Unlocking the archives

A pipeline for scanning, indexing, transcribing, and modelling entities of archival documents into Linked Open Data

Leon van Wissen
University of Amsterdam
l.vanwissen@uva.nl

Chiara Latronico
University of Amsterdam
c.latronico@uva.nl

Veruska Zamborlini
University of Amsterdam
v.zamborlini@uva.nl

Jirsi Reinders
Huygens Institute / Amsterdam City Archives
jirsi.reinders@huygens.knaw.nl

Charles van den Heuvel
Huygens Institute
charles.van.den.heuvel@huygens.knaw.nl

Keywords

Entity Extraction from Archival Documents, GLAM, Linked Open Data, Dutch Golden Age

Abstract

In the project *Golden Agents: Creative Industries and the Making of the Dutch Golden Age*, heterogeneous resources on the production of the creative industries in the Dutch Golden Age from heritage institutions (e.g. Rijksmuseum, KB, RKD) are brought together as linked data. Added to this, the digitization of the enormously rich collection of the notarial deeds in the Amsterdam City Archives will provide data on the consumption of cultural goods by the inhabitants of all layers of society in Amsterdam during the Dutch Golden Age. This archive currently plays a pioneering role in the massive digitization process of archival inventories. In the project *Alle*

Amsterdams Akten [All Amsterdam Deeds] handwritten notarial deeds are indexed on the level of inventories, documents, person names, and geolocations outside Amsterdam. At the same time, the full text of these documents is being made searchable by using the advanced Handwritten Text Recognition (HTR) tool Transkribus in the project *Crowd Leert Computer Lezen* [Crowd Teaches the Computer how to Read] in combination with corrections of the transcriptions by volunteers.

In the Golden Agents project, novel ways are explored to extract all entities of objects that are mentioned in such notary deeds between 1578 and 1750 that are relevant to get insight into the cultural goods of Amsterdamers in the Dutch Golden Age. TICCLAT (Reynaert et al. 2019) is used to find and extract these object entities, and once extracted and identified, almost all types of these objects can be linked to thesauri such as the Getty's Art & Architecture Thesaurus [AAT] and reconciled with textual/linguistic references to an item in an external (authored) dataset, such as the STCN, ICONCLASS, and those of the RKD. The development of the model that is used to express the combined, enriched, and created data is still work in progress. It will be compliant with major and widely used data models in the Galleries, Libraries, Archives, and Museums (GLAM) world, such as the CIDOC-CRM.

Here the full pipeline from archives to annotations is represented (Figure 1) that comprehends the successive stages of scanning, indexing, transcribing, correcting, aggregating, and modelling the entities of archival documents into RDF as Linked Open Data. It provides the creation of transparent datasets that can be replicated, evaluated and used for quantitative analyses in digital humanities research. Subsequently, Figure 2 and Figure 3 show a simplified RDF representation of a textual reference to a painting and a book as can be found in probate inventories.

These two examples show how references to objects in archival documents can be connected to external thesauri and datasets while keeping the provenance chain by pointing back to their location in the archival document. The same principle applies to other entity types, such as persons and locations. Besides connecting the entity to an external (authored) dataset (e.g. ECARTICO or ULAN), disambiguating the entity within the Amsterdam City Archives dataset can also be a next step. Within the Golden Agents projects, the Lenticular Lenses (Idrissou et al. 2018) tool is used for this.

References

Art & Architecture Thesaurus (AAT), https://www.getty.edu/research/tools/vocabularies/ .

Amsterdam City Archives, https://archief.amsterdam/.

All Amsterdam Deeds & Crowd Teaches the Computer how to Read, https://alleamsterdamseakten.nl/doemee/ .

CIDOC-CRM, http://www.cidoc-crm.org/ .

ECARTICO: linking cultural industries in the early modern Low Countries, http://www.vondel.humanities.uva.nl/ecartico/ .

Golden Agents: Creative industries and the making of the Dutch Golden Age, https://www.goldenagents.org/ .

ICONCLASS, a multilingual classification system for cultural content, http://www.iconclass.org/ .

Al Idrissou, Veruska Zamborlini, Chiara Latronico, Frank van Harmelen and Charles van den Heuvel, (2018) "Amsterdamers from the Golden Age to the Information Age via Lenticular Lenses," Short Paper, DHBenelux 2018, Amsterdam, http://2018.dhbenelux.org/wp-content/uploads/sites/8/2018/05/Al-Idrissou-Chiara-Latronico_GoldenAgentsLenticularLenses_DHBenelux2018.pdf .

Short Title Cataglogue of the Netherlands (STCN), https://www.kb.nl/organisatie/onderzoek-expertise/informatie-infrastructuur-diensten-voor-bibliotheken/short-title-catalogue-netherlands-stcn .

RKD-Netherlands Institute for Art History, https://rkd.nl/ .

Martin Reynaert, Janneke van der Zwaan, and Patrick Bos, (2019) "TICCLAT: a Dutch diachronic database of linked word variants." Short Paper, DHBenelux 2019, Liège, http://2019.dhbenelux.org/wp-content/uploads/sites/13/2019/08/DH_Benelux_2019_paper_46.pdf .

Transkribus, https://transkribus.eu/Transkribus/ .

Union List of Artist Names (ULAN), https://www.getty.edu/research/tools/vocabularies/ .

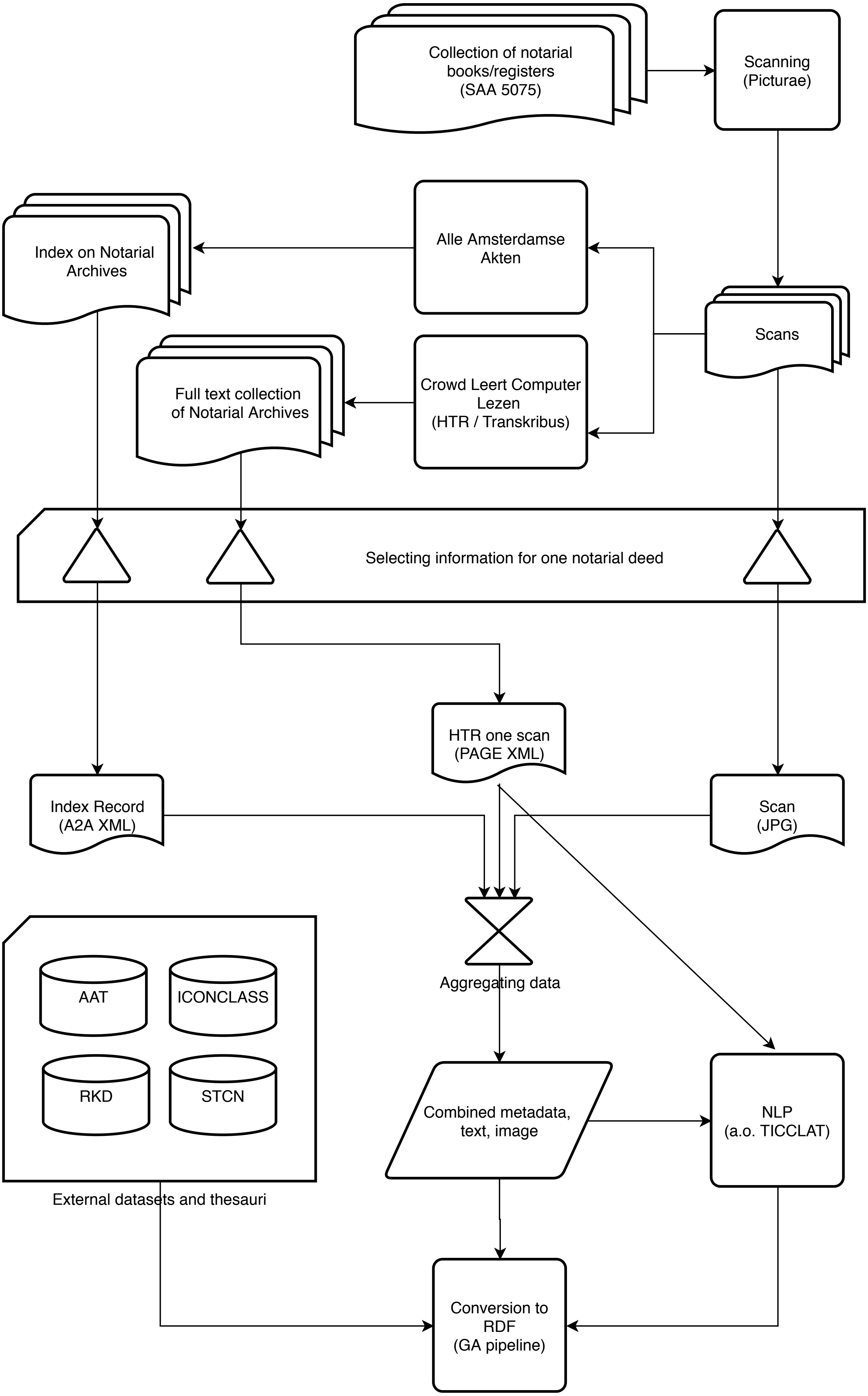
Collection of notarial books/registers (SAA 5075)
Scanning (Picturae)
Index on Notarial Archives
Alle Amsterdamse Akten
Scans
Full text collection of Notarial Archives
Crowd Leert Computer Lezen (HTR / Transkribus)
Selecting information for one notarial deed
HTR one scan (PAGE XML)
Index Record (A2A XML)
Scan (JPG)
Aggregating data
AAT
ICONCLASS
RKD
STCN
External datasets and thesauri
Combined metadata, text, image
NLP (a.o. TICCLAT)
Conversion to RDF (GA pipeline)

Figure 1 - Flowchart of the pipeline

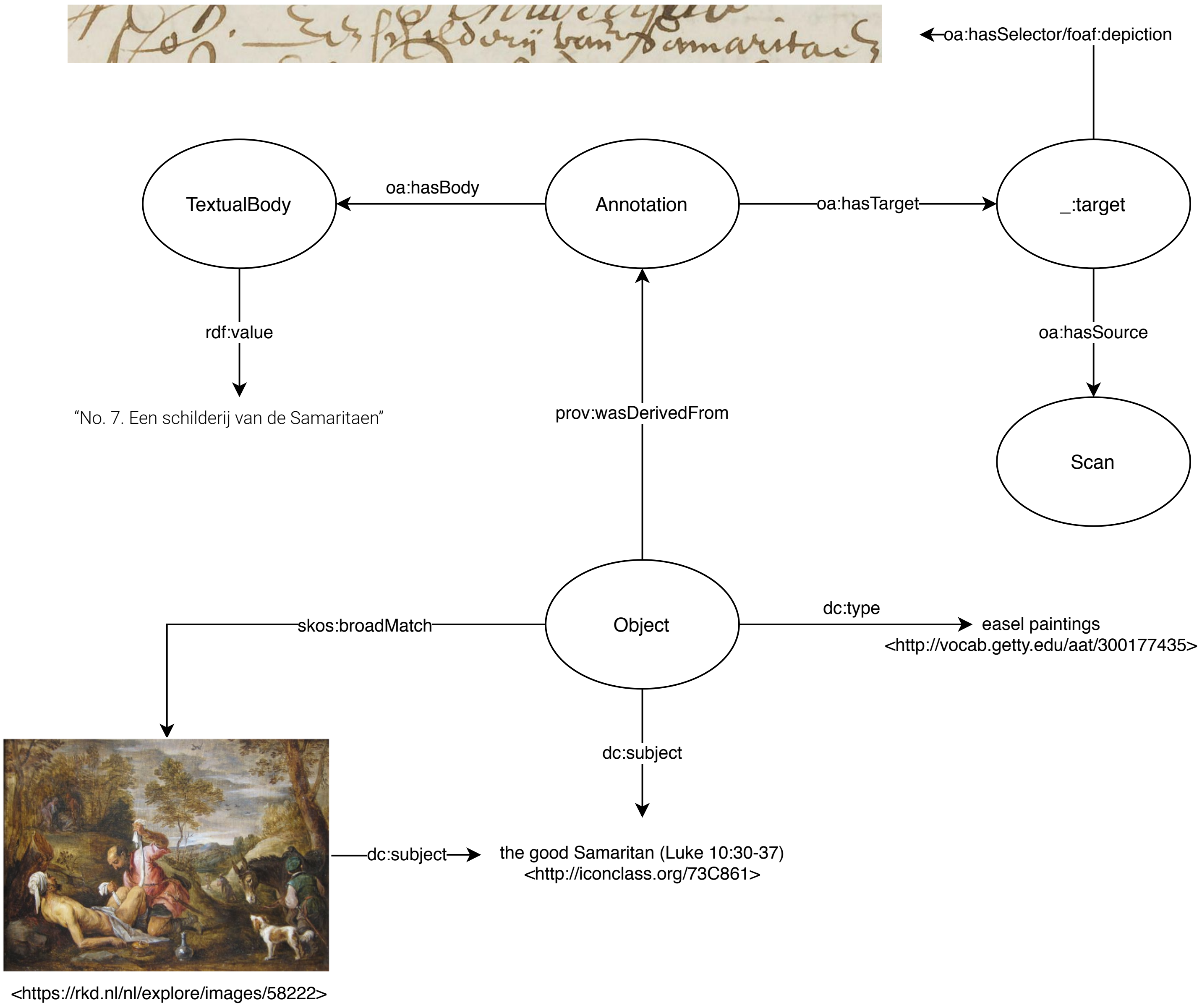
oa:hasSelector/foaf:depiction
TextualBody
oa:hasBody
Annotation
oa:hasTarget
_:target
rdf:value
oa:hasSource
"No. 7. Een schilderij van de Samaritaen"
prov:wasDerivedFrom
Scan
skos:broadMatch
Object
dc:type
easel paintings
<http://vocab.getty.edu/aat/300177435>
dc:subject
dc:subject
the good Samaritan (Luke 10:30-37)
<http://iconclass.org/73C861>
<https://rkd.nl/nl/explore/images/58222>

Figure 2 - Simplified RDF representation of a reference to a painting

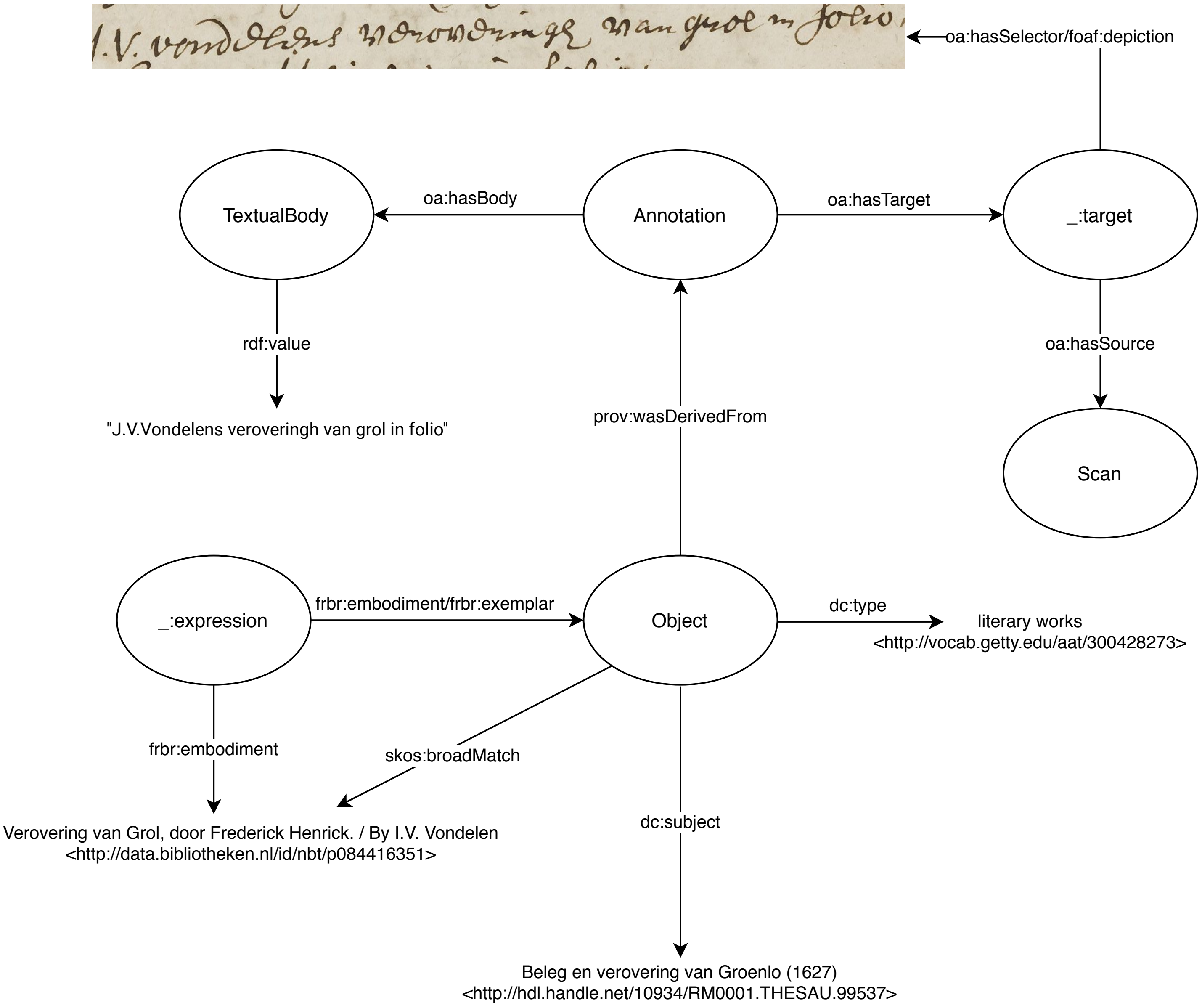
I.V. vondelens veroveringh van grol in folio
oa:hasSelector/foaf:depiction
TextualBody
oa:hasBody
Annotation
oa:hasTarget
_:target
rdf:value
"J.V.Vondelens veroveringh van grol in folio"
prov:wasDerivedFrom
oa:hasSource
Scan
_:expression
frbr:embodiment/frbr:exemplar
Object
dc:type
literary works
<http://vocab.getty.edu/aat/300428273>
frbr:embodiment
skos:broadMatch
Verovering van Grol, door Frederick Henrick. / By I.V. Vondelen
<http://data.bibliotheken.nl/id/nbt/p084416351>
dc:subject
Beleg en verovering van Groenlo (1627)
<http://hdl.handle.net/10934/RM0001.THESAU.99537>

Figure 3 - Simplified representation of a reference to a book/poem