Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0) license
    </license>
  − <source>
      https://github.com/tillgrallert/digital-muqtabas/blob/master/xml/oclc_4770057679-i_44.TEIP5.xml retrieved via Canonical
      Text Service http://cts.informatik.uni-leipzig.de/muqtabas/cts/ with CTS URN
      urn:cts:muqtabas:oclc.4770057679_i.44.TEIP5:2.1
    </source>
  </reply>
</GetPassage>
```

Review Feedback

What is to be made clear is that here we are dealing with the integration of static CTS; not the kind of technology that allows you to cite an arbitrary passage (which is better left out of the VLO) but of the identification of existing and identifiable parts of a resource.

Static URNs

- Document
  - urn:cts:pbc:bible.parallel.eng:
  - urn:cts:pbc:bible.parallel.eng.kingjames:
- Text part
  - urn:cts:pbc:bible.parallel.eng:1
  - urn:cts:pbc:bible.parallel.eng.kingjames:1.3.2

Dynamic URNs

- Text span (From one text part to another)
  - urn:cts:pbc:bible.parallel.eng:1.2-1.5.6
- Sub passage notation
  - urn:cts:pbc:bible.parallel.eng:1.2@the[2]-1.5.6@five

Review Feedback

it would make sense also to have a Web page giving (pointers to) the complete collection and have it represented in straight-forward HTML. Why was this not done / considered? Pls. explain.

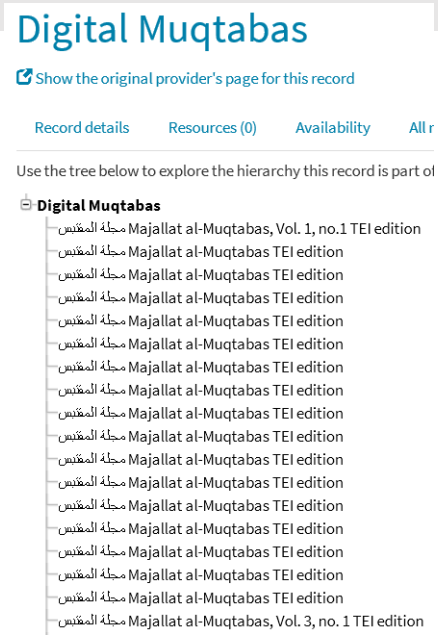
GetCapabilities	The text inventory with every CTS URN on document level and the corresponding meta information.

http://cts.informatik.uni-leipzig.de/muqtabas/cts/?request=GetCapabilities

Review Feedback

the VLO gives only a list of titles » مجلة المقتبسMajallat al-Muqtabas TEI edition« i.e. without any volume or number information, and this extensive listing of identical titles is quite useless. This should be fixed, otherwise the whole point is really lost.

Titles are based on the TEI/XML files



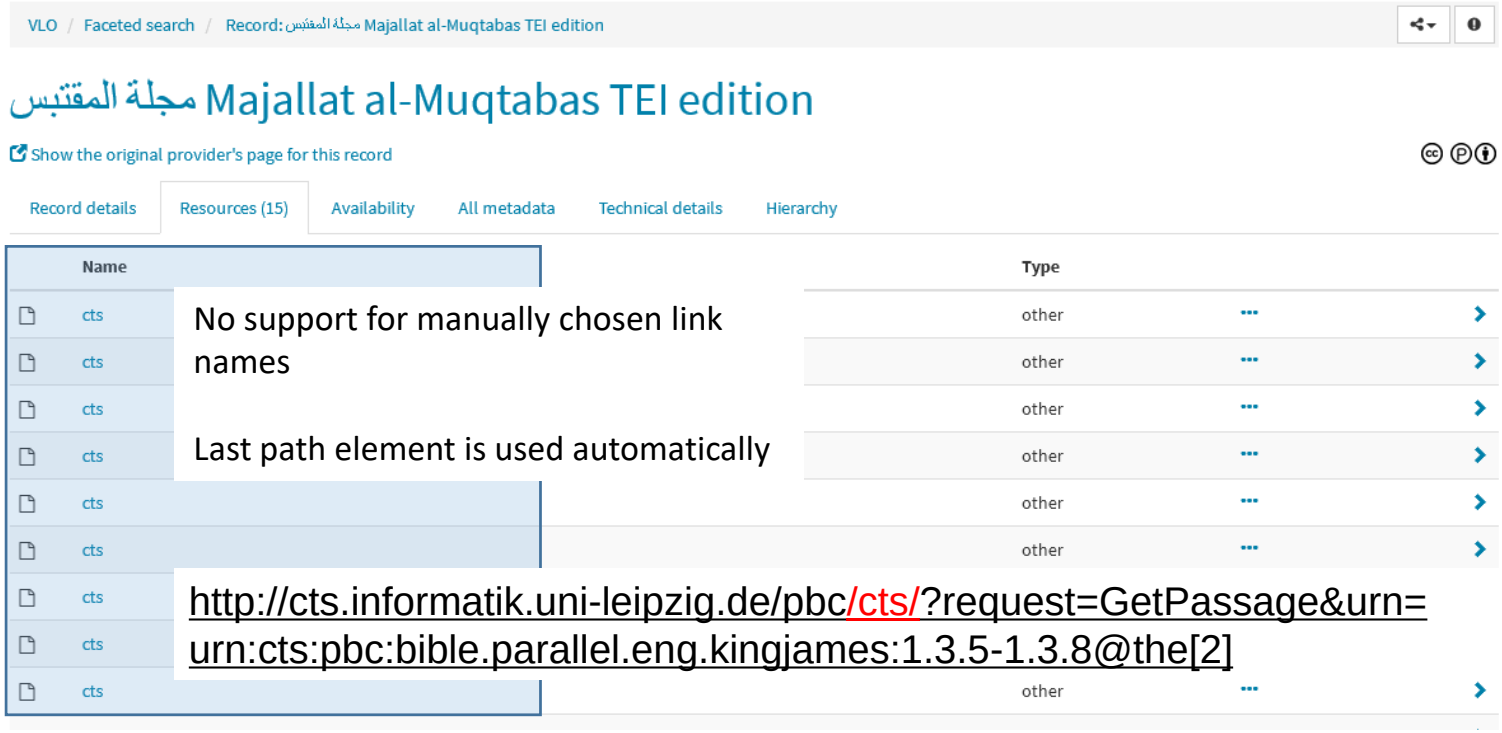
Review Feedback

the structure of the Muqtabas is not very well described in the paper

“CTS Anchor” XML tags:
front, back, div, p, head, lg, l

Review Feedback

each issue contains a list of CTS links, with no specification as to what textual unit they refer to.



No support for manually chosen link names

Last path element is used automatically

http://cts.informatik.uni-leipzig.de/pbc/cts/?request=GetPassage&urn=urn:cts:pbc:bible.parallel.eng.kingjames:1.3.5-1.3.8@the[2]

Review Feedback

the offered metadata are a bit poor (you don't encode the date, the genre, the fact that it is written text).

Yes. <genre> etc. should be added to text inventory (future work)

"written text" should be added automatically in future iterations


```
– <edition urn="urn:cts:muqtabas:oclc.4770057679_i.1.TEIP5:">
  – <title>
      مجلة المقتبس Majallat al-Muqtabas, Vol. 1, no.1 TEI edition
    </title>
    <author>محمد كرد علي Muḥammad Kurd ʿAlī</author>
  – <license>
      Distributed under a Creative Commons Attribution-ShareAlike 4.0
    </license>
  – <source>
      https://github.com/tillgrallert/digital-muqtabas/blob/master/xml/oc
      Canonical Text Service http://cts.informatik.uni-leipzig.de/muqtab
    </source>
    <publicationDate>2015</publicationDate>
    <language>arb</language>
    <contentType>xml</contentType>
  </edition>
```

Contact

OpenArabicPE

Dr. Till Grallert
E-Mail: grallert@orient-institut.org

Orient-Institut Beirut
Rue Hussein Beyhoum 44
Zokak el Blat
Beirut

Canonical Text Service

Jochen Tiepmar
E-Mail: jtiepmar@informatik.uni-leipzig.de

Scalable Data Solutions (ScaDS) Leipzig
Universität Leipzig
Ritterstraße 9-13
04109 Leipzig

CLARIN

Dr. Thomas Eckart
E-Mail: teckart@informatik.uni-leipzig.de

Dr. Dirk Goldhahn
E-Mail: dgoldhahn@informatik.uni-leipzig.de

Christoph Kuras
E-Mail: ckuras@informatik.uni-leipzig.de

NLP - Group
Universität Leipzig
Augustusplatz 10
04109 Leipzig