# (nederlab) Reflections on a platform for library textual data

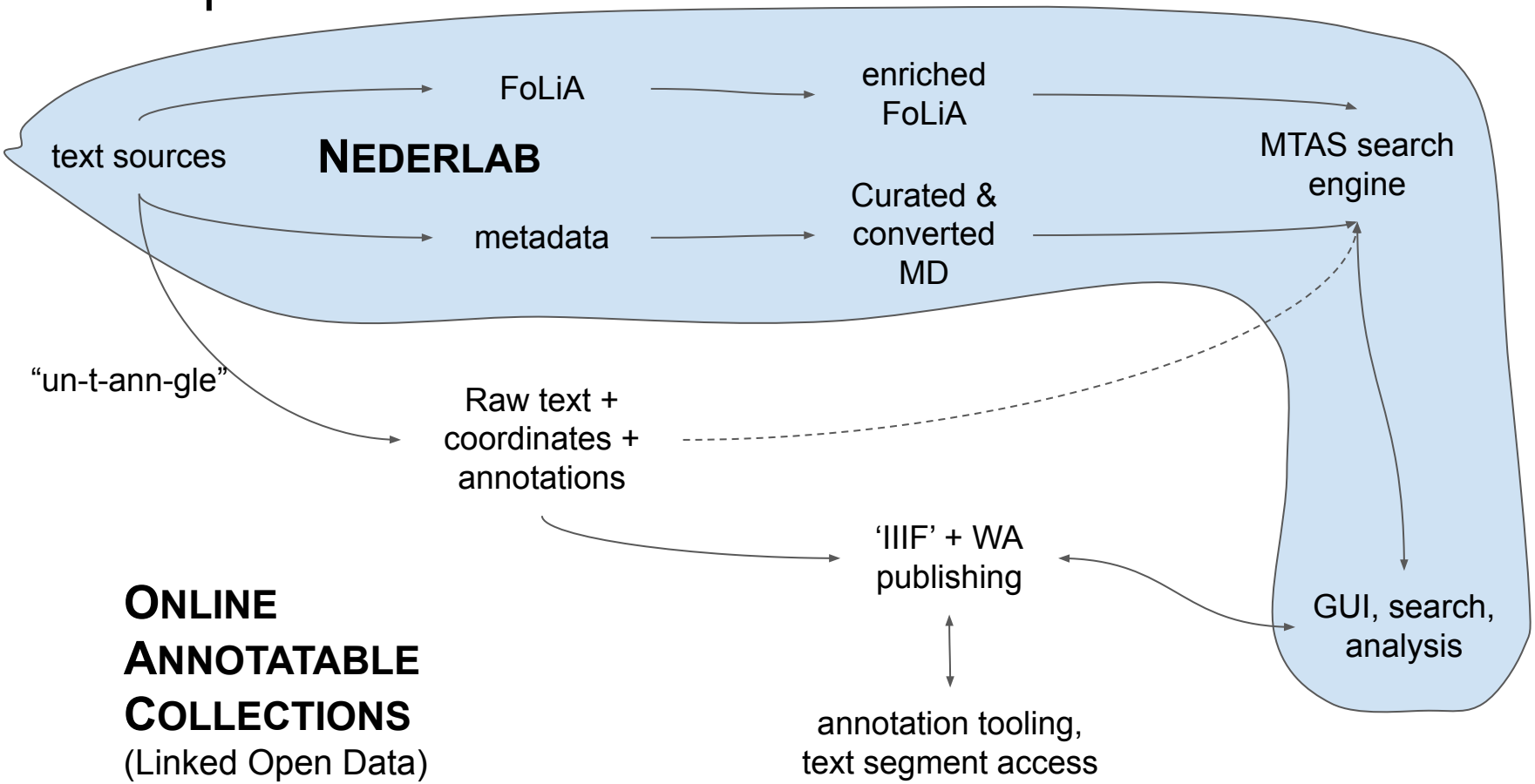
Hennie Brugman, KNAW Humanities Cluster, Digital Infrastructure

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




**ONLINE ANNOTATABLE COLLECTIONS**
(Linked Open Data)

# 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

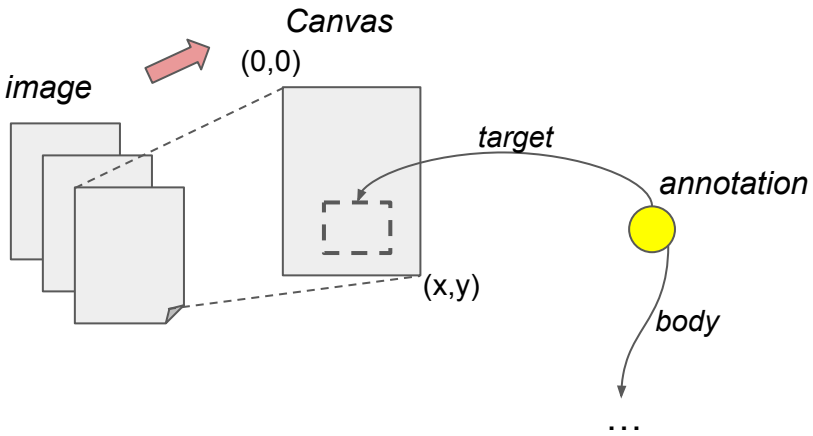
# Nederlab: lessons learned

- 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**

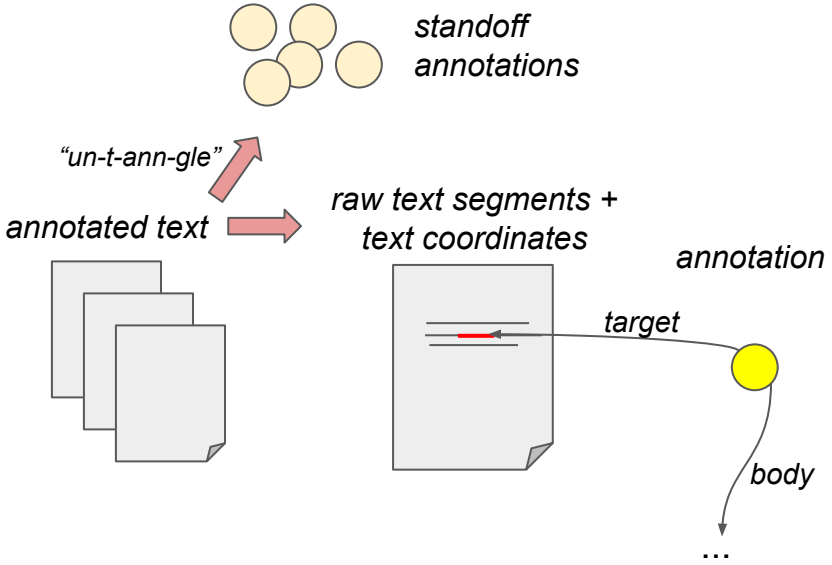
# ‘Online annotatable text collections’

- 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




1. text
2. structured data




coordinates: rectangle with 2 ints (width, height)

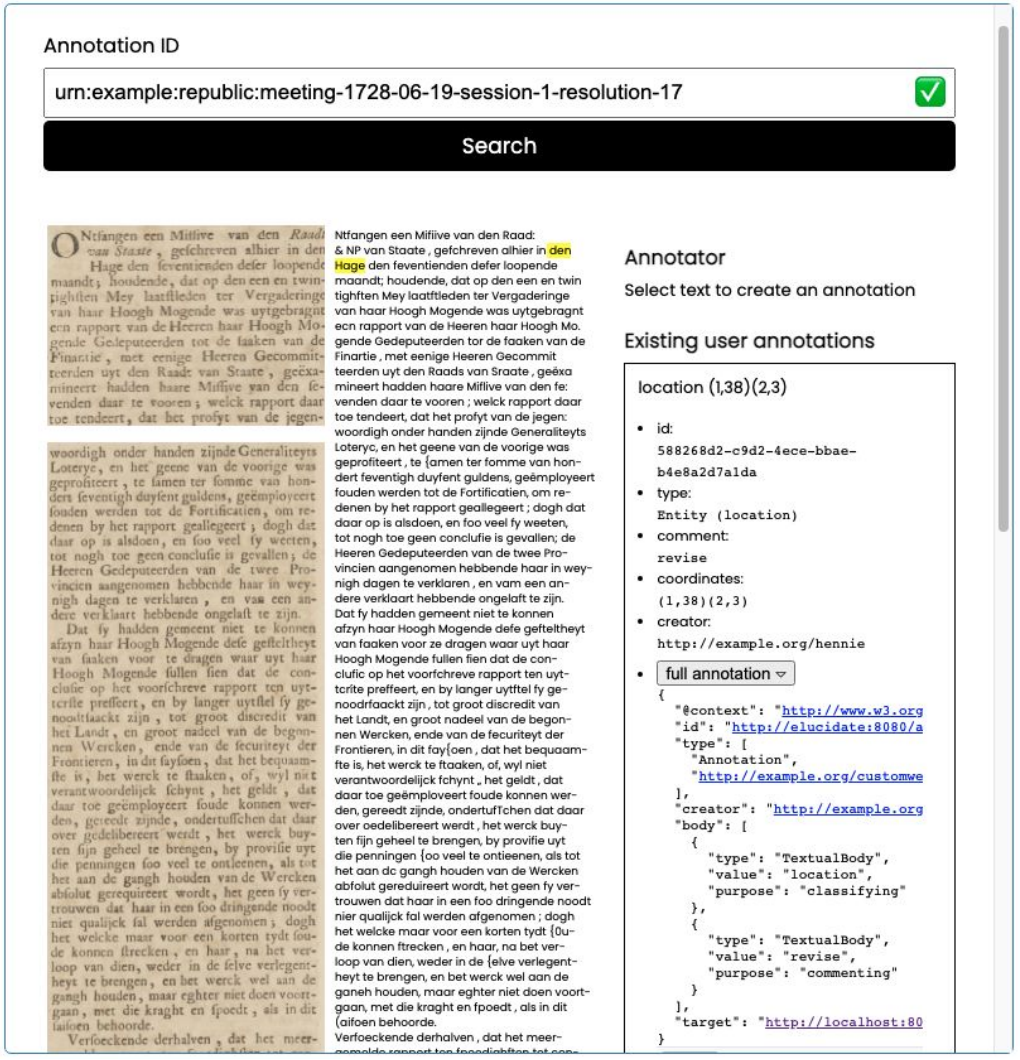
url:
https://images.diginfra.net/iiif/image.jpg/1451,2124,511,60/full/0/default.jpg

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 ‘IIIF’
- Retrieve overlapping annotations



# Current status

- TextRepo: our first ‘IIIF 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