

(nederlab) Reflections on a platform for library textual data

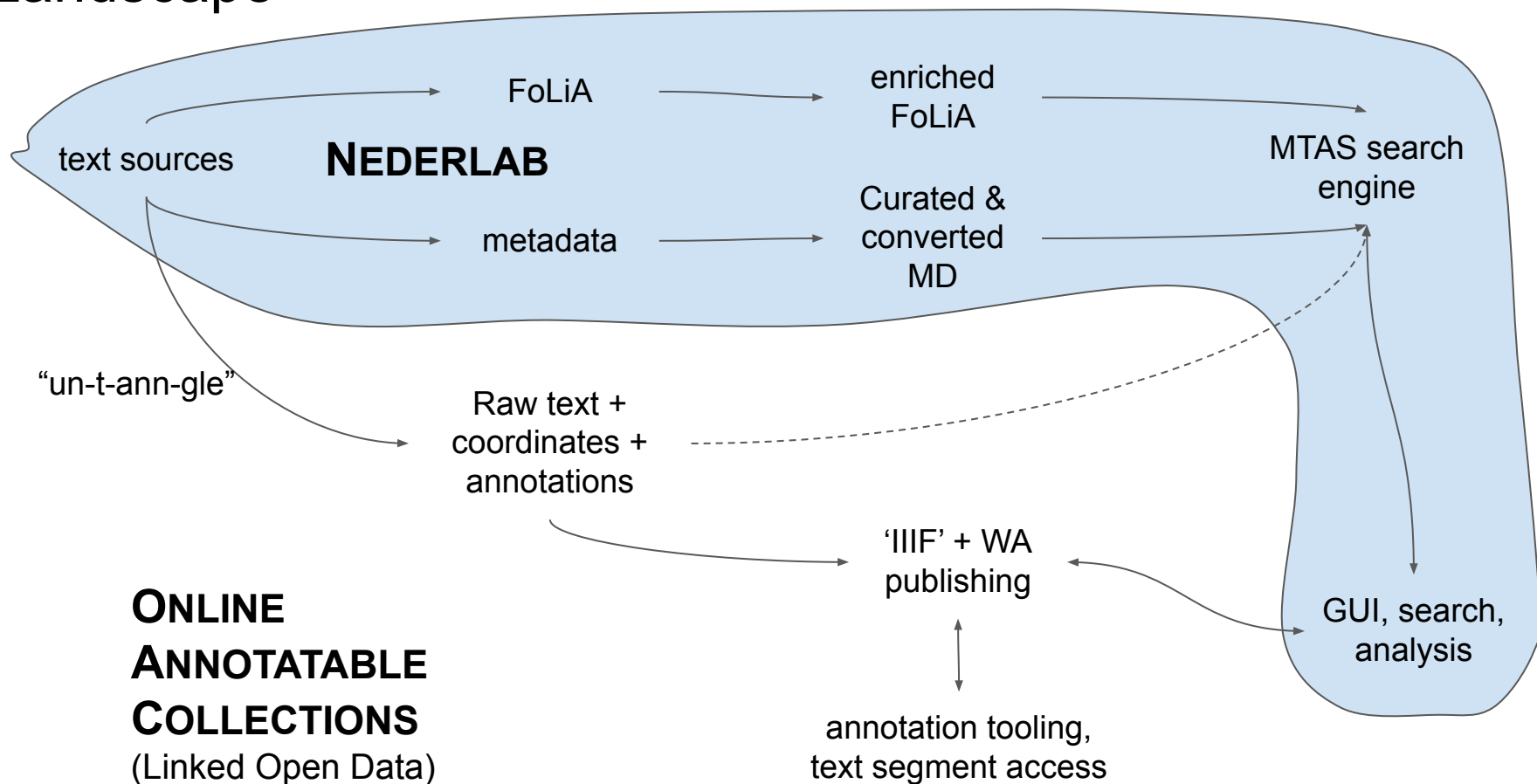
Hennie Brugman, KNAW Humanities Cluster, Digital Infrastructure

CLARIN & Libraries workshop, May 9-10, 2022

Overview

- The landscape
- Nederlab and lessons learned
- Online annotatable collections (and examples)
- Conclusions
- Issues and questions to discuss

Landscape



Nederlab

- Nederlab
 - Scholarly exploitation of historic dutch text collections
 - diachronic
 - project ended: 2018
 - Status: no continued software development, some collection updates, hosting upgrades
 - Regularly used by scholars
- Main features and statistics
 - 24 collections, 19 billion tokens, almost 100 different annotation layer (sub)types, 100 billion annotations
 - Query based: SOLR + MTAS plugin
 - Back end with built-in enriched text analysis
 - statistics, distributions, frequency lists, grouping, lexical query expansion

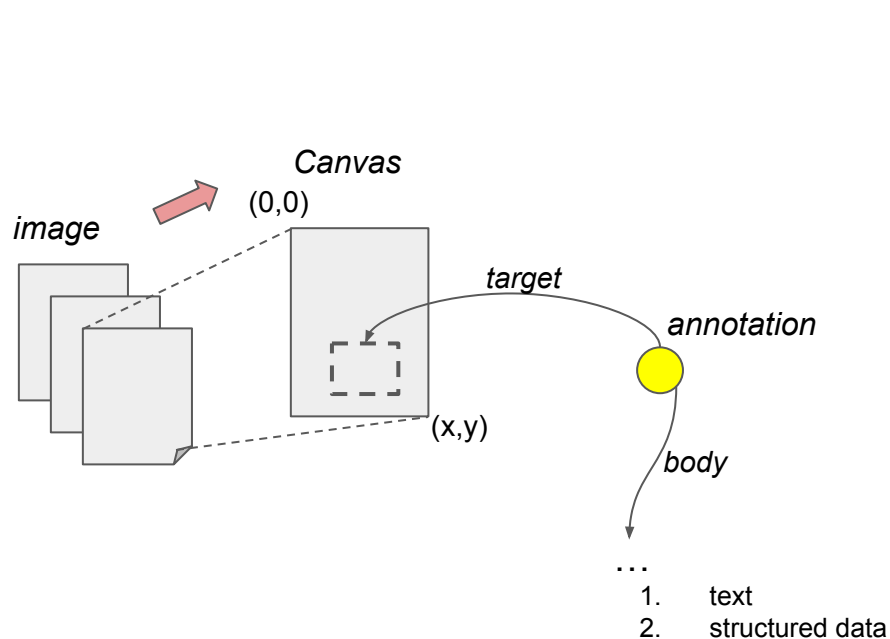
Nederlab: lessons learned

- Issues and lessons learned
 - **Project vs service**
 - Keep know how on board
 - **Cost of updates and upgrades**
 - Data quality
 - **Deliver enrichments back to providers**
 - **Rights issues**
 - **Increasing demand for access via API**
 - **Direct access to text files**
 - Usability, GUI
 - **Document segmentation, document parts**

'Online annotatable text collections'

- Shift focus from interactive research environments to online accessible data
 - Linked Open Data based (vs query based, as in Nederlab)
 - Smart back ends (APIs), small and simple ('micro'-) clients
- Annotation centered
 - Enrichments, entities, document structure as annotations
 - Annotation sets: first class data, scholarly autonomy
 - Scholarly discourse: 'annotation cloud' around online collections
 - Combine collection providers' annotations with user annotations
 - Standards based: Web Annotations + 'IIIF'
 - Combine text annotation with other media types
- Annotatable collections
 - Canvas-like coordinate systems, also for text
 - Substantial preprocessing required
- Experiments needed
 - scalability, costs (of processing, storage), maintainability, persistence

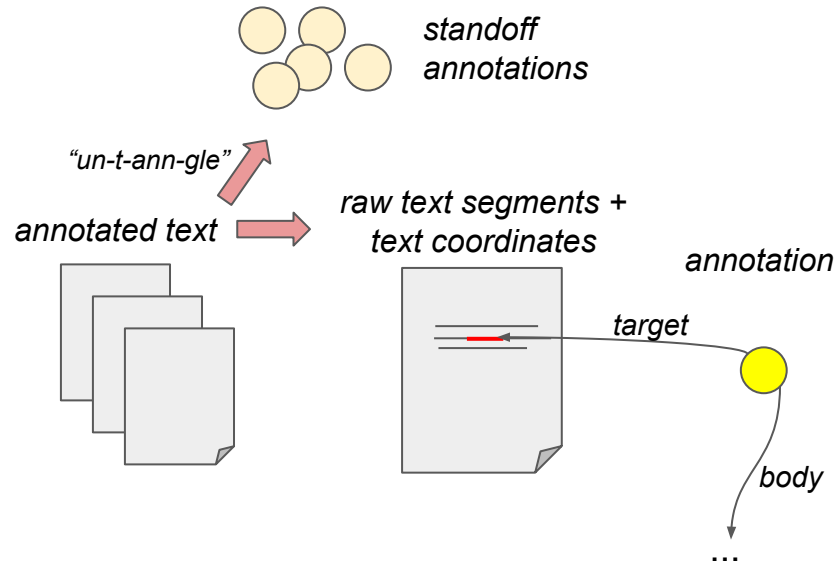
“IIIF”: images vs text



coordinates: rectangle with 2 ints (width, height)

url:

<https://images.diginfra.net/iiif/image.jpg/1451,2124,511,60/full/0/default.jpg>



coordinates: sequence of text segments with (persistent) anchors

url:

<http://host.nl/textrepo/.../ann-id/segments/index/56379,38,56380,3>

Example annotation

- From Republic project
- micro-client
- Logical object: a 'resolution' of Dutch parliament
- Multi target web annotation
- Image and text segment retrieved using 'IIF'
- Retrieve overlapping annotations

Annotation ID

urn:example:republic:meeting-1728-06-19-session-1-resolution-17



Search

O Ntangen een Miffive van den Raad van Staate, gefchreven alhier in den Hage den feventienden defer loopende maandt; houdende, dat op den een en twintighften Mey laaftlieden ter Vergaderinge van haar Hoogh Mogende was uytgebragt een rapport van de Heeren haar Hoogh Mogende Geleputeerden tot de taaken van de Finantie, met eenige Heeren Gecommitteerden uyt den Raad van Staate, geëxamineert hadden haare Miffive van den fevenden daer te vooren; welck rapport daer toe tendteert, dat het profyt van de jegen-

woordigh onder handen zijnde Generaliteyts Loterye, en het geene van de voorige was geproffiteert, te famen ter fomme van hondert feventigh duyfent guldens, geëmployeert fouden werden tot de Fortificaten, om redenen by het rapport geallegeert; dogh daer daer op is aldoen, en foo veel fy wetten, tot noch toe geen conclufie is gevallen; de Heeren Gedeputeerden van de twee Provincien aangenomen hebbende haar in weynigh dagen te verklaren, en van een andere verklaart hebbende ongeloft te zijn. Dat fy hadden gemeent niet te kunnen afzyn haar Hoogh Mogende defe gefelthet van faaken voor te dragen waar uyt haar Hoogh Mogende fullen sien dat de conclufie op het voorfchreve rapport ten uytterlicte preffteert, en by langer uytffel fy geenoedtraecht zijn, tot groot discredit van het Landt, en groot nadeel van de begonnen Wercken, ende van de fecuriteyt der Frontieren, in dit faayfen, dat het bequaamfte is het werck te ftaaken, of, wyl niet verantwoordelijck fchynt, het geldt, dat daar toe geëmployeert foude kunnen werden, gereedt zijnde, onderuffchen dat daar over gedelibereert werdt, het werck buyten fijn geheel te brengen, by provifie uyt die penningen foo veel te ontleen, als tot het aan de gangh houden van de Wercken abfolut gerequireert wordt, het geen fy vertrouwen dat haar in een foo dringende noode niet quajijck fal werden afgenomen; dogh het welcke maar voor een korten tydt foude kunnen ftrecken, en haar, na het verloop van dien, weder in de felve verlegentheyte te brengen, en het werck wel aan de gangh houden, maar eghter niet doen voortgaan, met die kragt en fpoedt, als in dit laaijen behoorde.

Verfoeckende derhalven, dat het meer-

Ntangen een Miffive van den Raad: & NP van Staate, gefchreven alhier in den Hage den feventienden defer loopende maandt; houdende, dat op den een en twintighften Mey laaftlieden ter Vergaderinge van haar Hoogh Mogende was uytgebragt een rapport van de Heeren haar Hoogh Mogende Gedeputeerden tot de faaken van de Finantie, met eenige Heeren Gecommitteerden uyt den Raads van Staate, geëxamineert hadden haare Miffive van den fevenden daer te vooren; welck rapport daer toe tendteert, dat het profyt van de jegen: wordigh onder handen zijnde Generaliteyts Loterye, en het geene van de voorige was geproffiteert, te famen ter fomme van hondert feventigh duyfent guldens, geëmployeert fouden werden tot de Fortificaten, om redenen by het rapport geallegeert; dogh daer daer op is aldoen, en foo veel fy wetten, tot noch toe geen conclufie is gevallen; de Heeren Gedeputeerden van de twee Provincien aangenomen hebbende haar in weynigh dagen te verklaren, en van een andere verklaart hebbende ongeloft te zijn. Dat fy hadden gemeent niet te kunnen afzyn haar Hoogh Mogende defe gefelthet van faaken voor te dragen waar uyt haar Hoogh Mogende fullen sien dat de conclufie op het voorfchreve rapport ten uytterlicte preffteert, en by langer uytffel fy geenoedtraecht zijn, tot groot discredit van het Landt, en groot nadeel van de begonnen Wercken, ende van de fecuriteyt der Frontieren, in dit faayfen, dat het bequaamfte is het werck te ftaaken, of, wyl niet verantwoordelijck fchynt, het geldt, dat daar toe geëmployeert foude kunnen werden, gereedt zijnde, onderuffchen dat daar over oedelibereert werdt, het werck buyten fijn geheel te brengen, by provifie uyt die penningen [oo veel te ontleen, als tot het aan de gangh houden van de Wercken abfolut gerequireert wordt, het geen fy vertrouwen dat haar in een foo dringende noode niet quajijck fal werden afgenomen; dogh het welcke maar voor een korten tydt [oude kunnen ftrecken, en haar, na het verloop van dien, weder in de [elve verlegentheyte te brengen, en bet werck wel aan de ganegh houden, maar eghter niet doen voortgaan, met die kragt en fpoedt, als in dit [alfoen behoorde.

Annotator

Select text to create an annotation

Existing user annotations

location (1,38)(2,3)

- id: 588268d2-c9d2-4ece-bbae-b4e8a2d7alda
- type: Entity (location)
- comment: revise
- coordinates: (1,38)(2,3)
- creator: http://example.org/hennie
- ```
{
 "@context": "http://www.w3.org",
 "id": "http://elucidate:8080/a",
 "type": [
 "Annotation",
 "http://example.org/customwe"
],
 "creator": "http://example.org",
 "body": [
 {
 "type": "TextualBody",
 "value": "location",
 "purpose": "classifying"
 },
 {
 "type": "TextualBody",
 "value": "revise",
 "purpose": "commenting"
 }
],
 "target": "http://localhost:80"
}
```



# Current status

- TextRepo: our first 'IIF for text' implementation
- eLucidate/AnnoRepo
- un-t-ann-gle
- Proof of concept (micro-)clients
- Projects & collections
  - WF Hermans (book versions, TEI)
  - Mondrian letters (letters, TEI)
  - Globalise (VOC, General Missives, PageXML)
  - Republic (Resolutions of States General, custom json)
  - CLARIAH Plus core shared service: FAIR Annotations

# Conclusions

- Both presented approaches have benefits and are complementary
- Invest more in accessible data, less in front-ends and search engines
  - More autonomy for scholars
  - Visualisation, analysis and search become more scholars' own responsibility
  - More sustainable
- Annotations
  - Are the perfect glue between/within collections
  - Can be valuable user contributed extensions of library collections
- Nederlab issues and lessons learned, revisited:
  - **Project vs service**
  - **Cost of updates and upgrades**
  - **Deliver enrichments back to providers**
  - **Rights issues**
  - **Increasing demand for access via API**
  - **Direct access to text files**
  - **Document segmentation, document parts**

# Questions and issues

- How well does this all scale?
- Text + text coordinates: where does this model break?
- Text versions
- Round trips to representations used by scholarly communities (TEI)
- Persistence
- Participation of collection providers