CLARIAH
Common Lab Research Infrastructure for the Arts and Humanities

CLARIAH MEDIA SUITE

Analyzing AV Sources as Data

Experiences Gained with Designing the CLARIAH Media Suite

Julia Noordegraaf
University of Amsterdam

Universiteit Utrecht

Universiteit van Amsterdam

VU University Amsterdam

Beeld en Geluid

Erasmus Universiteit Rotterdam

Introduction

- DSC prerequisite for broad take-up DH
- Learning by doing:
  - CREATE program
  - CLARIAH infrastructure
- Examples:
  - metadata criticism
  - data and tool criticism

Team WP5



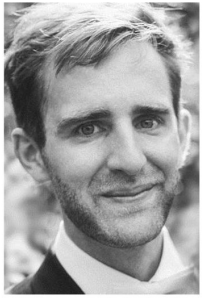














UNIVERSITEIT VAN AMSTERDAM

Universiteit Utrecht

VU VRIJE UNIVERSITEIT AMSTERDAM

BEELD EN GELUID

ERASMUS UNIVERSITEIT ROTTERDAM

CLARIAH
MEDIA SUITE

Data Tools

About Documentation Contact Login

The Media Suite is a research environment of the Dutch infrastructure for digital humanities and social sciences (CLARIAH) which aims to serve the needs of scholars who use audiovisual media by providing access to audiovisual collections and their contextual data.

READ MORE

Data

All available collections and their data are registered in a common inventory.

CHECKOUT THE DATA

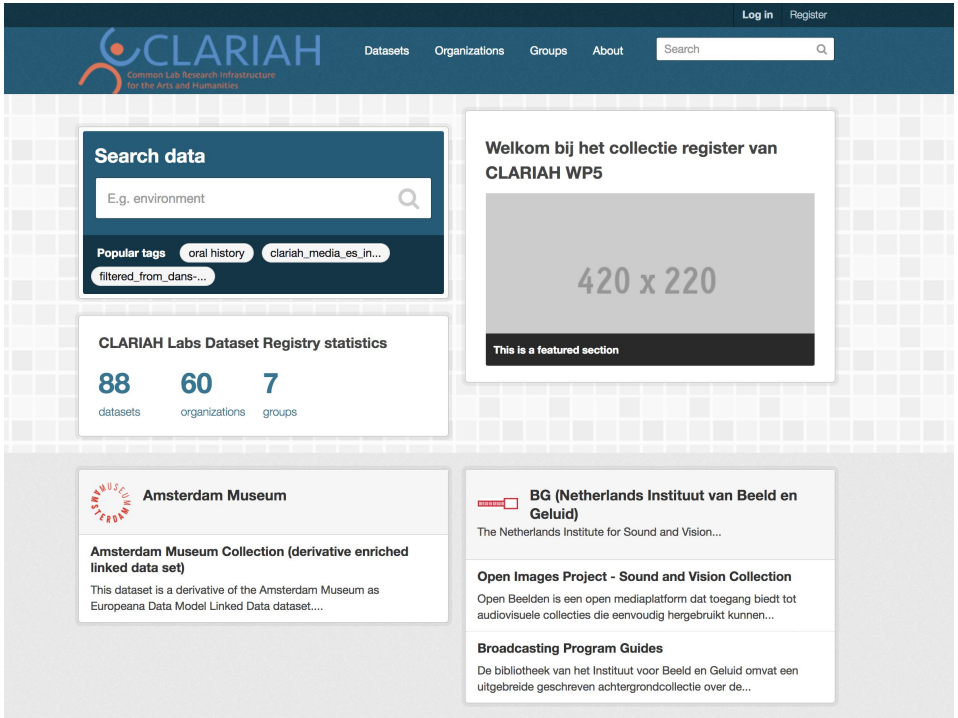
Tools

Tools allow researchers to perform tasks with the available data.

USE THE TOOLS

Data in the Media suite

Overview of available collections including collection descriptions (CKAN)

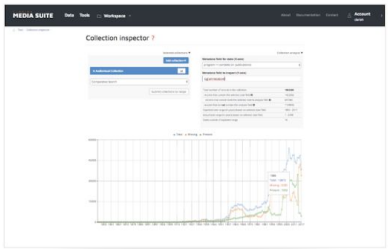


CLARIAH MEDIA SUITE

Data | Tools | Workspace | About | Documentation | Contact | Account jnoorde1

Tools /

?

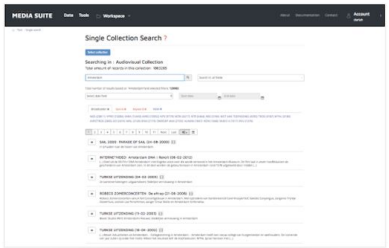


Inspect

DETAILS OPEN

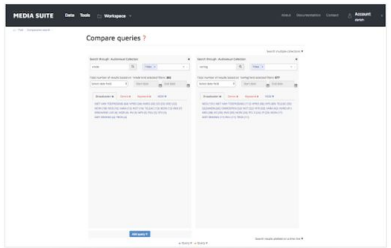
Explore

DETAILS OPEN



Search

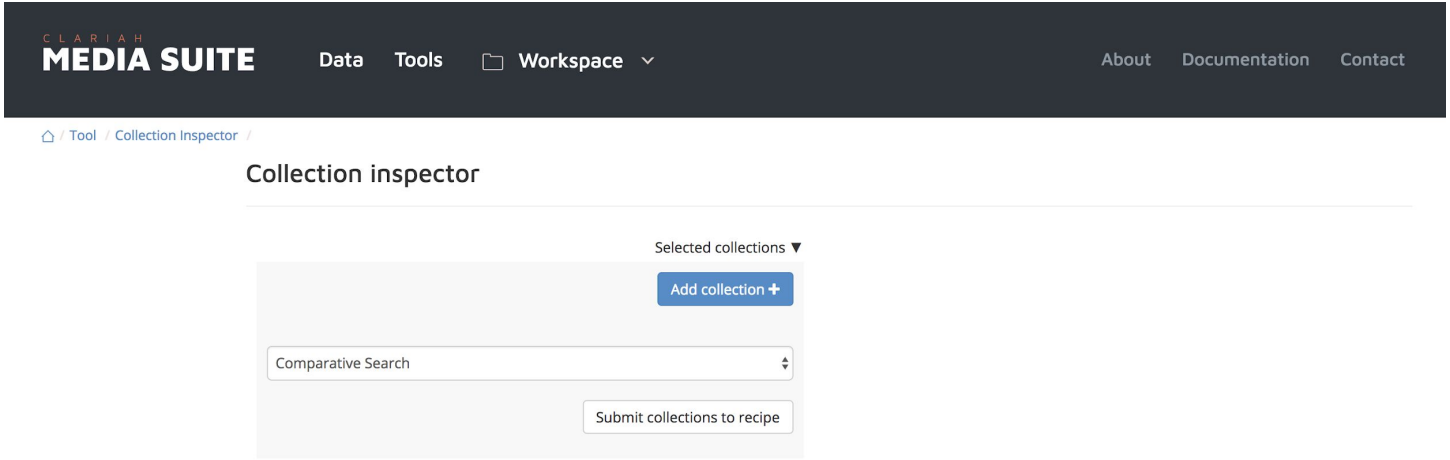
DETAILS OPEN



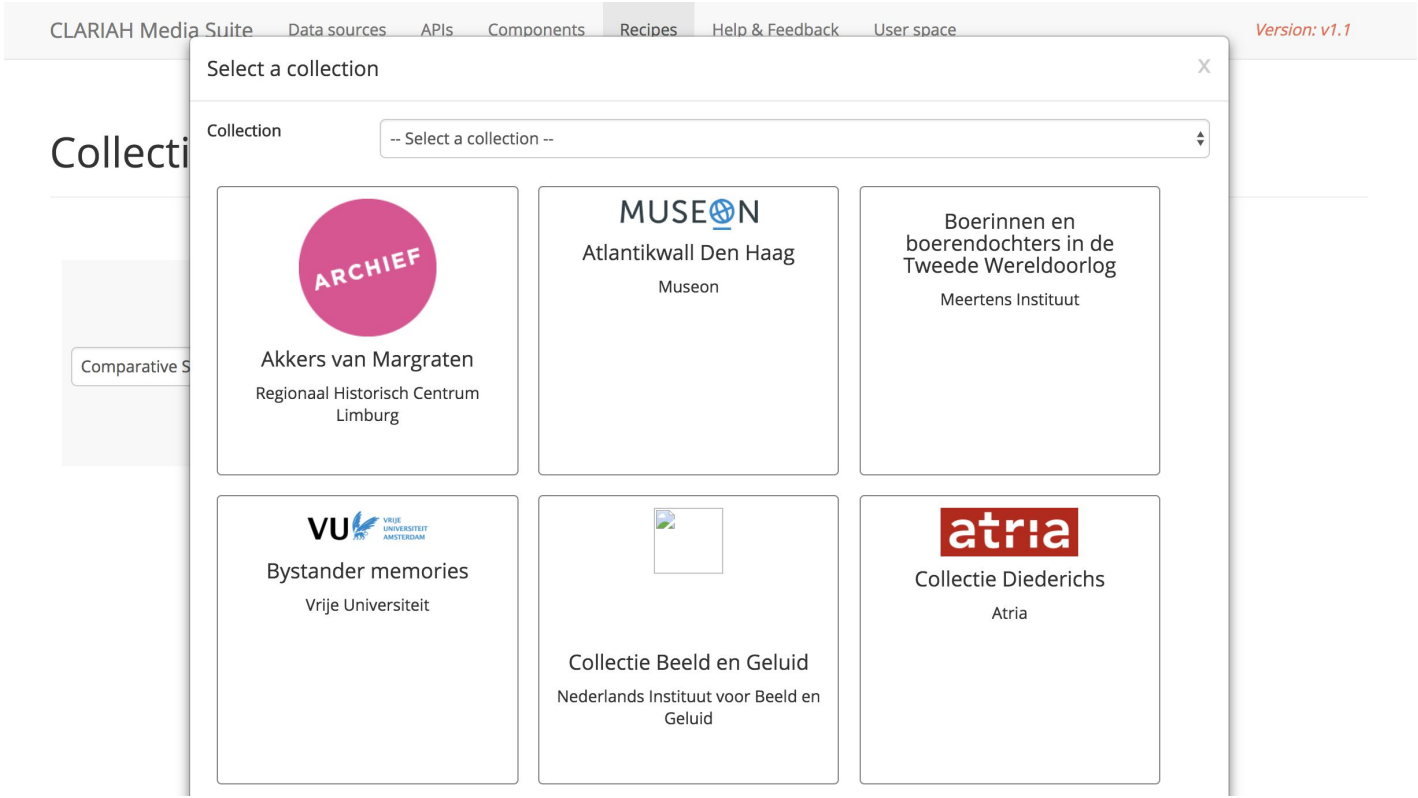
Compare

DETAILS OPEN

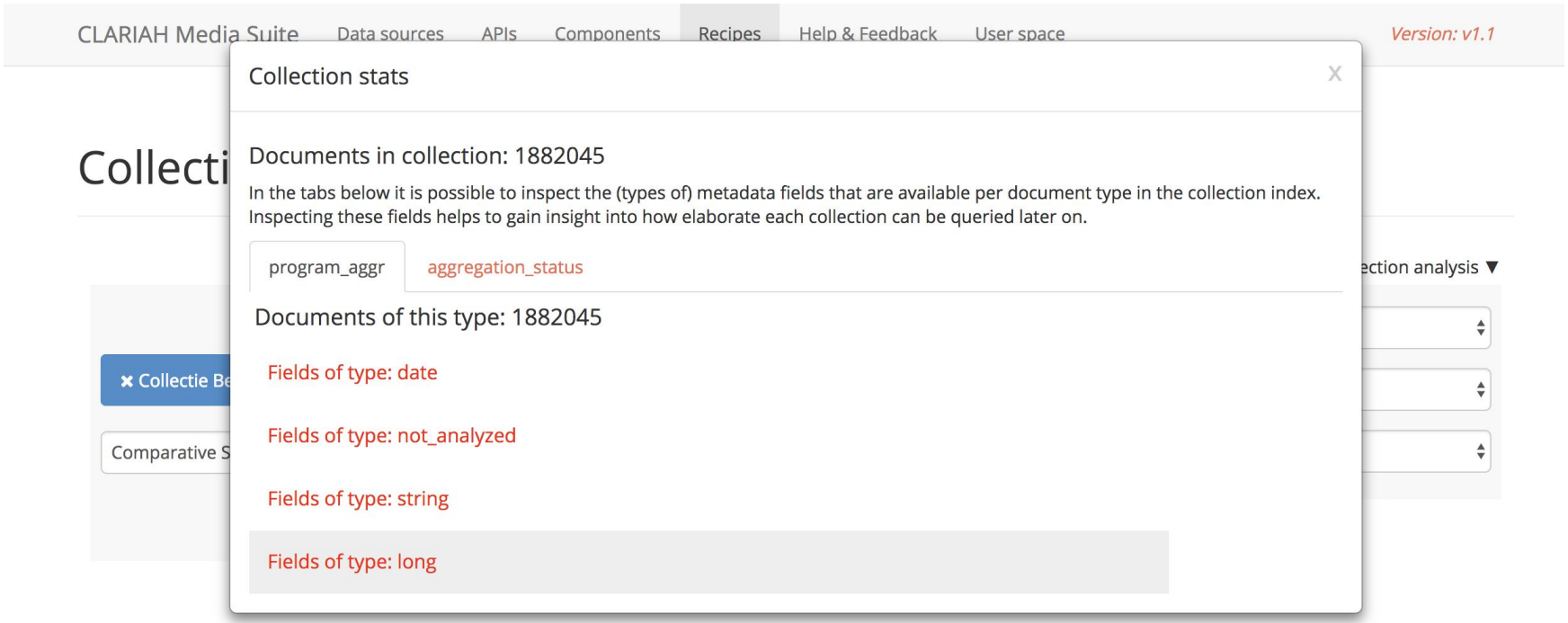
Collection Inspector Tool

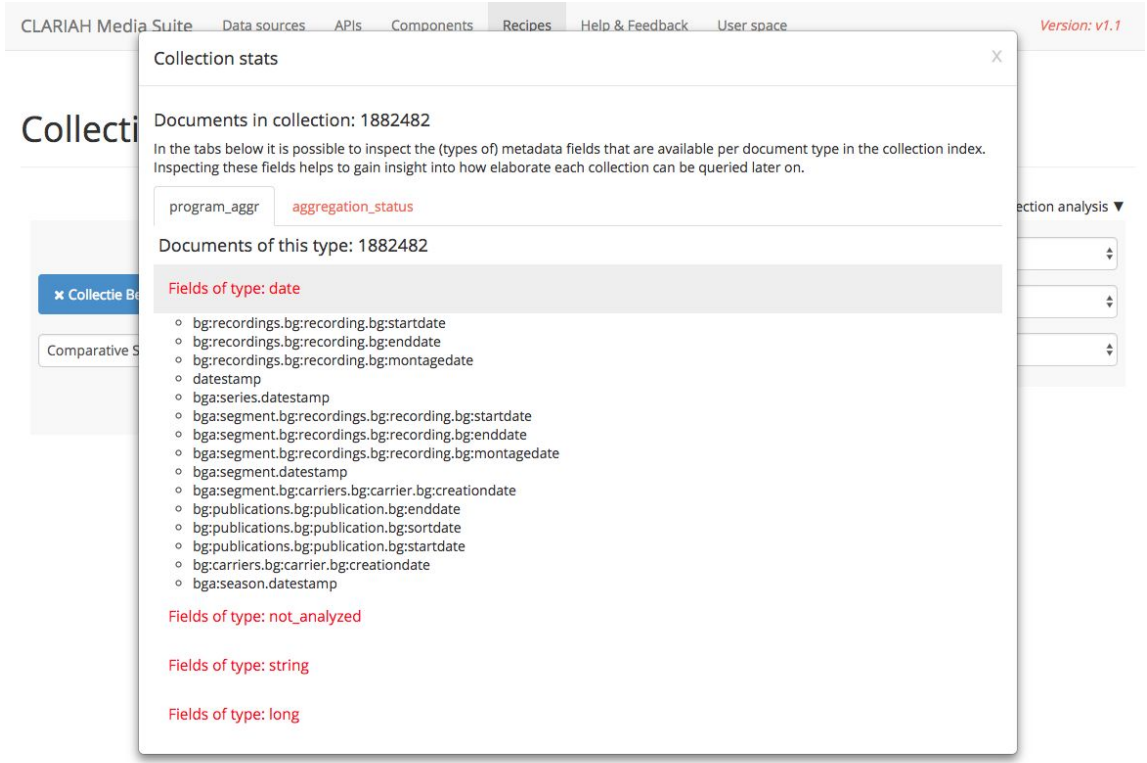


1. Selecting a collection

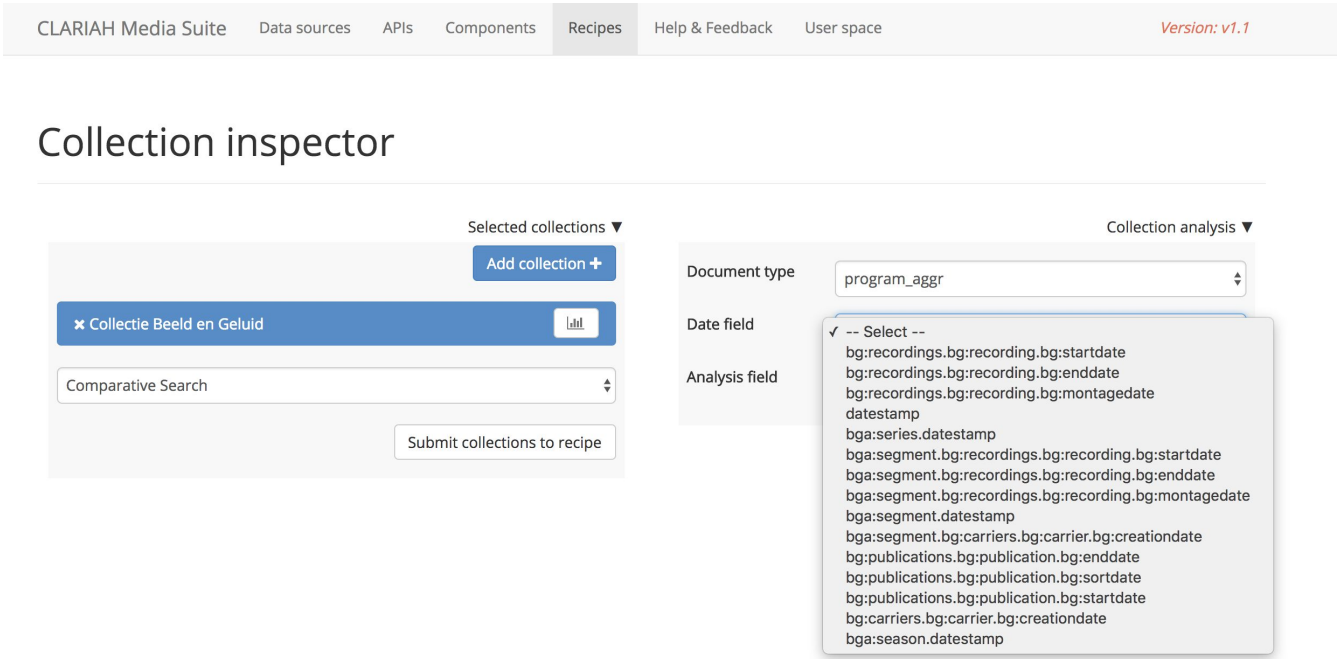


2. Inspecting available metadata fields




CLARIAH Media Suite
Data sources
APIs
Components
Recipes
Help & Feedback
User space
Version: v1.1
Collecti
ection analysis ▼
× Collectie Be
Comparative S
Collection stats
X
Documents in collection: 1882482
In the tabs below it is possible to inspect the (types of) metadata fields that are available per document type in the collection index. Inspecting these fields helps to gain insight into how elaborate each collection can be queried later on.
program_aggr
aggregation_status
Documents of this type: 1882482
Fields of type: date
bg:recordings.bg:recording.bg:startdate
bg:recordings.bg:recording.bg:enddate
bg:recordings.bg:recording.bg:montagedate
datestamp
bga:series.datestamp
bga:segment.bg:recordings.bg:recording.bg:startdate
bga:segment.bg:recordings.bg:recording.bg:enddate
bga:segment.bg:recordings.bg:recording.bg:montagedate
bga:segment.datestamp
bga:segment.bg:carriers.bg:carrier.bg:creationdate
bg:publications.bg:publication.bg:enddate
bg:publications.bg:publication.bg:sortdate
bg:publications.bg:publication.bg:startdate
bg:carriers.bg:carrier.bg:creationdate
bga:season.datestamp
Fields of type: not_analyzed
Fields of type: string
Fields of type: long

3. Selecting date & analysis fields



Collection inspector

Selected collections ▼

Add collection +

× Collectie Beeld en Geluid

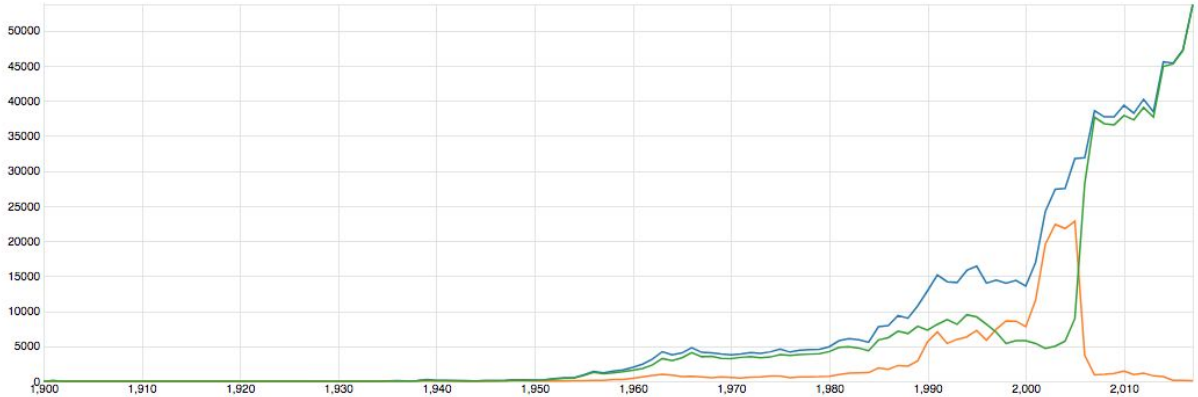
Comparative Search

Submit collections to recipe

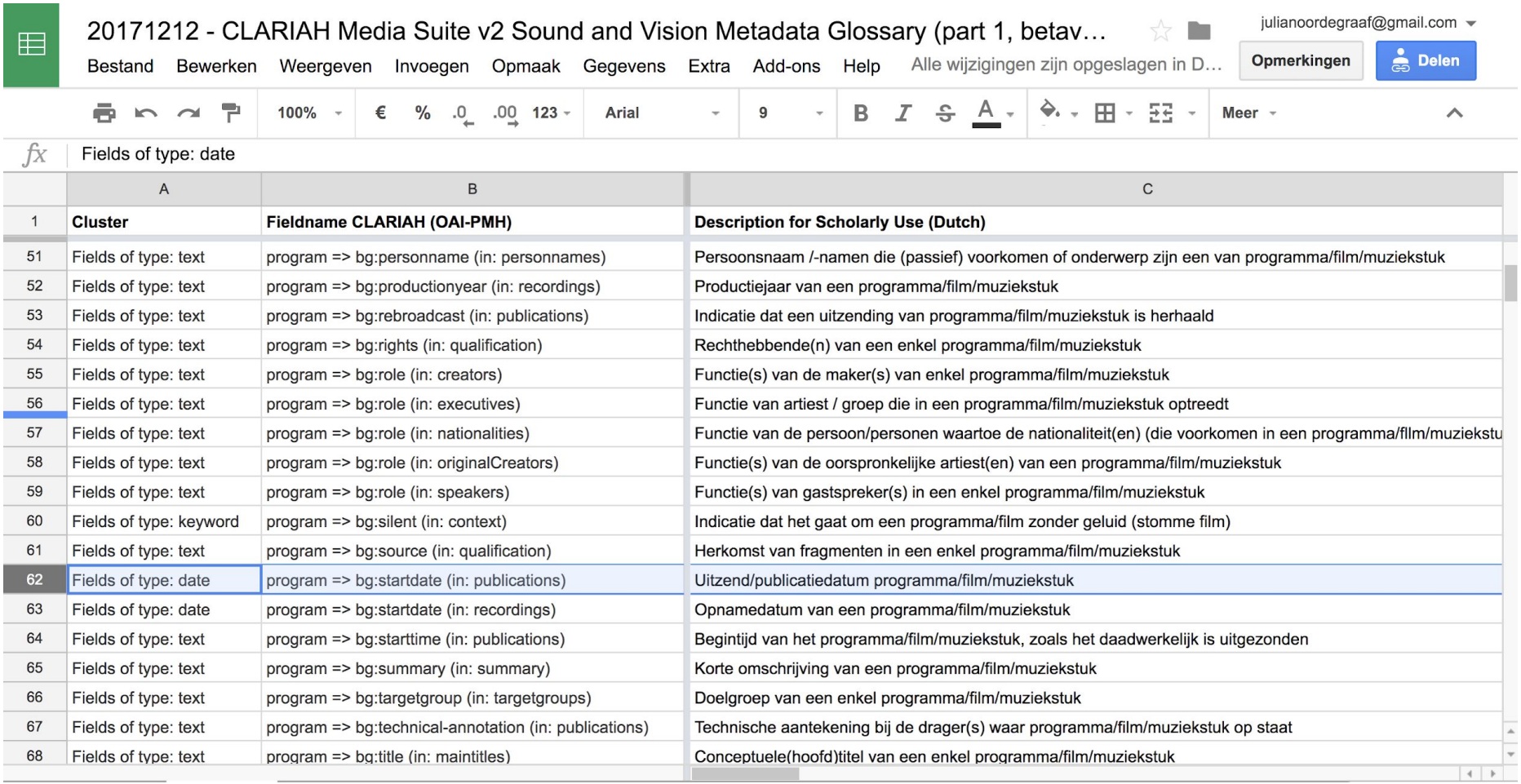
Collection analysis ▼

Document type	program_aggr
Date field	bg:publications.bg:publication.bg:startdate
Analysis field	bg:genres.bg:genre.raw

Document with/without date	346618 / 1535864
Document with/without analysis field	346618 / 1535864
Date range (years)	1900 - 2017
Dates outside range	2



Total | Present | Missing


20171212 - CLARIAH Media Suite v2 Sound and Vision Metadata Glossary (part 1, betav…

julianoordegraaf@gmail.com

Bestand Bewerken Weergeven Invoegen Opmaak Gegevens Extra Add-ons Help Alle wijzigingen zijn opgeslagen in D…

Opmerkingen Delen

100% € % .0 .00 123 Arial 9 Meer

fx Fields of type: date

	A	B	C
1	**Cluster**	**Fieldname CLARIAH (OAI-PMH)**	**Description for Scholarly Use (Dutch)**
51	Fields of type: text	program => bg:personname (in: personnames)	Persoonsnaam /-namen die (passief) voorkomen of onderwerp zijn een van programma/film/muziekstuk
52	Fields of type: text	program => bg:productionyear (in: recordings)	Productiejaar van een programma/film/muziekstuk
53	Fields of type: text	program => bg:rebroadcast (in: publications)	Indicatie dat een uitzending van programma/film/muziekstuk is herhaald
54	Fields of type: text	program => bg:rights (in: qualification)	Rechthebbende(n) van een enkel programma/film/muziekstuk
55	Fields of type: text	program => bg:role (in: creators)	Functie(s) van de maker(s) van enkel programma/film/muziekstuk
56	Fields of type: text	program => bg:role (in: executives)	Functie van artiest / groep die in een programma/film/muziekstuk optreedt
57	Fields of type: text	program => bg:role (in: nationalities)	Functie van de persoon/personen waartoe de nationaliteit(en) (die voorkomen in een programma/fllm/muziekstu
58	Fields of type: text	program => bg:role (in: originalCreators)	Functie(s) van de oorspronkelijke artiest(en) van een programma/film/muziekstuk
59	Fields of type: text	program => bg:role (in: speakers)	Functie(s) van gastspreker(s) in een enkel programma/film/muziekstuk
60	Fields of type: keyword	program => bg:silent (in: context)	Indicatie dat het gaat om een programma/film zonder geluid (stomme film)
61	Fields of type: text	program => bg:source (in: qualification)	Herkomst van fragmenten in een enkel programma/film/muziekstuk
62	Fields of type: date	program => bg:startdate (in: publications)	Uitzend/publicatiedatum programma/film/muziekstuk
63	Fields of type: date	program => bg:startdate (in: recordings)	Opnamedatum van een programma/film/muziekstuk
64	Fields of type: text	program => bg:starttime (in: publications)	Begintijd van het programma/film/muziekstuk, zoals het daadwerkelijk is uitgezonden
65	Fields of type: text	program => bg:summary (in: summary)	Korte omschrijving van een programma/film/muziekstuk
66	Fields of type: text	program => bg:targetgroup (in: targetgroups)	Doelgroep van een enkel programma/film/muziekstuk
67	Fields of type: text	program => bg:technical-annotation (in: publications)	Technische aantekening bij de drager(s) waar programma/film/muziekstuk op staat
68	Fields of type: text	program => bg:title (in: maintitles)	Conceptuele(hoofd)titel van een enkel programma/film/muziekstuk

Media History Digital Library

Online Access to the Histories of Cinema, Broadcasting & Sound

Welcome to the Media History Digital Library

We are a non-profit initiative dedicated to digitizing collections of classic media periodicals that belong in the public domain for full public access. The project is supported by owners of materials who loan them for scanning, and donors who contribute funds to cover the cost of scanning. We have currently scanned over 1.3 million pages, and that number is growing.

Our Collections feature Extensive Runs of several important trade papers and fan magazines. Click on the arrows below to learn more about these periodicals and select volumes to download and read. You'll

Recent Blog Posts

- Digital Close Up (1927-1933) and the SCMS Anne Friedberg Innovative Scholarship Award
- New Year's at the Movies
- Magazine of the Week – The New York Clipper (December 1919)
- Our In Media Res Theme Week Begins Today!
- 'Cine-Mundial' (1916-1946) goes digital!

Donate

Support our mission to digitize public domain media periodicals and provide online access to the histories of cinema, broadcasting, and recorded sound.

Donate

Collections Gateway

All Collections (1903-1995)

Hollywood Studio System Collection

- Download This Page & Item Details
- Read in Context
- Read in Context with Search Highlighted

(takes longer to load; not yet compatible with advanced search)

2. Scandinavian Film 1952

... THE EARLY DANISH CINEMA ***ASTA NIELSEN*** WHITE SLAVE TRAFFIC 1911 THE ABYSS 1910 ...

- Download This Page & Item Details
- Read in Context
- Read in Context with Search Highlighted

(takes longer to load; not yet compatible with advanced search)

on the same theme, *The Last Victim of the Slave Traffic* (1911) and *Dealer in Girls* (1912). It had been suggested that the international success of these films prompted Carl Laemmle to make his *Traffic in Souls* in 1913.

There were clearly other reasons, however, for the appeal of the Danish films. Contemporary critics wrote of their carefully prepared stories, naturalistic settings, restrained acting and skilful direction. Already it seems that the Danish films were acquiring a reputation for clear photography and much of it reflected the love of the countryside characteristic of the Scandinavian peoples. In addition, the standard of acting appeared to be considerably above the crude, over-emphatic style common in other countries.

Two players helped to win an international audience for the early Danish films, Asta Nielsen and Valdemar Psilander. Asta Nielsen's first film, *The Abyss* (1910), was written and directed by Peter Urban Gad, who later married the actress and accompanied her to Germany. It was an immediate success and audiences everywhere responded to a sensitive, expressive style of acting which contrasted sharply with the grimacing antics of her contemporaries. Even as a girl, we are told, her face was already a tragic mask, almost impassive yet strangely expressive, with great burning eyes. Through the films she made in Denmark before going to Germany her name became known to cinema audiences everywhere and the character she created—a beautiful and intelligent woman in the grip of destiny—was one of the first to be identified and recognized. Although most of her films, and the most important, were made in Germany, she helped to lay the foundation of the Danish cinema, and through her interpretations of Nordic legend and romance many thousands of cinema-goers came nearer to an understanding of Scandinavian culture. In Germany, Asta Nielsen's most notable films were a silent version of *Hamlet* and Pabst's *The Joyless Street*. She had no sympathy with the Nazi régime and returned in 1937 to Copenhagen where she still lives.

Valdemar Psilander, a handsome, manly actor, made his first film for Nordisk, *At the Prison Gates*, in 1911. His restrained, expressive acting gave him something in common with Asta Nielsen, with whom he appeared in *The Ballet Dancer* (1911) and *The Black Dream* (1912). His most popular film was *The Clown* (1916), directed by A. W. Sandberg, of which nearly four hundred copies were sold. Paradoxically when his popularity was at its height he became a victim of melancholia and committed suicide in 1917. Just before his death he had formed an independent company which was taken over by Olaf Fönss, an actor who achieved some popularity in Danish films before going to Germany.

There were several other interesting achievements in this highly-productive period of the Danish cinema. Benjamin Christensen, a director who later worked in Sweden and Hollywood, made *The Mysterious X* for Dansk Biografkompagni in 1913. In the same year Danmark Studios made *The Island of the Dead*, based on Böcklin's painting. *The Four Devils*, based on a short story by Hermann Bang, was produced by the Kinografen Studio. Nordisk, however, remained the most

active production company and in the first years of the war made more than three hundred films.

But the market for Danish films was rapidly contracting as the war made trading conditions more and more difficult. By the end of the war it consisted of only Scandinavia and Germany. In film-making as in other matters, Denmark took seriously its position as a neutral country and in several productions sought to press the cause of peace. The Nordisk company made such films as *Pax Aeterna*, *Pro Patria* and *Ned med Våbnene* (*Down with the Weapons*). Typical of their unrealistic flavour was *Himmelskibet* (*The Sky Ship*), written by Ole Olsen. This described a rocket flight to Mars by a group of young idealists and their return with a Martian emissary to plead for peace on earth. This was a well-intentioned but misguided production trend. It is curious to note that the Second World War produced similar sermons on peace from neutral Switzerland.

The end of the war found the Danish film struggling against much more powerfully equipped competitors. Hollywood had achieved a world market for its films during the war, the Swedish cinema had out-paced its smaller Scandinavian neighbour in development, and the German cinema was about to enter on its golden period. In an attempt to re-establish its popularity in the English-speaking world, the Danish companies produced film versions of novels by Dickens and Captain Marryat. A. W. Sandberg made *Our Mutual Friend* (1919), *Great Expectations* (1921), *David Copperfield* (1922), and *Little Dorrit* (1924). These films had some success in Scandinavia where their wistful sentimentality had an appeal; but for audiences in Britain and America they failed to capture the essential flavour of Dickens's work. The success of David Lean's films twenty-five years later suggests that an author's ideas can best be interpreted in his own country.

It was through much less pretentious material that the Danish cinema regained part at least of its world market. The comedies which Lau Lauritzen made for the new Palladium company, with Carl Schenström and Harald Madsen, could not be equated with the clowning of Chaplin; but they were lively, friendly affairs which had a warm appeal for audiences all over the world in the troubled 'twenties. The comedians, tall, thin and serious Schenström and small, plump and merry Madsen, were given a number of names: 'Fyrtårnet og Bivognen' ('Light-house and Tram-Trailer') in Denmark, 'Long and Short' in Britain, 'Double-patte and Patachon' in France. When most of the more ambitious Danish films had disappeared from the world's cinemas, these comedies were still shown regularly to audiences to whom the comedians became as familiar as Laurel and Hardy and Abbott and Costello. In ten years they made about forty films, an output equalled by few other teams of comedians. Their director, Lau Lauritzen, who knew how to get the last broad laugh out of a farcical situation, died in 1938.

One figure links the Danish cinema of yesterday and to-day. Carl Theodor Dreyer began writing scripts for Nordisk about 1912 when he was a young journalist. His first film, *The President*, adapted from a novel by Karl Emil

Valdemar Psilander, a handsome, manly actor, made his first film for Nordisk. At the Prison Gates, in 1911. His restrained, expressive acting gave him something in common with Asta Nielsen, with whom he appeared in The Ballet Dancer (1911) and The Black Dream (1912). His most popular film was The Clown (1916), directed by A. W. Sandberg, of which nearly four hundred copies were sold. Paradoxically when his popularity was at its height he became a victim of melancholia and committed suicide in 1917. Just before his death he had formed an independent company which was taken over by Olaf Fonss, an actor who achieved some popularity in Danish films before going to Germany. *Page 2*

Table I
RADIO STATION RESULTS.

Rank	Station ID	Market
1	WGN	Chicago
2	WJZ	New York
3	KDKA	Pittsburgh
4	WMCA	New York
5	KYW	Philadelphia
6	WLS	Chicago
7	WBBM	Chicago
8	WBZ-WBZA	Boston & Springfield
9	WCAU	Philadelphia
10	WHN	New York
11	KHJ	Los Angeles
12	WSB	Atlanta
13	KNX	Los Angeles
14	KFI	Los Angeles
15	WGY	Schenectady
16	WWJ	Detroit
17	WCCO	Minneapolis
18	WIP	Philadelphia
19	KGO	San Franscisco
20	WFAA	Dallas

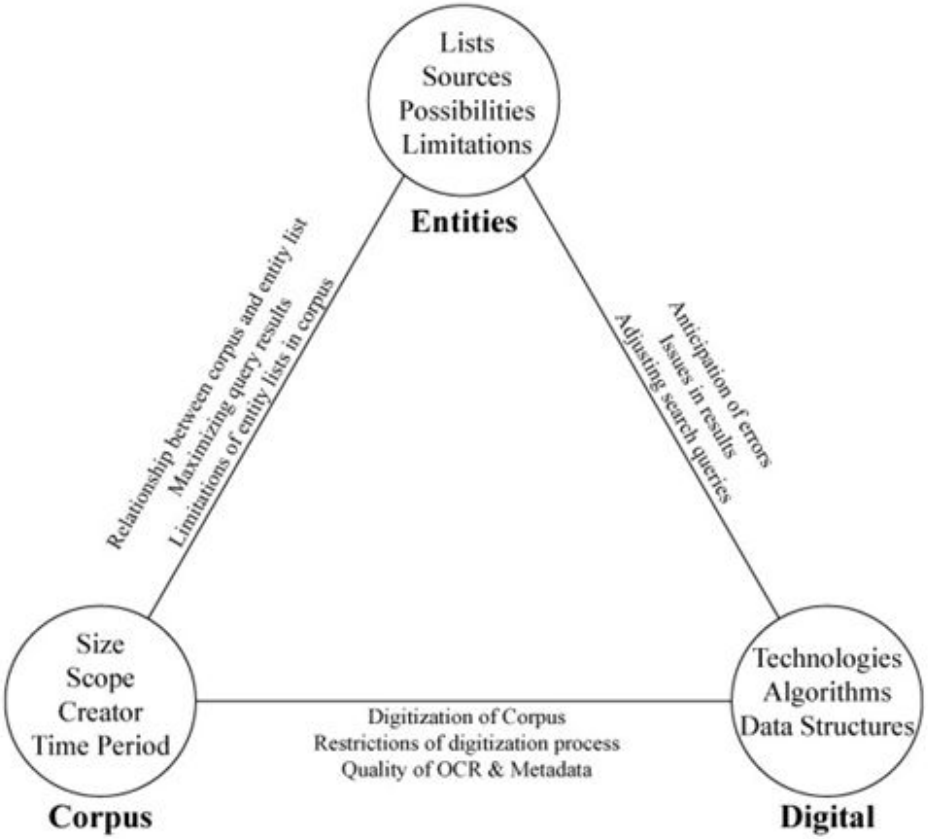
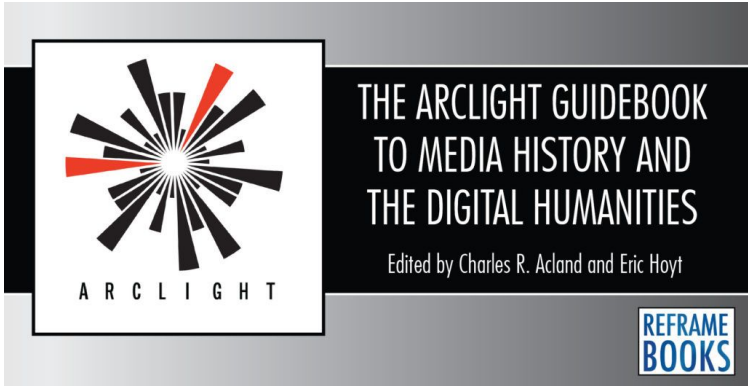


Figure 2. The SES triangle method of interpretation.

Hoyt, Eric, Kit Hughes, Derek Long, Anthony Tran, and Kevin Ponto. 2014. "Scaled Entity Search: A Method for Media Historiography and Response to Critiques of Big Humanities Data Research." In *2014 IEEE International Conference on Big Data (Big Data)*, 51–59.



ENTITIES

Selection of entity lists:
How and why did you select this grouping to compare? If you did not generate the entity list yourself, where did it come from? What sources were used to generate the data? How does this list open up new possibilities for research? How does it limit or close down other possibilities?

CORPUS

The corpus being queried:
What is the size and scope of the corpus? Who created it and why? What are its strengths and weaknesses in terms of the time periods covered and diversity of publications?

DIGITAL

Technologies, algorithms, and data structures that comprise the process:
What schema, fields and facets were used in creating the search index? What historical materials, processes, and experiences do not easily lend themselves to digitization and what effect does their omission have on results? How does making materials machine-readable change the research process?

ENTITIES + CORPUS

What is the relationship between the list of entities you are querying and the corpus? How could you design an entity list that plays to the strengths of the corpus? At the same time, if we only design research questions and entity lists on the basis of what is likely to generate interesting results in the corpus, how does this limit scholarship?

CORPUS + DIGITAL

How did the digitization process change the nature of the corpus? What is the quality of the OCR text? How did intellectual property restrictions and other factors influence what material was digitized and what was left out? How granular is the metadata that describes the corpus and is it consistent? Is the underlying corpus data openly accessible, viewable, and reusable?

ENTITIES + DIGITAL

The Entities-Digital Relationship: What issues of disambiguation, false positives, and false negatives can you anticipate before querying the entities? What issues do you recognize in examining the queried results? How do you adjust the search queries to try to mitigate these problems? Do you make these adjustments consistently or selectively?

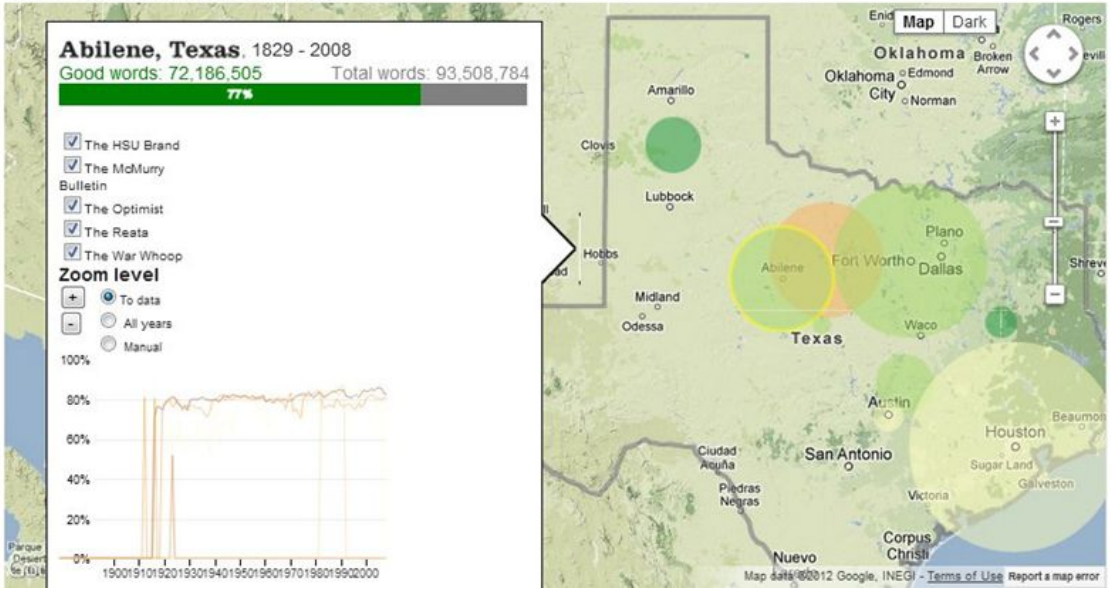
Assessing Digitization Quality

Scans of Texas Newspapers, 1829-2008

This visualization plots the quantity and quality of 232,567 pages of historical Texas newspapers, as they spread out over time and space. The graphs plot the overall quantity of information available by year and the quality of the corpus (by comparing the number of words we can recognize to the total number scanned). The map shows the geography of the collection, grouping all newspapers by their publication city, and can show both the quantity and quality of the newspapers from various locations. Clicking on a particular city will provide a detailed view of the individual newspapers, where you can examine both the quantity and quality of information. A timeline of historical events related to Texas is also available for context.

Time Quantity of Recognized and Unrecognized Text, 1829-2008

Space Collection Quantity and Quality by Location



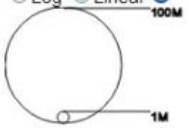
Legend

Ratio of Good to Bad Words

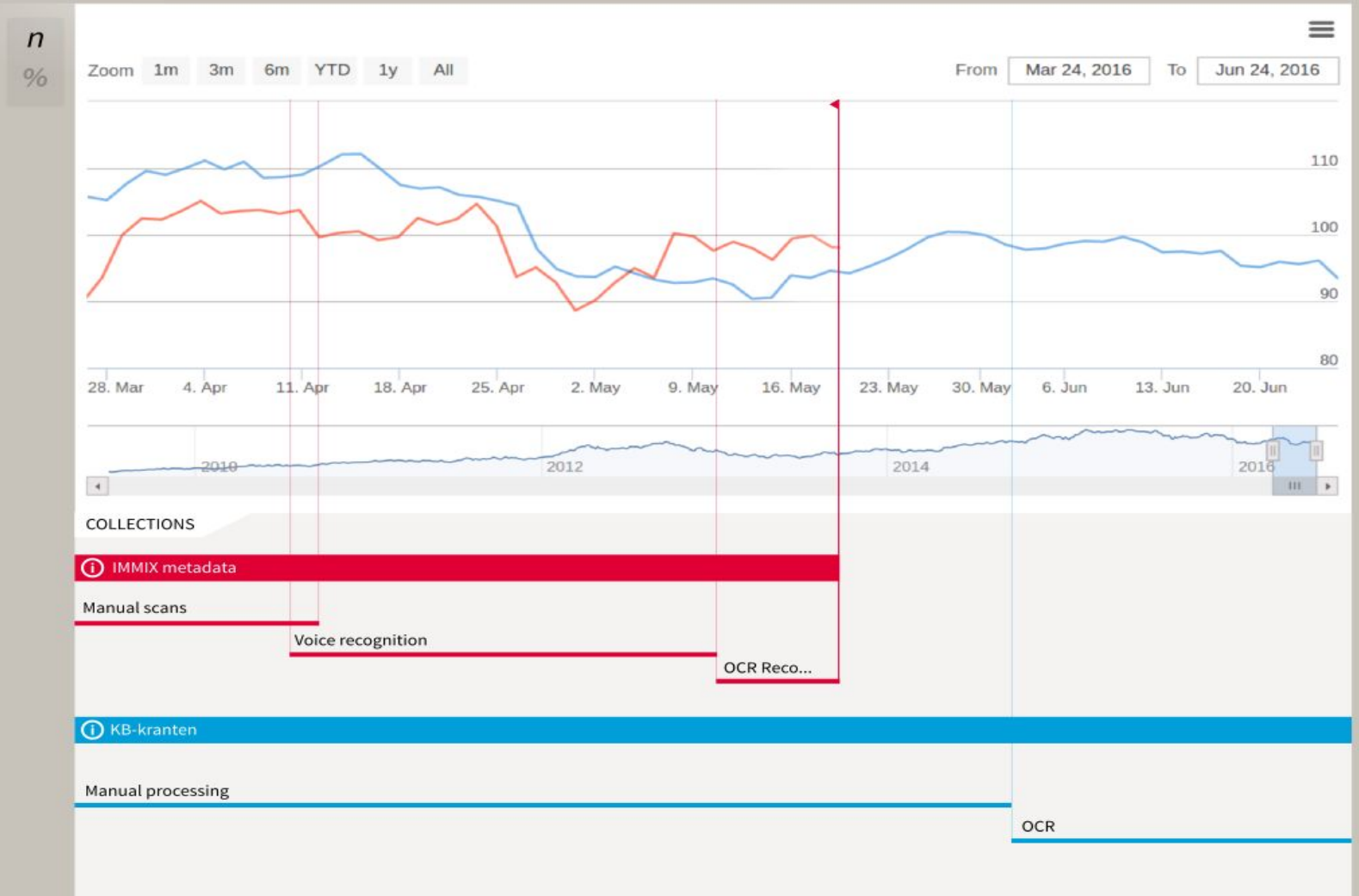
*Showing cities having publications with **50.0% - 100%** correct words.*

Circle Scaling:

Log Linear

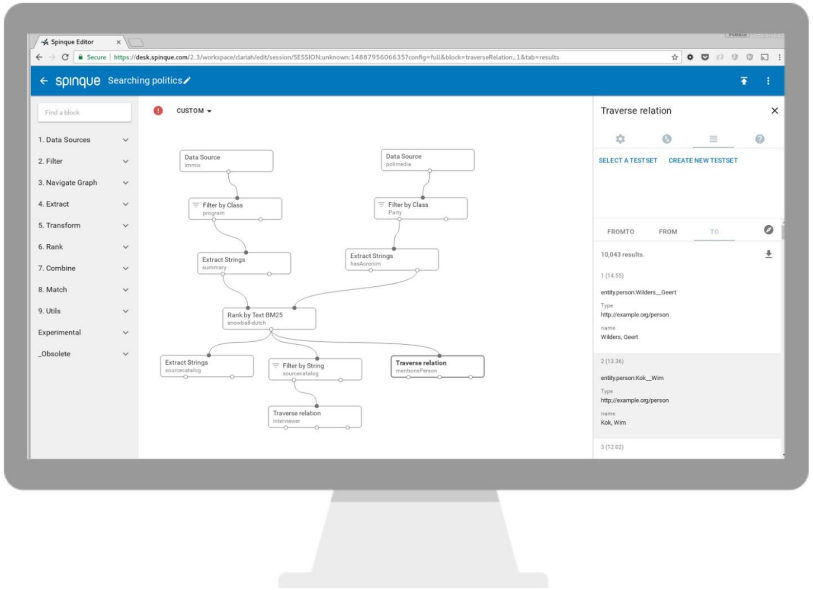


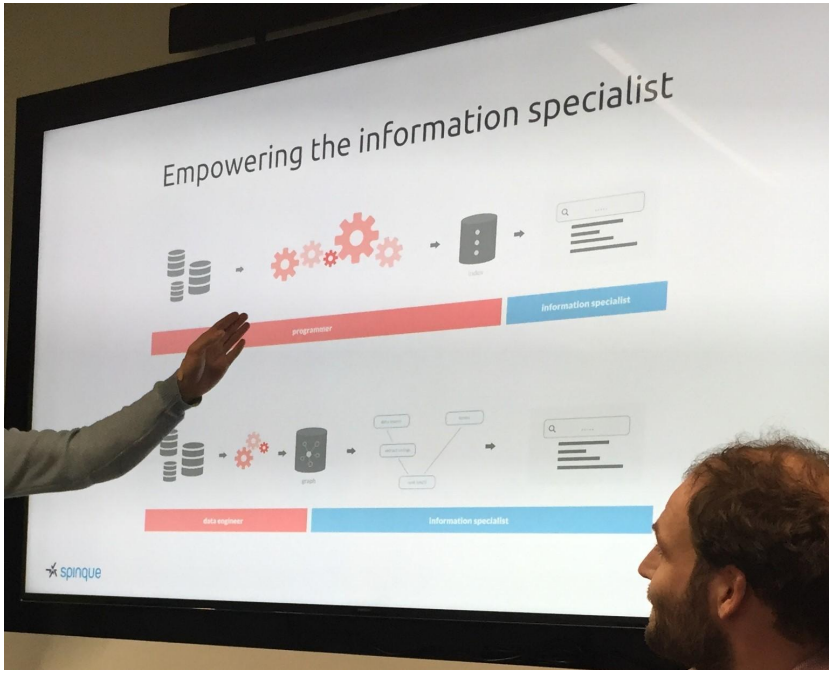
Circle size is relative to total number of words scanned. Color indicates overall scan quality.

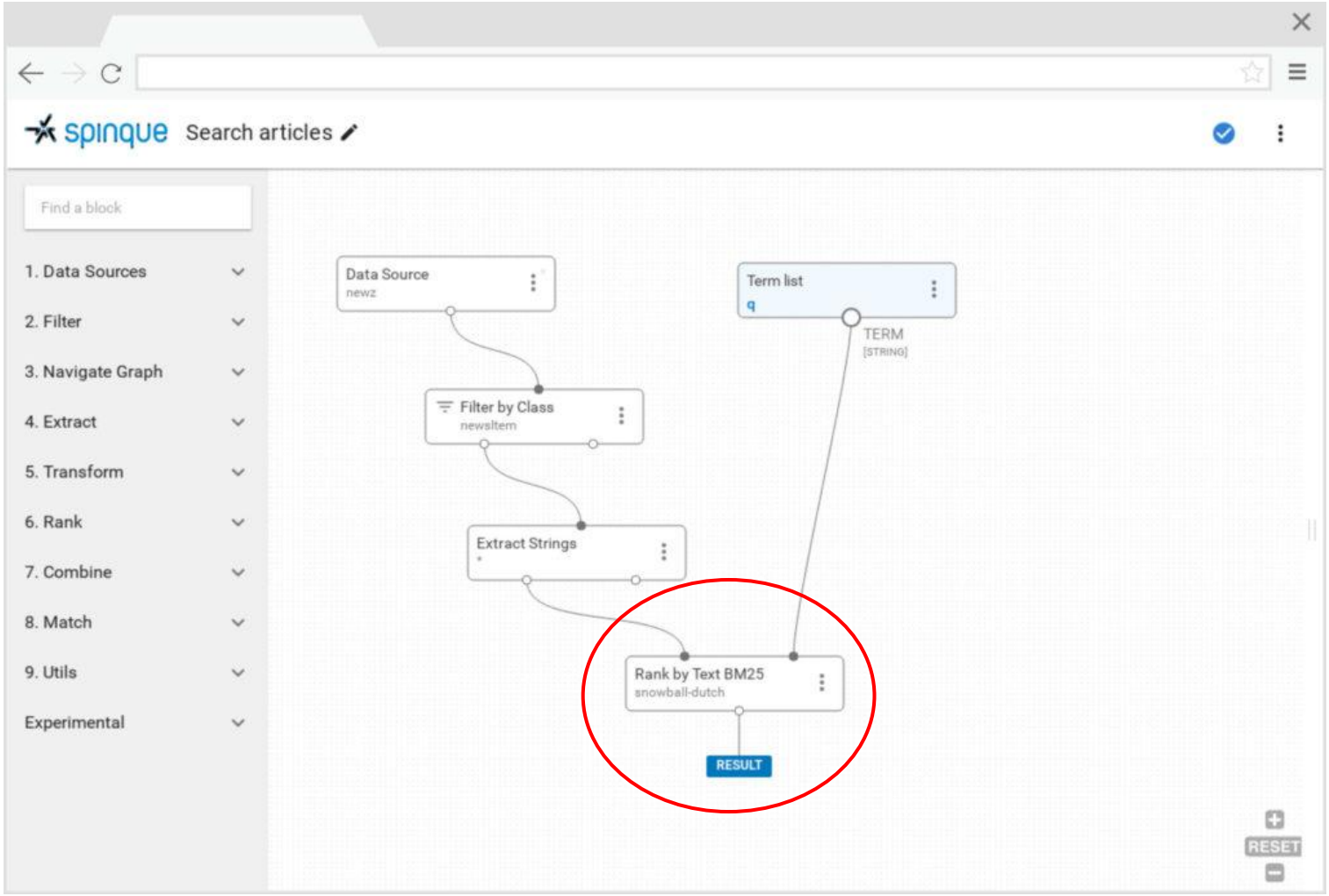

n
%
Zoom 1m 3m 6m YTD 1y All
From Mar 24, 2016 To Jun 24, 2016
110
100
90
80
28. Mar 4. Apr 11. Apr 18. Apr 25. Apr 2. May 9. May 16. May 23. May 30. May 6. Jun 13. Jun 20. Jun
2010 2012 2014 2016
COLLECTIONS
IMMIX metadata
Manual scans
Voice recognition
OCR Reco...
KB-kranten
Manual processing
OCR

Make your own search engine

Spinque Desk






spinque
Search articles
Find a block
1. Data Sources
2. Filter
3. Navigate Graph
4. Extract
5. Transform
6. Rank
7. Combine
8. Match
9. Utils
Experimental
Data Source
newz
Term list
q
TERM
[STRING]
Filter by Class
newsItem
Extract Strings
Rank by Text BM25
snowball-dutch
RESULT
RESET

Conclusions

- More data = more transparency needed
- Black-boxing versus improving digital literacy
- Solutions:
  - aggregate views on collections and metadata quality
  - tools for data criticism
  - tools for tool criticism

Launch MSv2: 20 December 2017

http://mediasuite.clariah.nl/

Julia Noordegraaf

j.j.noordegraaf@uva.nl

http://www.clariah.nl/werkpakketten/focusgebieden/media-studies