Linked Data Scopes

Victor de Boer Ivette Bonestroo Marijn Koolen and Rik Hoekstra

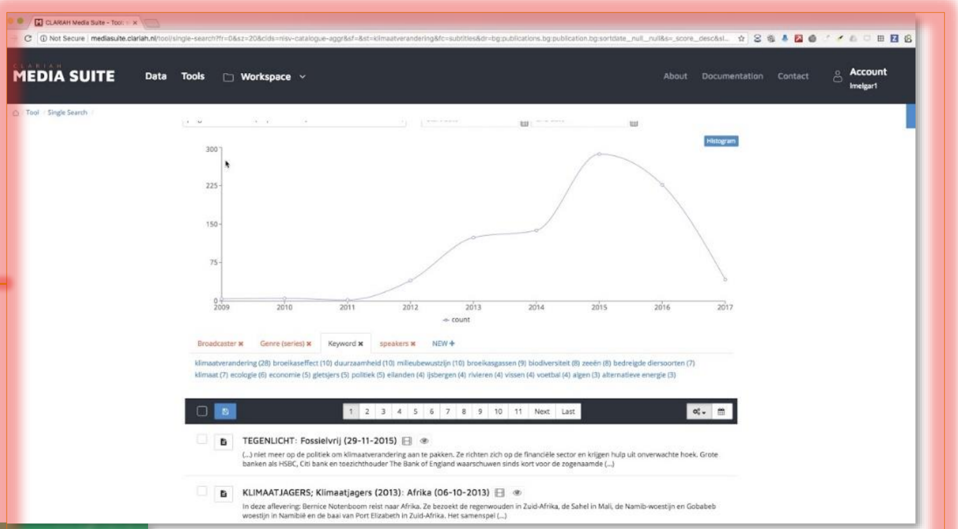
VU VRIJE UNIVERSITEIT AMSTERDAM

KNAW Humanities Cluster

huygens ing

CLARIAH Common Lab Research Infrastructure for the Arts and Humanities

(Digital) Humanities research has increasingly a data component

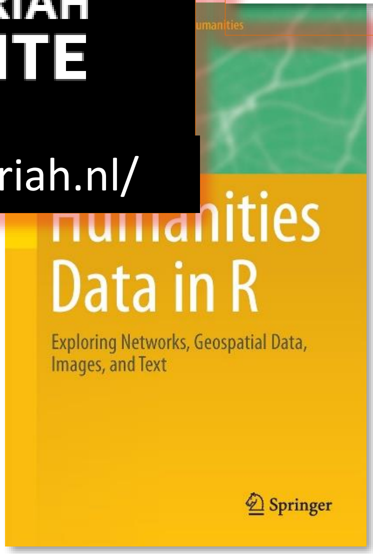


TEGENLICHT: Fossielvrij (29-11-2015)

KLIMAATJAGERS; Klimaatjagers (2013): Afrika (06-10-2013)

CLARIAH MEDIA SUITE

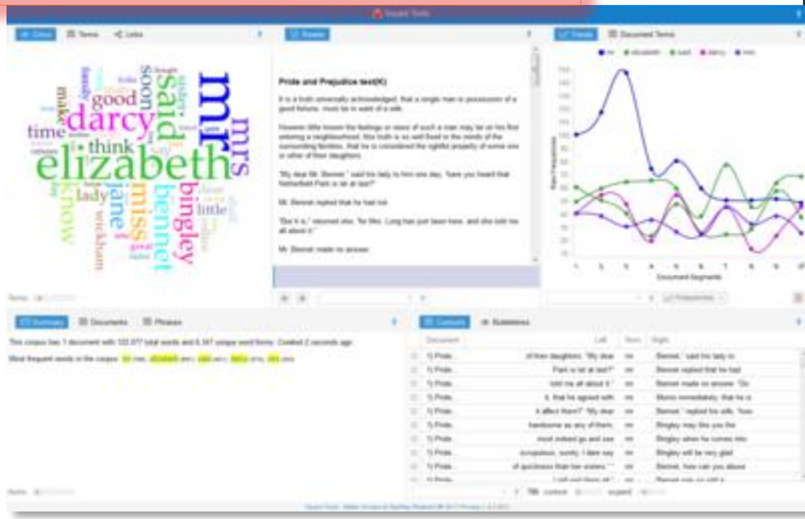
https://mediasuite.clariah.nl/

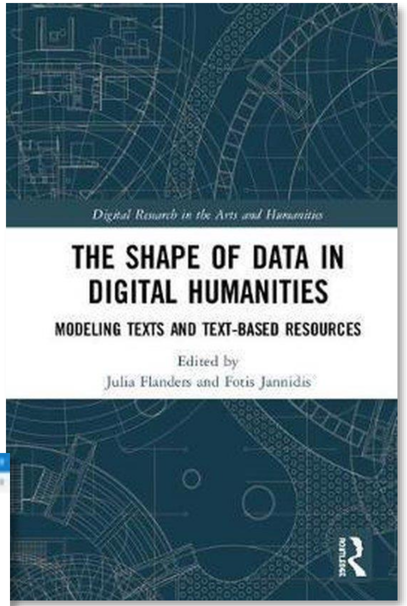


Humanities Data in R

Exploring Networks, Geospatial Data, Images, and Text

Springer





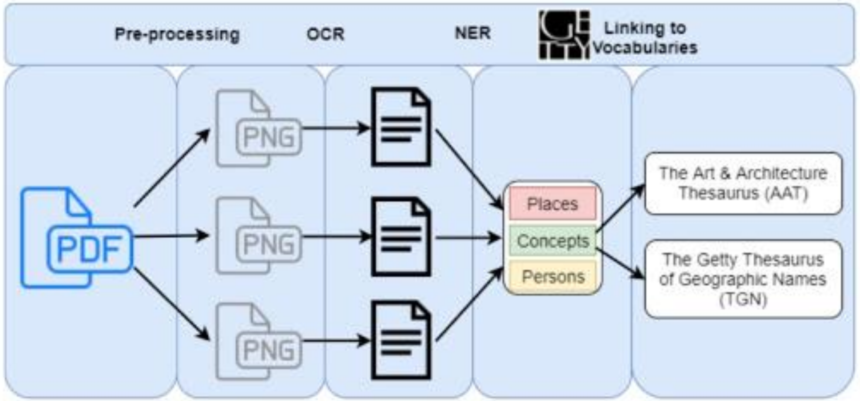
Digital Research in the Arts and Humanities

THE SHAPE OF DATA IN DIGITAL HUMANITIES

MODELING TEXTS AND TEXT-BASED RESOURCES

Edited by
Julia Flanders and Fotis Jannidis

Variety of processing and data manipulation workflows/pipelines


Pre-processing
OCR
NER
Linking to Vocabularies
PDF
PNG
PNG
PNG
Places
Concepts
Persons
The Art & Architecture Thesaurus (AAT)
The Getty Thesaurus of Geographic Names (TGN)

tions.

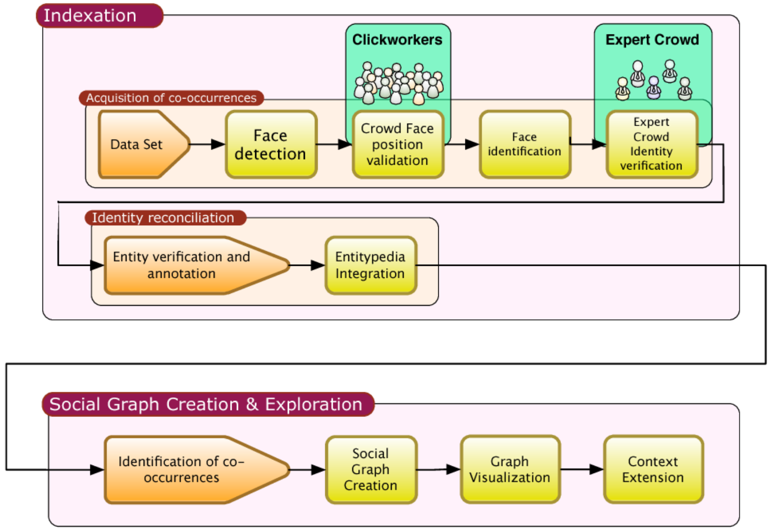

Indexation
Clickworkers
Expert Crowd
Acquisition of co-occurrences
Data Set
Face detection
Crowd Face position validation
Face identification
Expert Crowd Identity verification
Identity reconciliation
Entity verification and annotation
Entitypedia Integration
Social Graph Creation & Exploration
Identification of co-occurrences
Social Graph Creation
Graph Visualization
Context Extension

Fig. 3 High-level view of the HeF indexation pipeline

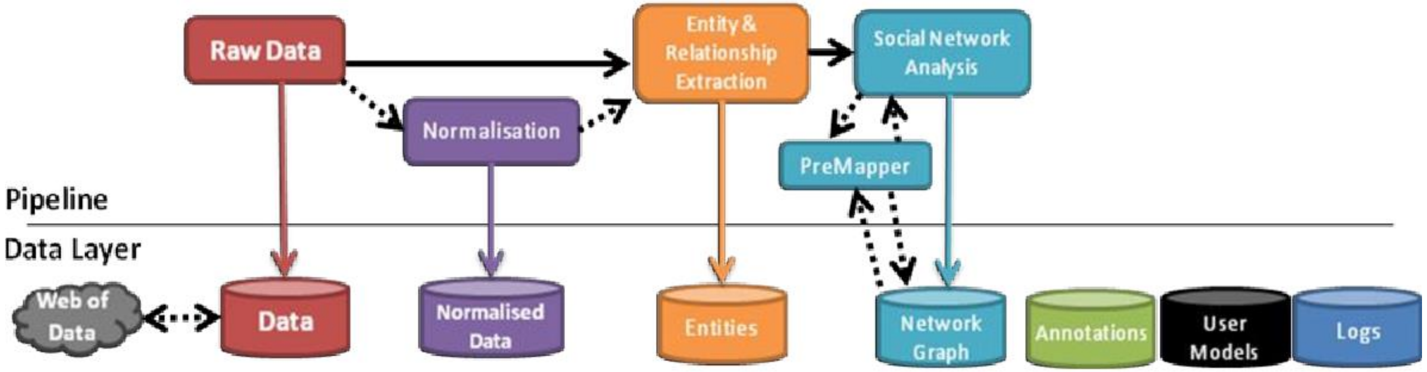
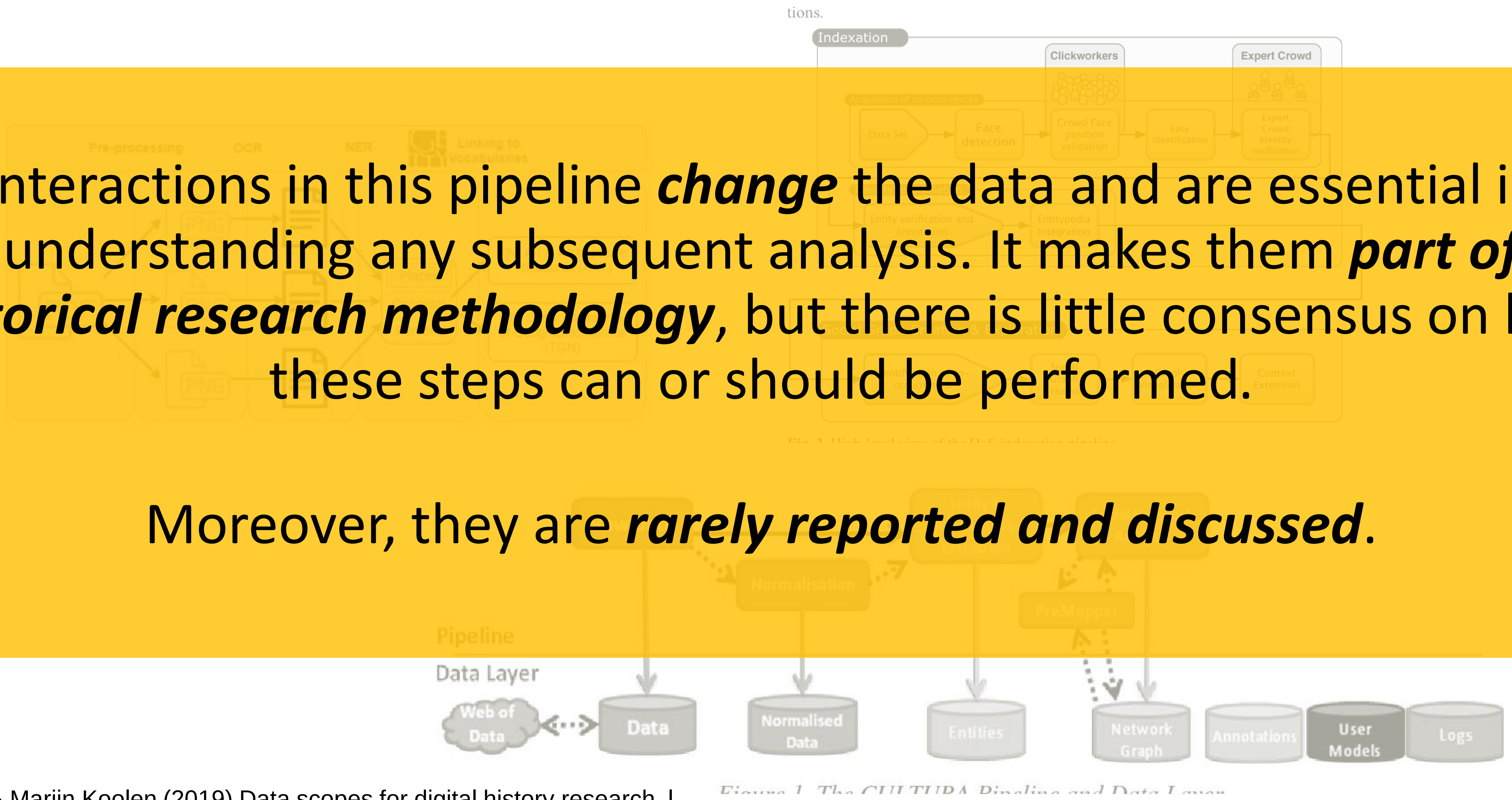


Figure 1. The CULTURA Pipeline and Data Layer

Are these steps recognizable in resulting publications?



Interactions in this pipeline ***change*** the data and are essential in understanding any subsequent analysis. It makes them ***part of historical research methodology***, but there is little consensus on how these steps can or should be performed.

Moreover, they are ***rarely reported and discussed***.

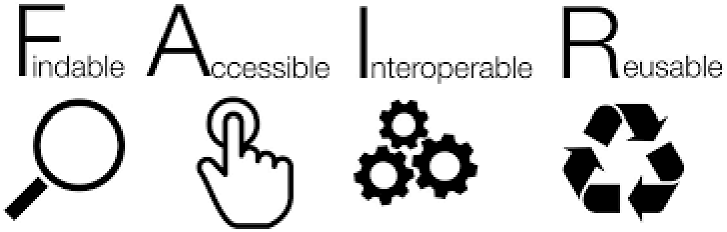
Rik Hoekstra & Marijn Koolen (2019) Data scopes for digital history research, I

Goal: increase reusability and transparency in humanities research output

Through explicit descriptions of data transformations users of the datasets can assess “the context in which the data was created, its quality and validity, and the appropriate conditions for use.” (Groth et al. 2012)

Describe and share

- – Methodology
- – Intermediate results
- – Datasets
- – Data enrichments



Groth, P., Y. Gil, J. Cheney, and S. Miles. 2012. Requirements for provenance on the web.

Preliminary: Data scopes for digital history research

Data scopes are a method to ‘characterize the interaction between researchers and their data and the transformation of a cluster of data into a research instrument.’ (Hoekstra & Koolen,2019)

HISTORICAL METHODS
2019, VOL. 52, NO. 2, 79–94
https://doi.org/10.1080/01615440.2018.1484676

ARTICLES

Routledge
Taylor & Francis Group

Check for updates

Data scopes for digital history research

Rik Hoekstra and Marijn Koolen
Huygens Institute for the History of the Netherlands – KNAW, Amsterdam, Netherlands

ABSTRACT
The term Macroscope has recently been introduced as an instrument to study historical big data using digital tools. In this paper we argue the need for a more elaborate set of concepts to describe and reason about the interactions to select, enrich, connect, analyse and evaluate historical data using digital tools. Interactions change the data and are essential in understanding any subsequent analysis. It makes them part of historical research methodology, but there is little consensus on how these steps can or should be performed. Moreover, they are rarely reported and discussed. We introduce the term data scope as an instrument encompassing these choices and interactions. Elaborating on these processes encourages deeper reflection on and discussion of the interactions and their consequences for research outcomes.

KEYWORDS
Digital history; methodology; historical research process; data interaction

Introduction

Researchers need instruments to collect and study their data. Astronomers use telescopes, biologists use microscopes. With the coming of big data, a number of researchers decided to use the term macroscope to qualify the instrument needed for studying data clusters, large amounts of data from one or more datasets. This macroscopic approach is very different from traditional historical research practice on which scholarly argumentation and citation is based, but is still undertheorized, as Ted Underwood argues (Underwood 2014). Historians conceive their data as a reaction to their research questions. They also interact with their datasets and interactively enrich and enlarge them as research progresses. Each step of selection, enrichment, and classification represents a choice that is based on explorations and interpretations of the data. These interactions change the data and are essential in understanding any subsequent analysis, which makes them part of research methodology, but there is little consensus on how these steps can or should be performed. Moreover, they are rarely reported and discussed.

Historical research

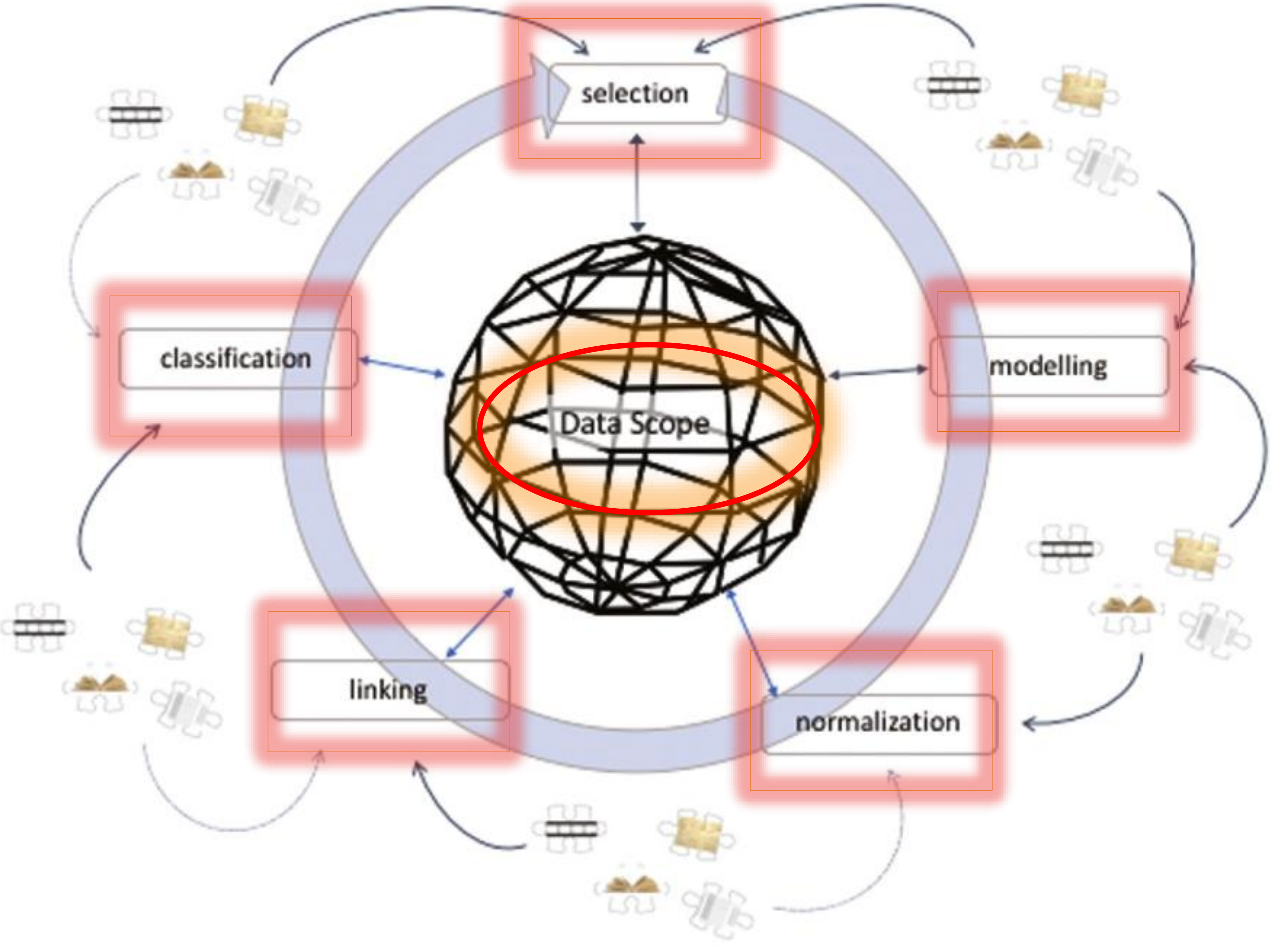
historians increasingly use digital environments and data, research practices shift. Digital environments offer different ways to search and discover sources, that, as Lara Putnam warns us in “The Transnational and the Text-Searchable” (Putnam 2016), make it easy to find highly relevant documents, but often present search results on their own without their context, missing the built-in contextualization of discovery in physical archives. This is also reflected in referencing practices that are still based on hard-copy, even when argumentation is based on a digital search process. “To take a single example of this disconnect between research process and representation, many of us use and cite eighteenth and nineteenth-century newspapers as simple hard-copy references without mention of how we navigated to the specific article, page and issue. In doing so, we actively misrepresent the limitations within which we are working” (Hitchcock 2013, 12).

Several scholars, including Fred Gibbs and Trevor Owens (2013), Ted Underwood (2014), Graham, Milligan and Weingart (2015) and

Rik Hoekstra & Marijn Koolen (2019) Data scopes for digital history research, Historical Methods: A Journal of Quantitative and Interdisciplinary History, 52:2, 79-94,
DOI: 10.1080/01615440.2018.1484676

Data scope denotes 5 data manipulation steps

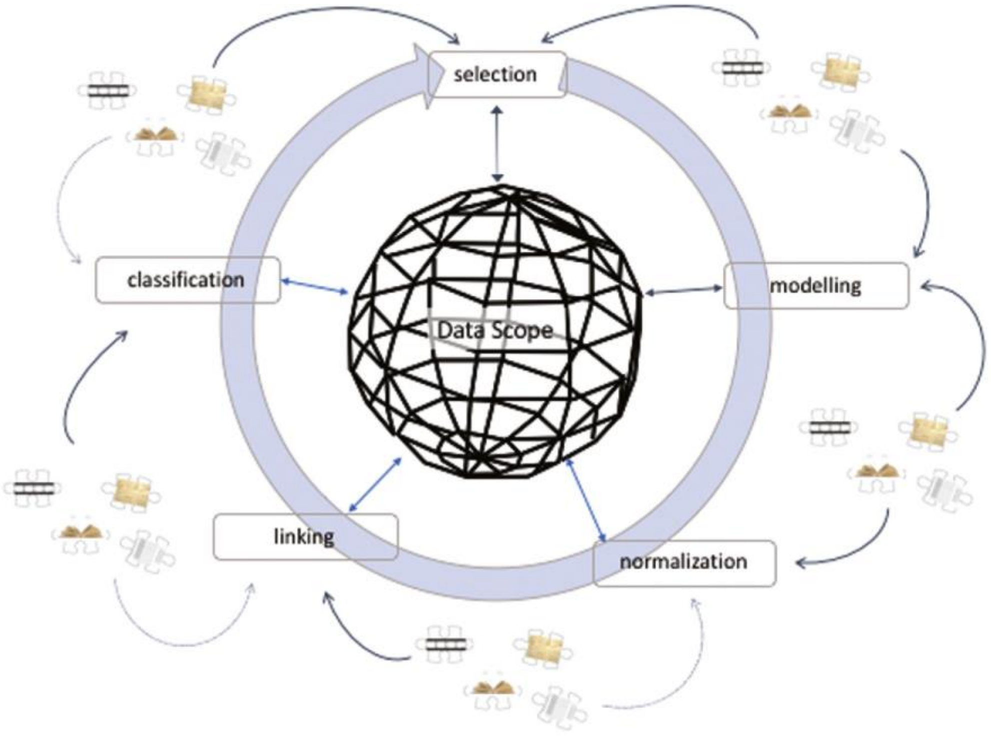
1. **Selection**: which data and sources are selected? (corpus forming)
2. **Modelling**: how are the relevant elements in sources represented? (using implicit or explicit models)
3. **Normalization**: how are surface forms mapped to a normalized form? (e.g. "Firstname, Lastname")
4. **Linking**: what explicit internal and external connections are established? (Includes deduplication, NE resolution etc.)
5. **Classification**: how are objects grouped or categorized? (includes internal or external schemes or theories)



Rik Hoekstra & Marijn Koolen (2019) Data scopes for digital history research,

This paper: Contribution 1

From qualitative conceptualization …



…to reusable, explicit representation

Machine readable
FAIR
Shareable

W3C® RDF

The Data scopes ontology

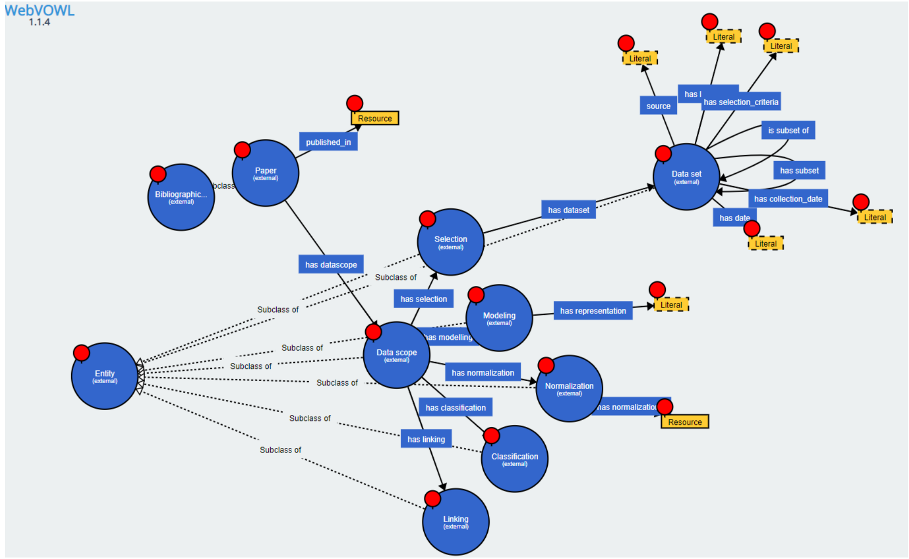
<http://w3id.org/datascope#>
permanent identifiers

Central class is **dsont:DataScope**

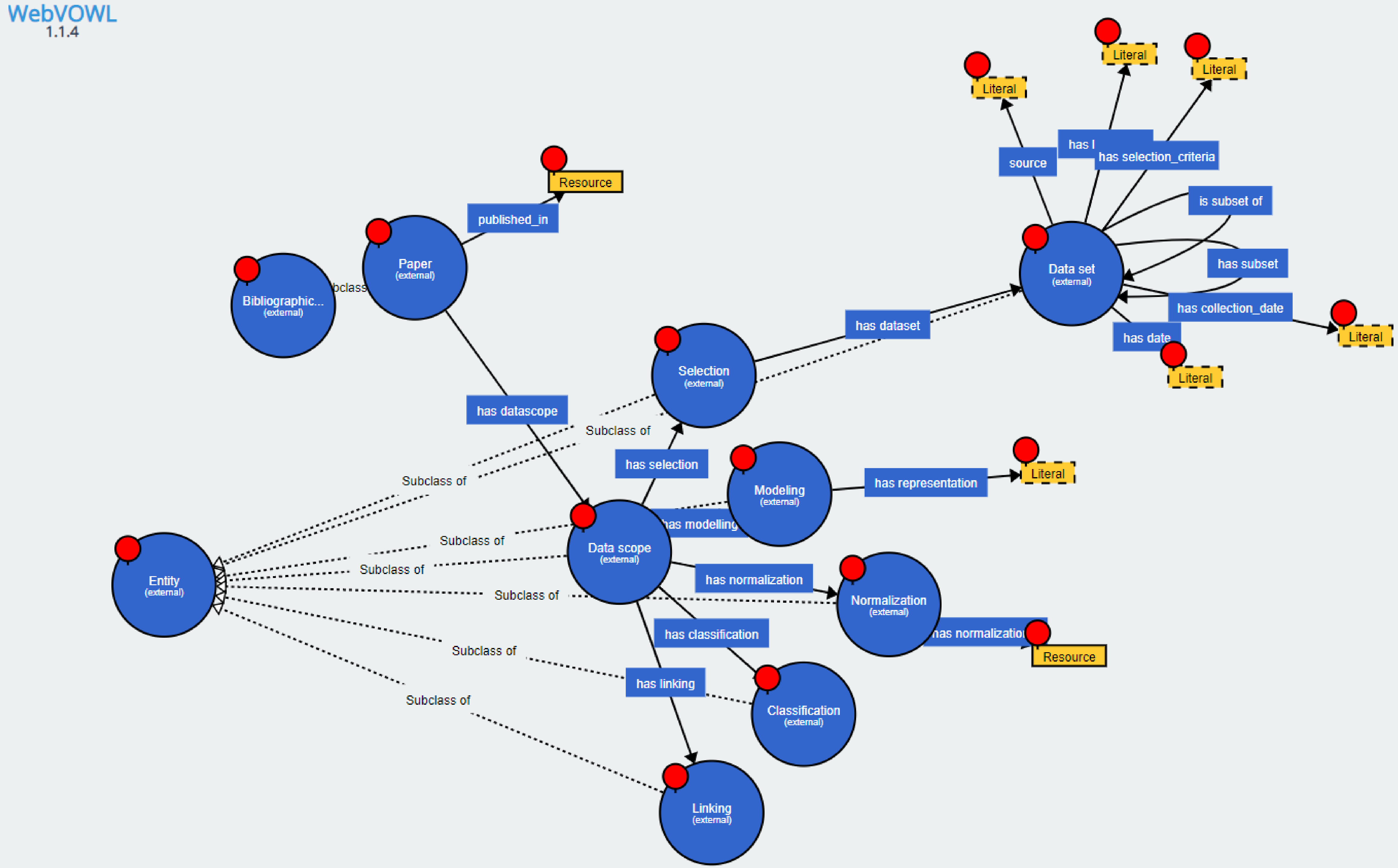
Classes for the five main manipulation activities

Additional classes and properties for Datasets, selection steps, linksets, Concept Schemes etc.

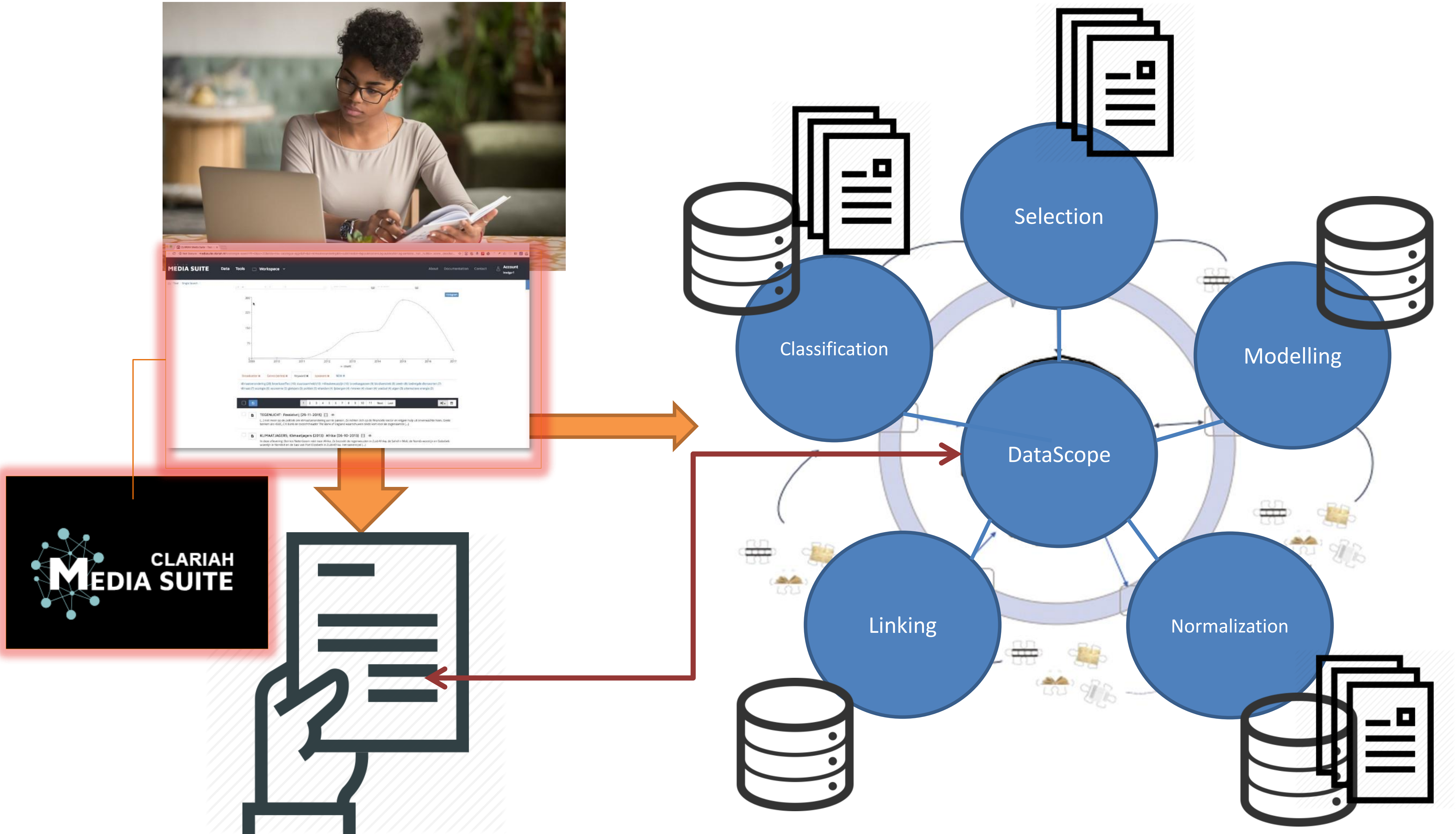
Aligned with Dcterms, PROV



http://biktorrr.github.io/datascope/


WebVOWL
1.1.4
Literal
Literal
Literal
Resource
has l
has selection_criteria
source
published_in
is subset of
Paper
(external)
Data set
(external)
bclass
has subset
Bibliographic...
(external)
has collection_date
has dataset
Literal
has date
Selection
(external)
Literal
has datascope
Subclass of
has selection
Literal
Modeling
(external)
has representation
Subclass of
has modelling
Data scope
(external)
Subclass of
Entity
(external)
Subclass of
has normalization
Subclass of
Normalization
(external)
has normalizatio
has classification
Resource
Subclass of
has linking
Classification
(external)
Subclass of
Linking
(external)

Use of The Data scopes ontology

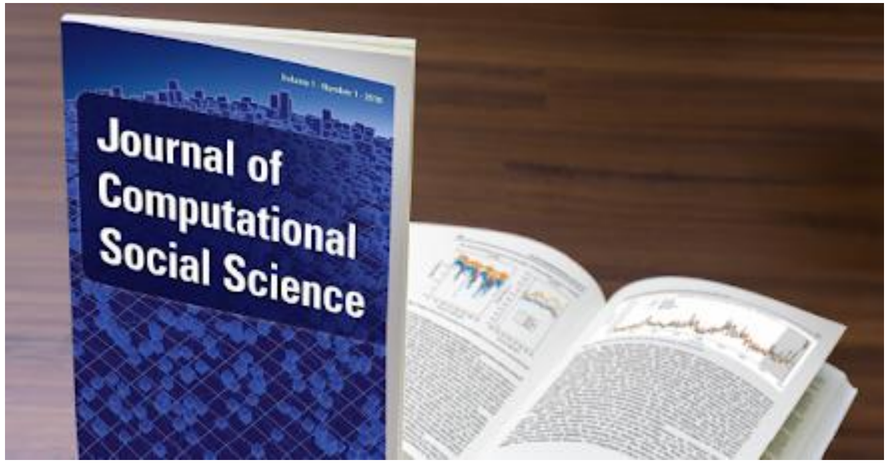


This paper: Contribution 2

Initial validation of the model in two research domains: Digital Humanities and Computational Social Sciences

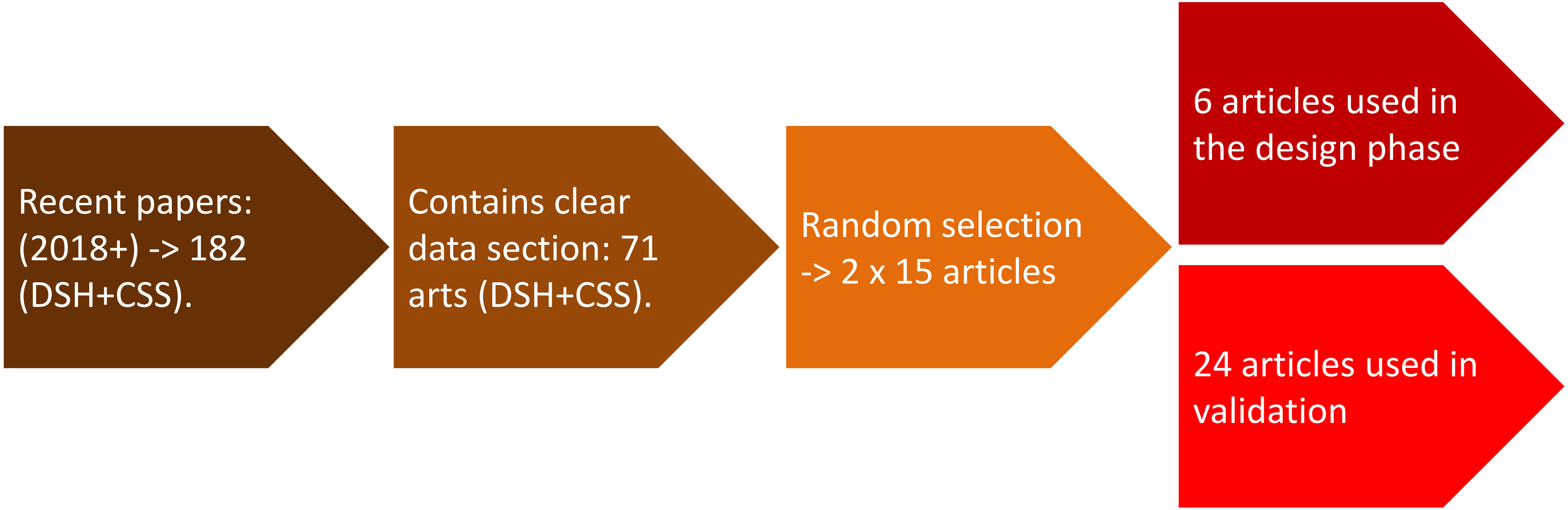
DSH
Digital Scholarship in the Humanities
Journal of the Alliance of Digital Humanities Organizations
OXFORD

Digital Scholarship for the Humanities (DSH)



Computational Social Science (CSS)

Method



Recent papers: (2018+) -> 182 (DSH+CSS).

Contains clear data section: 71 arts (DSH+CSS).

Random selection -> 2 x 15 articles

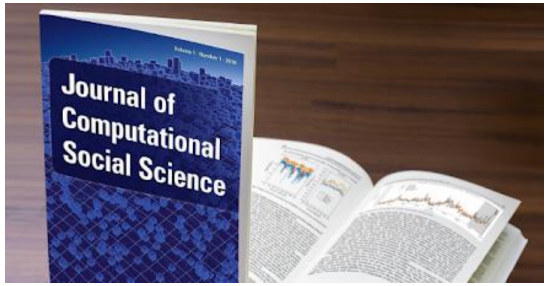
6 articles used in the design phase

24 articles used in validation

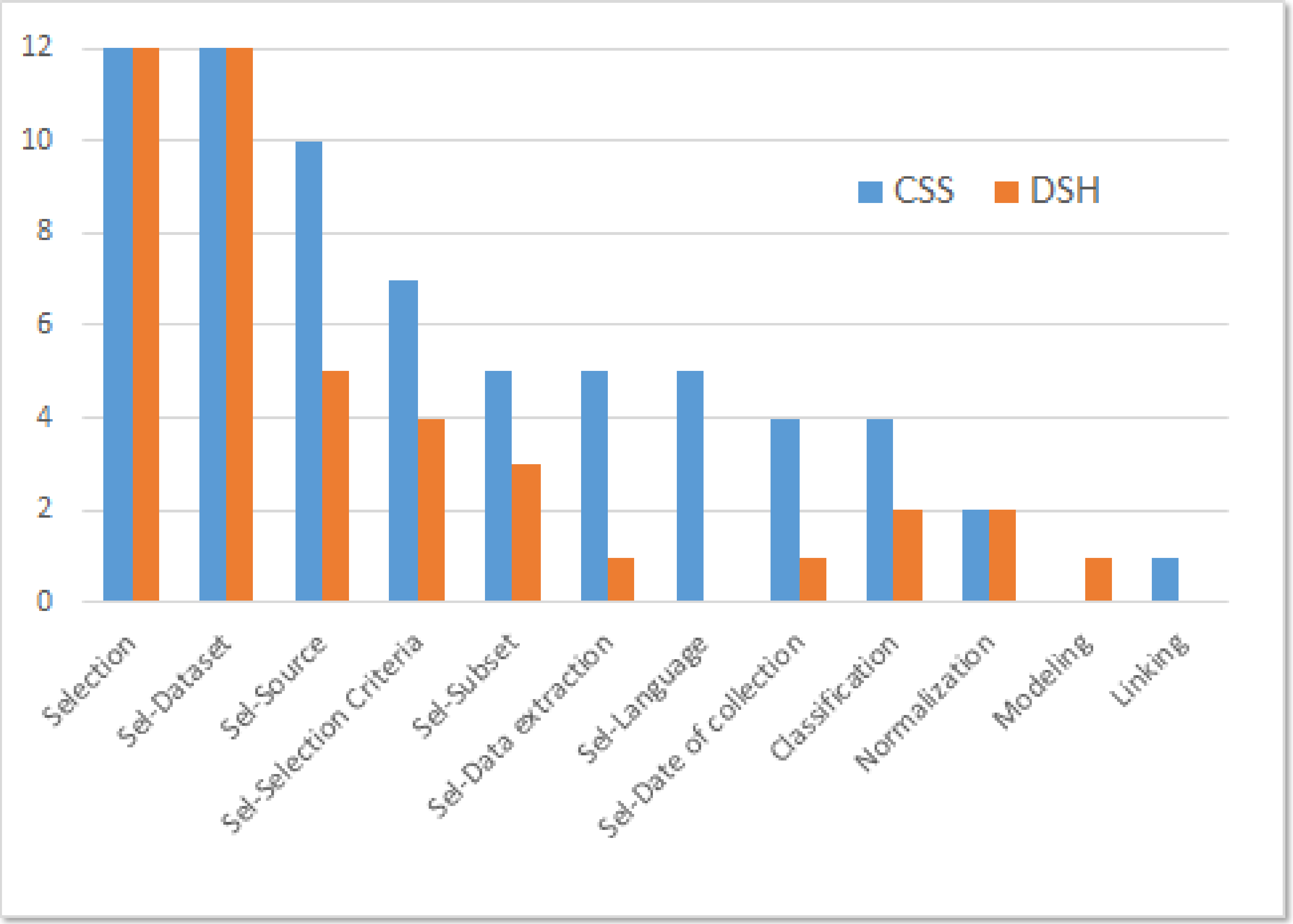
Set of coding guidelines

Two independent coders

DSH
Digital Scholarship in the Humanities
Journal of the Alliance of Digital Humanities Organizations
OXFORD



Results



Each element appears at least once

But most papers do not have complete descriptions

Selection most used (as hypothesized by Hoekstra & Koolen)

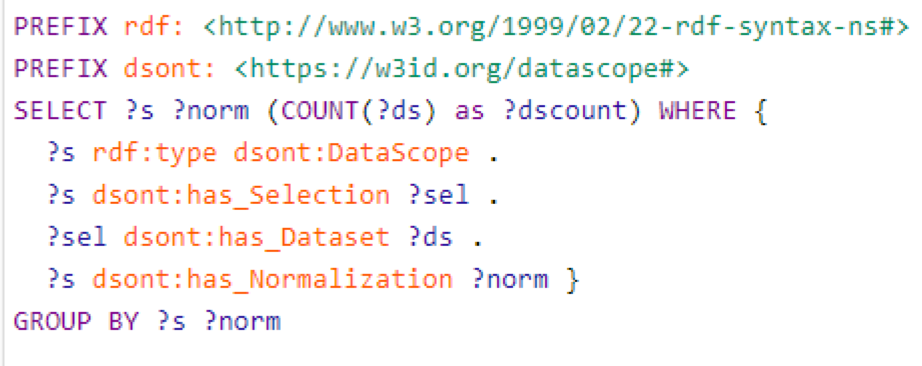
Can be used to compare fields

CSS papers have more complete descriptions

Results represented as RDF data

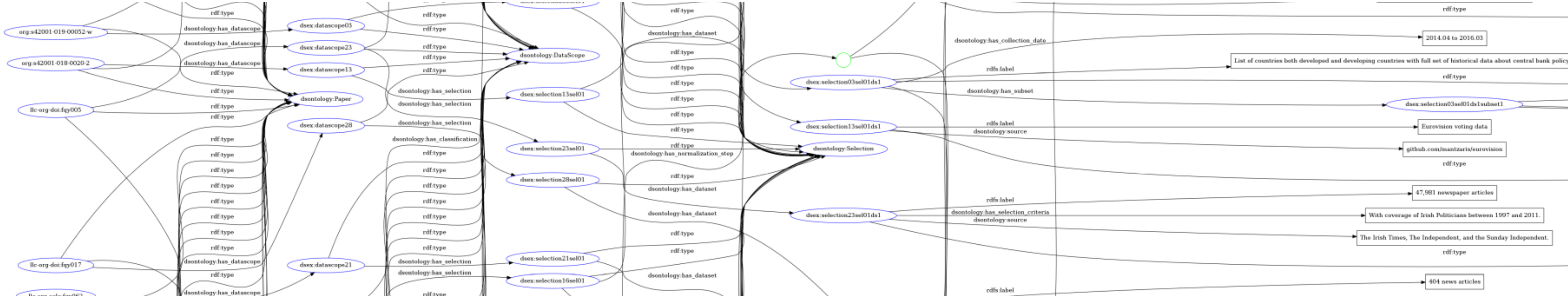
511 RDF triples describing the datascopes of 24 papers, relating back to the original papers

SPARQLable at
https://semanticweb.cs.vu.nl/test/query



```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX dsont: <https://w3id.org/datascope#>
SELECT ?s ?norm (COUNT(?ds) as ?dscount) WHERE {
  ?s rdf:type dsont:DataScope .
  ?s dsont:has_Selection ?sel .
  ?sel dsont:has_Dataset ?ds .
  ?s dsont:has_Normalization ?norm }
GROUP BY ?s ?norm
```

“Show me the # of datasets for which a normalization step is registered “



Next steps

Further refinement and validation

Integrate these with existing tools

- CLARIAH mediasuite
- OpenRefine…

Allow direct and FAIR publishing

- Using Nanopublications/nanobench

Linking to other ontologies

- Workflow ontologies…

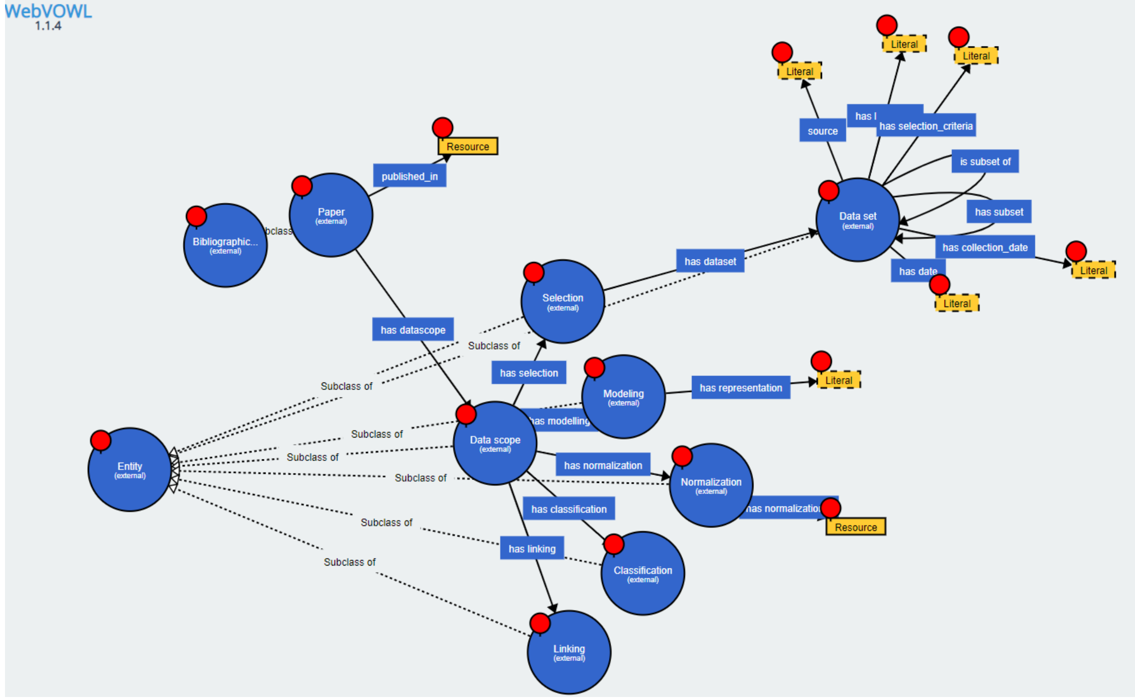
Thank you!

http://biktorrr.github.io/datascope/

http://mediasuite.clariah.nl

@victordeboer

v.de.boer@vu.nl


WebVOWL
1.1.4
Resource
published_in
Paper
(external)
Bibliographic...
(external)
bclass
has datascope
Subclass of
Subclass of
Selection
(external)
has selection
Entity
(external)
Subclass of
Subclass of
Subclass of
Subclass of
Subclass of
Data scope
(external)
has modelling
Modeling
(external)
has representation
Literal
has normalization
Normalization
(external)
has normalizatio
Resource
has classification
Classification
(external)
has linking
Linking
(external)
has dataset
Data set
(external)
source
Literal
has l
has selection_criteria
Literal
Literal
is subset of
has subset
has collection_date
Literal
has date
Literal