LD4LT@LDK-2021, Zaragoza, Spain, Sep 4, 2021

Distributed Text Services

Christian Chiarcos

Applied Computational Linguistics (ACoLi)

chiarcos@informatik.uni-frankfurt.de

:acoli dc:isPartOf <http://uni-frankfurt.de>.

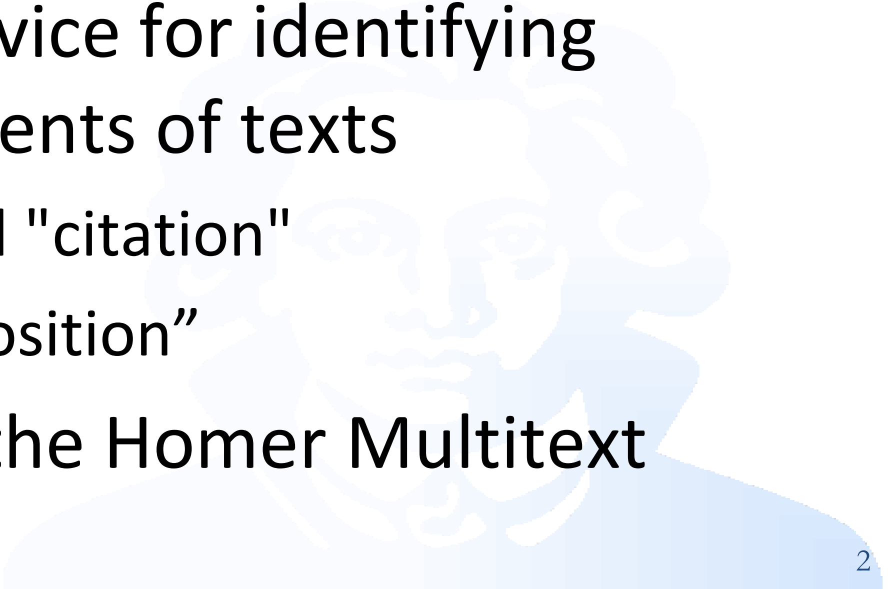
GOETHE UNIVERSITÄT FRANKFURT AM MAIN

[yʉg] LiODi Linked Open Dictionaries

Canonical Text Services (CTS)

http://cite-architecture.org/cts/

- protocol to identify and retrieve passages of text cited by canonical reference.
- specification: network service for identifying texts and retrieving fragments of texts
  - using notions of "work" and "citation"
  - rather than “string” and “position”
- Initially implemented for the Homer Multitext project



CTS URNs

http://cite-architecture.org/ctsurn/

urn:cts:[NAMESPACE]:[WORK]:[PASSAGE]

- urn:cts:pbc:bible.parallel.eng
  - WORK has a hierarchical structure

    TEXTGROUP.WORK.VERSION.EXEMPLAR
  - PASSAGE is optional

CTS URNs

http://cite-architecture.org/ctsurn/

urn:cts:[NAMESPACE]:[WORK]:[PASSAGE]

- urn:cts:pbc:bible.parallel.eng
- urn:cts:pbc:bible.parallel.eng:1.3.2
  - reference to a citable node, implicit encoding of a hierarchy
  - designed for TEI documents

CTS URNs

http://cite-architecture.org/ctsurn/

urn:cts:[NAMESPACE]:[WORK]:[PASSAGE]

- urn:cts:pbc:bible.parallel.eng
- urn:cts:pbc:bible.parallel.eng:1.3.2
- urn:cts:pbc:bible.parallel.eng:1.2-1.5.6
  - dynamic URI: a span between two nodes

CTS URNs

http://cite-architecture.org/ctsurn/

urn:cts:[NAMESPACE]:[WORK]:[PASSAGE]

- urn:cts:pbc:bible.parallel.eng
- urn:cts:pbc:bible.parallel.eng:1.3.2
- urn:cts:pbc:bible.parallel.eng:1.2-1.5.6
- urn:cts:pbc:bible.parallel.eng:1.2@the[2]-1.5.6@five
  - @ => subsections (here, tokens)

CTS URNs

http://cite-architecture.org/ctsurn/

urn:cts:[NAMESPACE]:[WORK]:[PASSAGE]

- urn:cts:pbc:bible.parallel.eng
- urn:cts:pbc:bible.parallel.eng:1.3.2
- urn:cts:pbc:bible.parallel.eng:1.2-1.5.6
- urn:cts:pbc:bible.parallel.eng:1.2@the[2]-1.5.6@five
  - [...] => index

CTS URLs

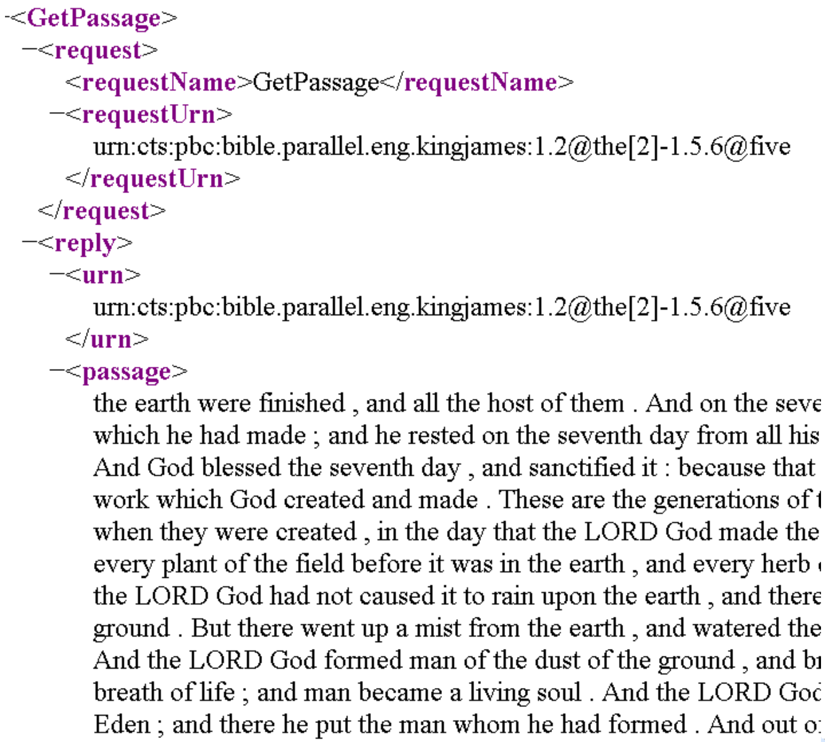
http://cite-architecture.github.io/cts_spec/

- URNs are valid URIs, but they don‘t resolve
  - ⇒ wrapping into a URI, using a URN resolver
  - ❑ urn:cts:pbc:bible.parallel.eng:1.2@the[2]-1.5.6@five
  - ❑ http://cts.informatik.uni-leipzig.de/pbc/cts/?request=GetPassage&urn=urn:cts:pbc:bible.parallel.eng.kingjames:1.2@the[2]-1.5.6@five
    - as defined in CTS protocol

CTS URLs

http://cite-architecture.github.io/cts_spec/

http://cts.informatik.uni-leipzig.de/pbc/cts/?request=GetPassage&urn=urn:cts:pbc:bible.parallel.eng.kingjames:1.2@the[2]-1.5.6@five



```
-<GetPassage>
  -<request>
      <requestName>GetPassage</requestName>
    -<requestUrn>
        urn:cts:pbc:bible.parallel.eng.kingjames:1.2@the[2]-1.5.6@five
      </requestUrn>
    </request>
  -<reply>
    -<urn>
        urn:cts:pbc:bible.parallel.eng.kingjames:1.2@the[2]-1.5.6@five
      </urn>
    -<passage>
        the earth were finished , and all the host of them . And on the seve
        which he had made ; and he rested on the seventh day from all his
        And God blessed the seventh day , and sanctified it : because that
        work which God created and made . These are the generations of t
        when they were created , in the day that the LORD God made the
        every plant of the field before it was in the earth , and every herb
        the LORD God had not caused it to rain upon the earth , and there
        ground . But there went up a mist from the earth , and watered the
        And the LORD God formed man of the dust of the ground , and br
        breath of life ; and man became a living soul . And the LORD Goc
        Eden ; and there he put the man whom he had formed . And out o:
```

examples from
http://cts.informatik.uni-leipzig.de

Return values

- CTS is designed to work with TEI documents
  - defines a special-purpose XML syntax for responses
  - ⇒ CTS URLs can resolve, but they **cannot** resolve to RDF data
- CTS Protocol requires CTS URNs
  - cannot support other (more widely used) citation systems

DTS – Distributed Text Service

https://distributed-text-services.github.io

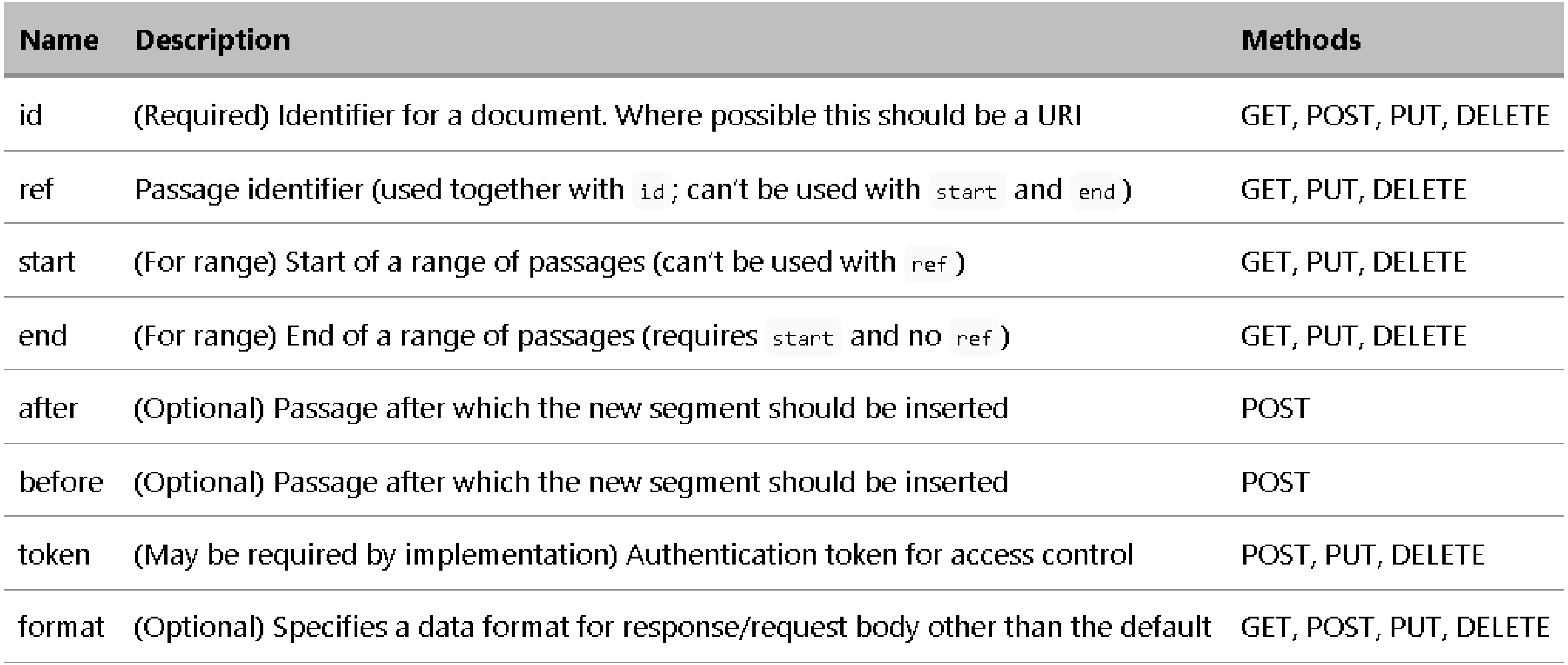
- REST API for collections of TEI documents
- inspired, informed and influenced by CTS
- but:
  - more generic, no prescriptions for identifier system, or structure of documents
- real-world systems also support CTS URNs

  https://dts.perseids.org/collections?id=urn:perseids:latinLit

DTS Operations

- Navigate across texts, navigate within texts, retrieve textual content
- e.g., retrieve
  - lists of collection members
  - metadata about individual collection items
  - lists of citeable passages within a text
  - lists of citeable passages within a text as groups of client-defined sizes
  - metadata about the citation structure of a document
  - single text passage at any level of the citation hierarchy
  - range of text passages with a clearly defined start and end passage
  - entire text

Retrieval („Document Endpoint“)



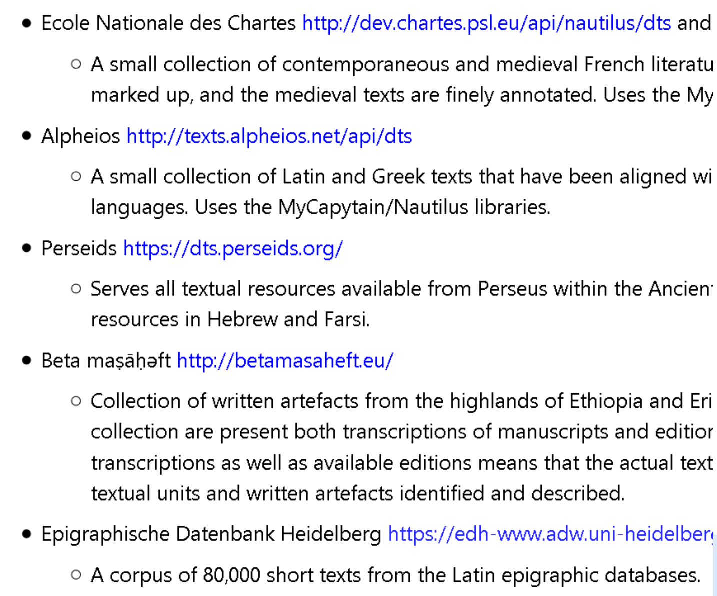
Name	Description	Methods
id	(Required) Identifier for a document. Where possible this should be a URI	GET, POST, PUT, DELETE
ref	Passage identifier (used together with `id`; can't be used with `start` and `end`)	GET, PUT, DELETE
start	(For range) Start of a range of passages (can't be used with `ref`)	GET, PUT, DELETE
end	(For range) End of a range of passages (requires `start` and no `ref`)	GET, PUT, DELETE
after	(Optional) Passage after which the new segment should be inserted	POST
before	(Optional) Passage after which the new segment should be inserted	POST
token	(May be required by implementation) Authentication token for access control	POST, PUT, DELETE
format	(Optional) Specifies a data format for response/request body other than the default	GET, POST, PUT, DELETE

DTS + Linked Data

- REST API
- JSON-LD
  - ⇒ enables full-fledged Linked Data

Sample end points:

https://distributed-text-services.github.io/specifications/

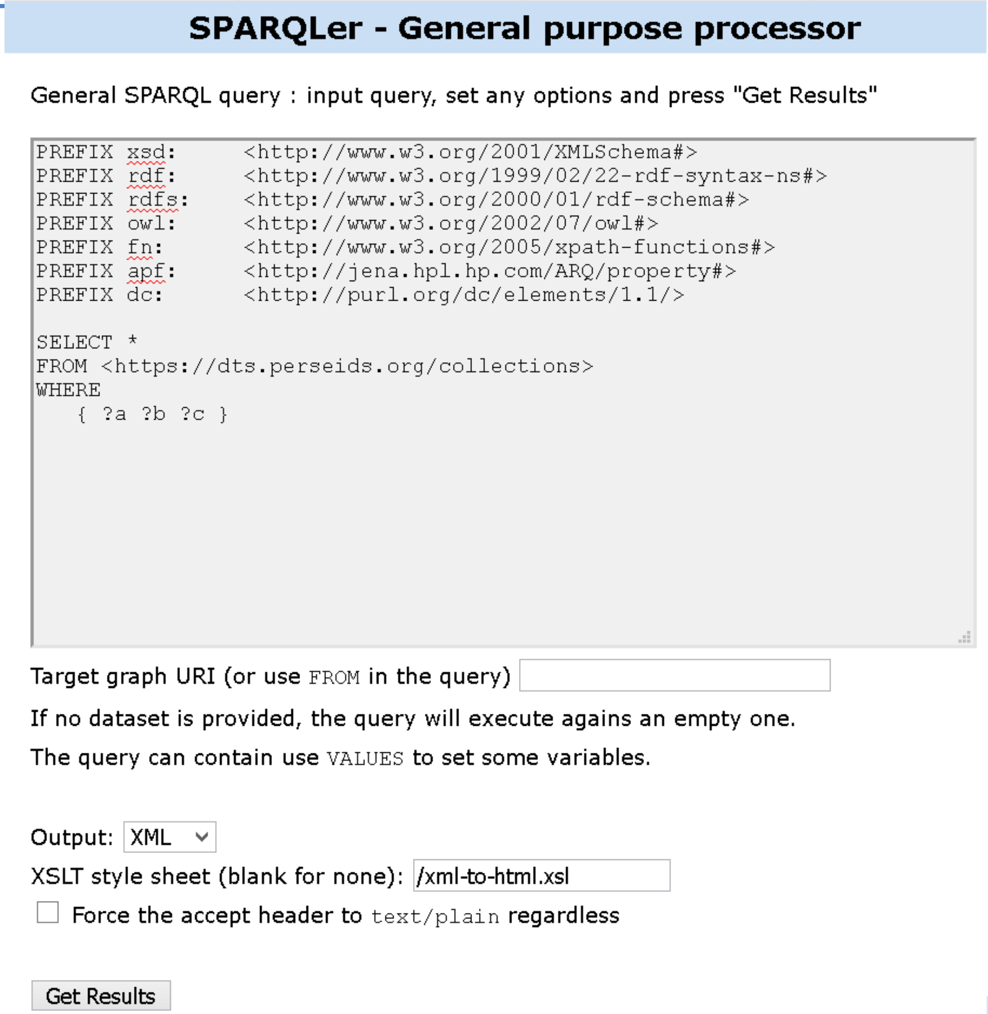


- Ecole Nationale des Chartes http://dev.chartes.psl.eu/api/nautilus/dts and
  - A small collection of contemporaneous and medieval French literatu marked up, and the medieval texts are finely annotated. Uses the My
- Alpheios http://texts.alpheios.net/api/dts
  - A small collection of Latin and Greek texts that have been aligned wi languages. Uses the MyCapytain/Nautilus libraries.
- Perseids https://dts.perseids.org/
  - Serves all textual resources available from Perseus within the Ancien resources in Hebrew and Farsi.
- Beta maṣāḥəft http://betamasaheft.eu/
  - Collection of written artefacts from the highlands of Ethiopia and Eri collection are present both transcriptions of manuscripts and editior transcriptions as well as available editions means that the actual text textual units and written artefacts identified and described.
- Epigraphische Datenbank Heidelberg https://edh-www.adw.uni-heidelber
  - A corpus of 80,000 short texts from the Latin epigraphic databases.

Accessing a DTS endpoint

https://dts.perseids.org/collections

- accessed via sparql.org
  - web service around Apache Jena



SPARQLer - General purpose processor

General SPARQL query : input query, set any options and press "Get Results"

```
PREFIX xsd:      <http://www.w3.org/2001/XMLSchema#>
PREFIX rdf:      <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs:     <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl:      <http://www.w3.org/2002/07/owl#>
PREFIX fn:       <http://www.w3.org/2005/xpath-functions#>
PREFIX apf:      <http://jena.hpl.hp.com/ARQ/property#>
PREFIX dc:       <http://purl.org/dc/elements/1.1/>

SELECT *
FROM <https://dts.perseids.org/collections>
WHERE
   { ?a ?b ?c }
```

Target graph URI (or use FROM in the query)

If no dataset is provided, the query will execute agains an empty one.

The query can contain use VALUES to set some variables.

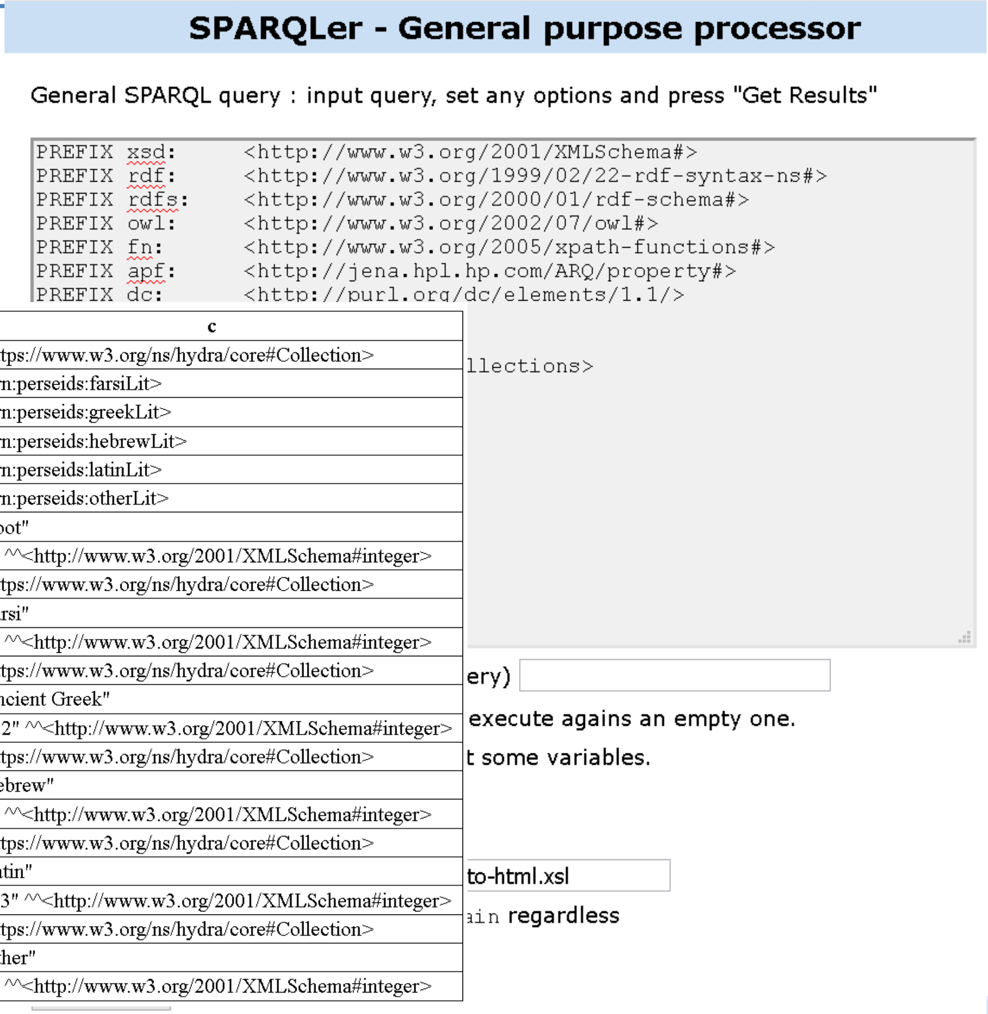
Output: XML

XSLT style sheet (blank for none): /xml-to-html.xsl

□ Force the accept header to text/plain regardless

Accessing a DTS endpoint

https://dts.perseids.org/collections



SPARQLer - General purpose processor

General SPARQL query : input query, set any options and press "Get Results"

```
PREFIX xsd:      <http://www.w3.org/2001/XMLSchema#>
PREFIX rdf:      <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs:     <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl:      <http://www.w3.org/2002/07/owl#>
PREFIX fn:       <http://www.w3.org/2005/xpath-functions#>
PREFIX apf:      <http://jena.hpl.hp.com/ARQ/property#>
PREFIX dc:       <http://purl.org/dc/elements/1.1/>
```

```
llections>
```

ery)

execute agains an empty one.

t some variables.

to-html.xsl

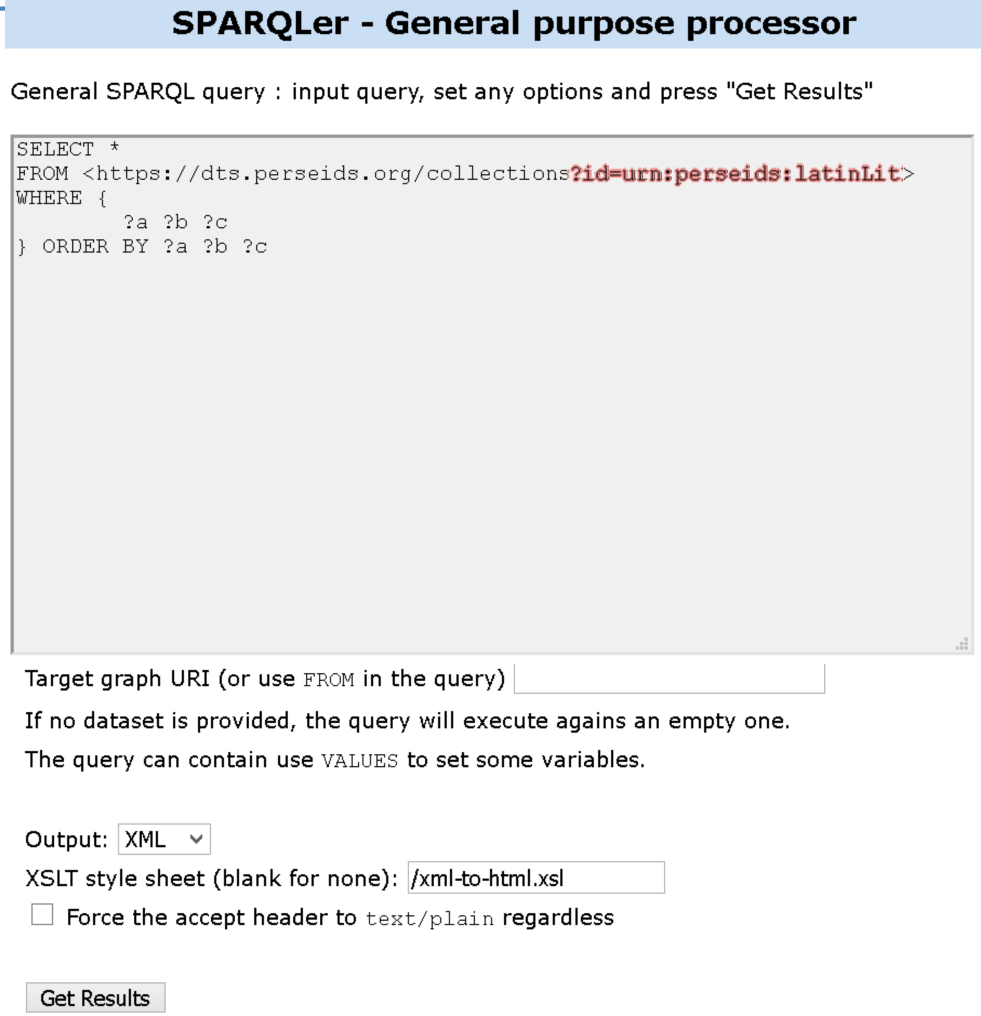
ain regardless

a	b	c
<https://dts.perseids.org/default>	<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>	<https://www.w3.org/ns/hydra/core#Collection>
<https://dts.perseids.org/default>	<https://www.w3.org/ns/hydra/core#member>	<urn:perseids:farsiLit>
<https://dts.perseids.org/default>	<https://www.w3.org/ns/hydra/core#member>	<urn:perseids:greekLit>
<https://dts.perseids.org/default>	<https://www.w3.org/ns/hydra/core#member>	<urn:perseids:hebrewLit>
<https://dts.perseids.org/default>	<https://www.w3.org/ns/hydra/core#member>	<urn:perseids:latinLit>
<https://dts.perseids.org/default>	<https://www.w3.org/ns/hydra/core#member>	<urn:perseids:otherLit>
<https://dts.perseids.org/default>	<https://www.w3.org/ns/hydra/core#title>	"Root"
<https://dts.perseids.org/default>	<https://www.w3.org/ns/hydra/core#totalItems>	"5" ^^<http://www.w3.org/2001/XMLSchema#integer>
<urn:perseids:farsiLit>	<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>	<https://www.w3.org/ns/hydra/core#Collection>
<urn:perseids:farsiLit>	<https://www.w3.org/ns/hydra/core#title>	"Farsi"
<urn:perseids:farsiLit>	<https://www.w3.org/ns/hydra/core#totalItems>	"1" ^^<http://www.w3.org/2001/XMLSchema#integer>
<urn:perseids:greekLit>	<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>	<https://www.w3.org/ns/hydra/core#Collection>
<urn:perseids:greekLit>	<https://www.w3.org/ns/hydra/core#title>	"Ancient Greek"
<urn:perseids:greekLit>	<https://www.w3.org/ns/hydra/core#totalItems>	"212" ^^<http://www.w3.org/2001/XMLSchema#integer>
<urn:perseids:hebrewLit>	<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>	<https://www.w3.org/ns/hydra/core#Collection>
<urn:perseids:hebrewLit>	<https://www.w3.org/ns/hydra/core#title>	"Hebrew"
<urn:perseids:hebrewLit>	<https://www.w3.org/ns/hydra/core#totalItems>	"4" ^^<http://www.w3.org/2001/XMLSchema#integer>
<urn:perseids:latinLit>	<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>	<https://www.w3.org/ns/hydra/core#Collection>
<urn:perseids:latinLit>	<https://www.w3.org/ns/hydra/core#title>	"Latin"
<urn:perseids:latinLit>	<https://www.w3.org/ns/hydra/core#totalItems>	"113" ^^<http://www.w3.org/2001/XMLSchema#integer>
<urn:perseids:otherLit>	<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>	<https://www.w3.org/ns/hydra/core#Collection>
<urn:perseids:otherLit>	<https://www.w3.org/ns/hydra/core#title>	"Other"
<urn:perseids:otherLit>	<https://www.w3.org/ns/hydra/core#totalItems>	"2" ^^<http://www.w3.org/2001/XMLSchema#integer>

Accessing a DTS endpoint

https://dts.perseids.org/collections

- accessed via sparql.org
- FROM
  - ⇒ GET request at DTS end point
  - ⇒ results can be loaded into named graph



SPARQLer - General purpose processor

General SPARQL query : input query, set any options and press "Get Results"

```
SELECT *
FROM <https://dts.perseids.org/collections?id=urn:perseids:latinLit>
WHERE {
        ?a ?b ?c
} ORDER BY ?a ?b ?c
```

Target graph URI (or use `FROM` in the query)

If no dataset is provided, the query will execute agains an empty one.

The query can contain use `VALUES` to set some variables.

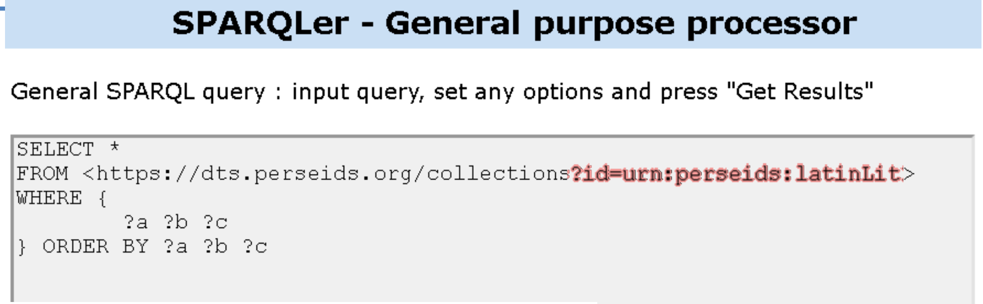
Output: XML

XSLT style sheet (blank for none): /xml-to-html.xsl

□ Force the accept header to `text/plain` regardless

Accessing a DTS endpoint

https://dts.perseids.org/collections



SPARQLer - General purpose processor

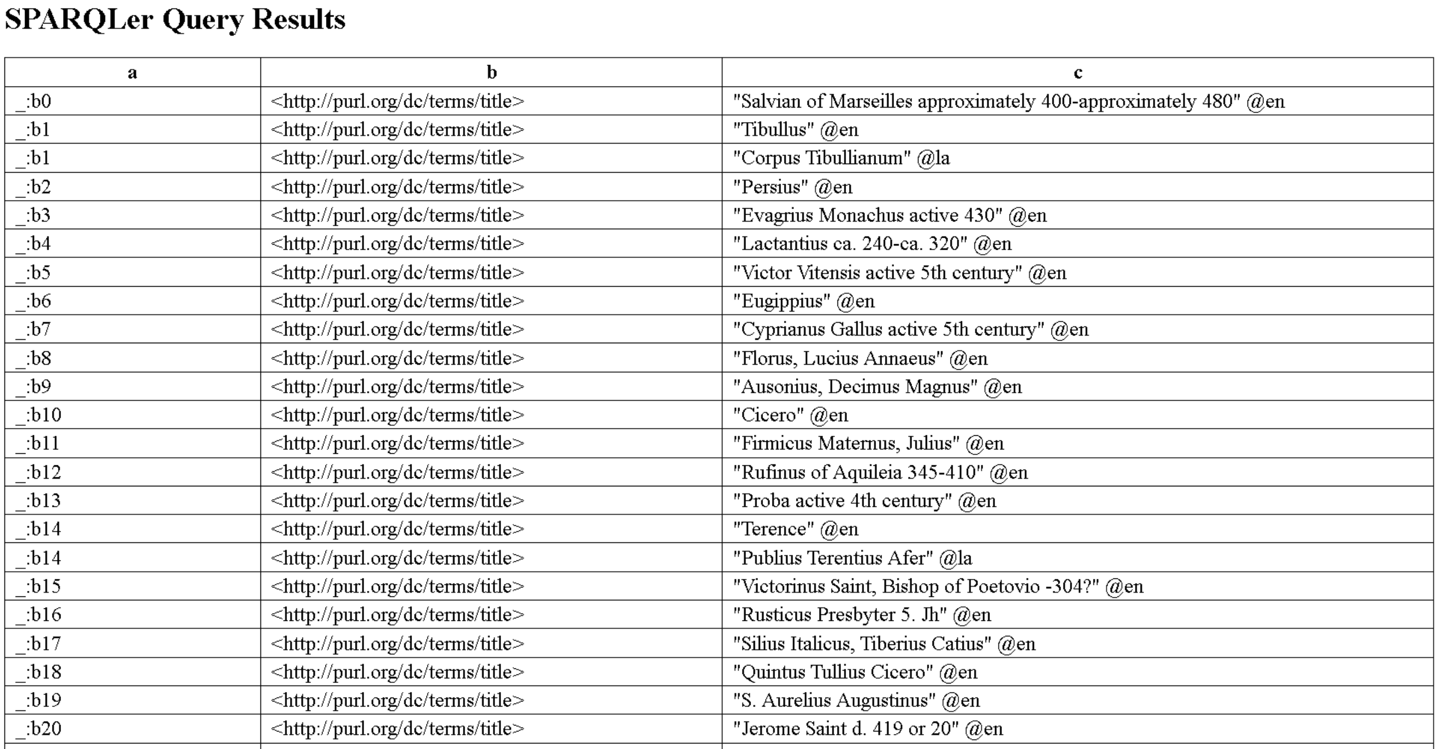
General SPARQL query : input query, set any options and press "Get Results"

```
SELECT *
FROM <https://dts.perseids.org/collections?id=urn:perseids:latinLit>
WHERE {
        ?a ?b ?c
} ORDER BY ?a ?b ?c
```

ins an empty one.

bles.

ss



SPARQLer Query Results

a	b	c
_:b0	<http://purl.org/dc/terms/title>	"Salvian of Marseilles approximately 400-approximately 480" @en
_:b1	<http://purl.org/dc/terms/title>	"Tibullus" @en
_:b1	<http://purl.org/dc/terms/title>	"Corpus Tibullianum" @la
_:b2	<http://purl.org/dc/terms/title>	"Persius" @en
_:b3	<http://purl.org/dc/terms/title>	"Evagrius Monachus active 430" @en
_:b4	<http://purl.org/dc/terms/title>	"Lactantius ca. 240-ca. 320" @en
_:b5	<http://purl.org/dc/terms/title>	"Victor Vitensis active 5th century" @en
_:b6	<http://purl.org/dc/terms/title>	"Eugippius" @en
_:b7	<http://purl.org/dc/terms/title>	"Cyprianus Gallus active 5th century" @en
_:b8	<http://purl.org/dc/terms/title>	"Florus, Lucius Annaeus" @en
_:b9	<http://purl.org/dc/terms/title>	"Ausonius, Decimus Magnus" @en
_:b10	<http://purl.org/dc/terms/title>	"Cicero" @en
_:b11	<http://purl.org/dc/terms/title>	"Firmicus Maternus, Julius" @en
_:b12	<http://purl.org/dc/terms/title>	"Rufinus of Aquileia 345-410" @en
_:b13	<http://purl.org/dc/terms/title>	"Proba active 4th century" @en
_:b14	<http://purl.org/dc/terms/title>	"Terence" @en
_:b14	<http://purl.org/dc/terms/title>	"Publius Terentius Afer" @la
_:b15	<http://purl.org/dc/terms/title>	"Victorinus Saint, Bishop of Poetovio -304?" @en
_:b16	<http://purl.org/dc/terms/title>	"Rusticus Presbyter 5. Jh" @en
_:b17	<http://purl.org/dc/terms/title>	"Silius Italicus, Tiberius Catius" @en
_:b18	<http://purl.org/dc/terms/title>	"Quintus Tullius Cicero" @en
_:b19	<http://purl.org/dc/terms/title>	"S. Aurelius Augustinus" @en
_:b20	<http://purl.org/dc/terms/title>	"Jerome Saint d. 419 or 20" @en

Summary

- CTS URIs
  - yet another text addressing system
  - *canonical*, can abstract from/generalize over multiple string representations
- DTS protocol
  - resolvable (CTS and other) URIs
  - JSON-LD responses, full RDF integration