

Data Scopes

Towards transparent data research in digital history

Marijn Koolen

Tutorial - Research School Political History
Huygens ING, Amsterdam, 12/02/2021

Slides: <http://bit.ly/OPG-2021-Digital-History>

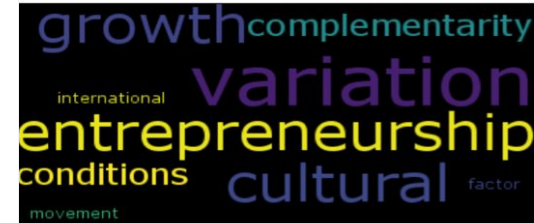
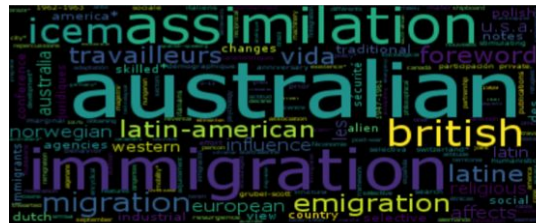
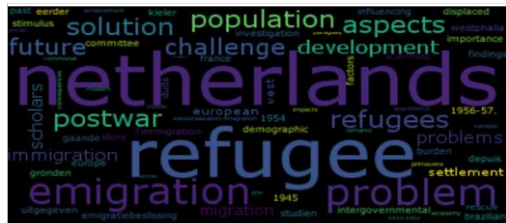
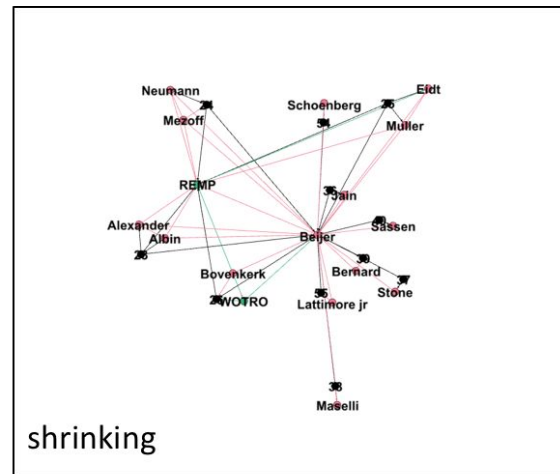
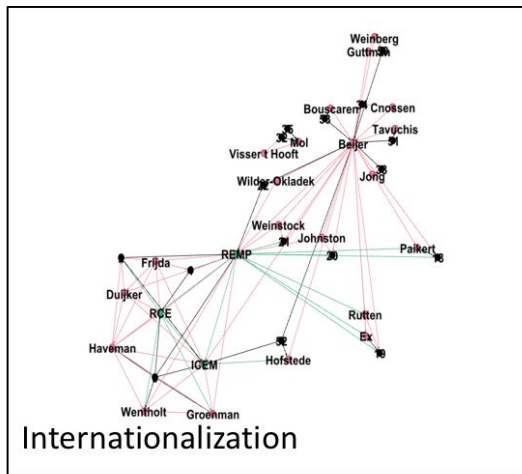
Data Scopes

- Conceptual toolbox for doing and communicating about digital research
- Data scope: you want to analyse a certain aspect of your materials,
 - but the “raw” data is not suitable for direct analysis.
- You have to do something with the data. Questions:
 - What do I have to do to make data suitable?
 - How do I do that?
 - What should I document of this process to share it with others?

Example: discourse coalition migration

- **Research question:**
 - What is the development of the discourse about the management of migrants?
- **Context:**
 - research project about Dutch emigration 1945-1992
- **Sub questions:**
 - Who were involved in the international discourse about the management of migrants?
 - How did the discourse change over time?
 - How can we relate these changes?
 - To what extent do politics and science interact with and influence each other?
- **From documents to datasets:**
 - Journals and bulletins on international migration
 - Data about people and their connections (authors, editors, funders)
 - Data about the topics of discourse (titles of journal articles)

Discourse Themes



1950s

1960s

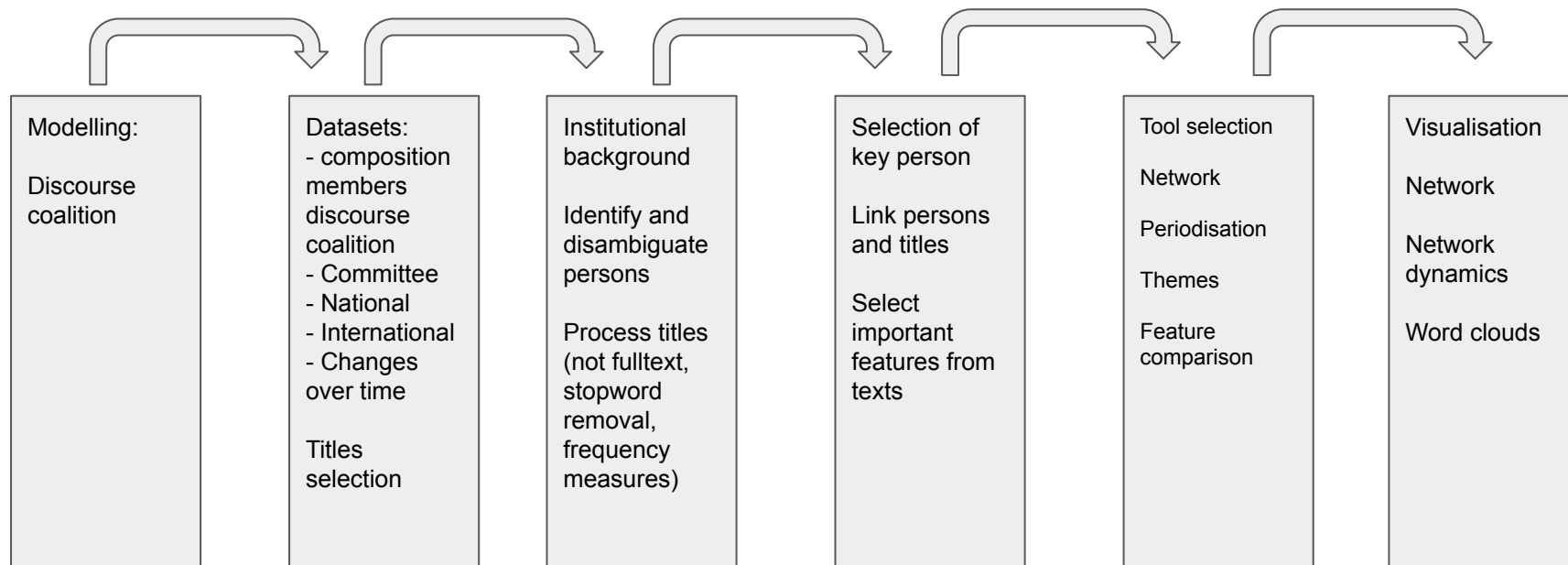
1970s

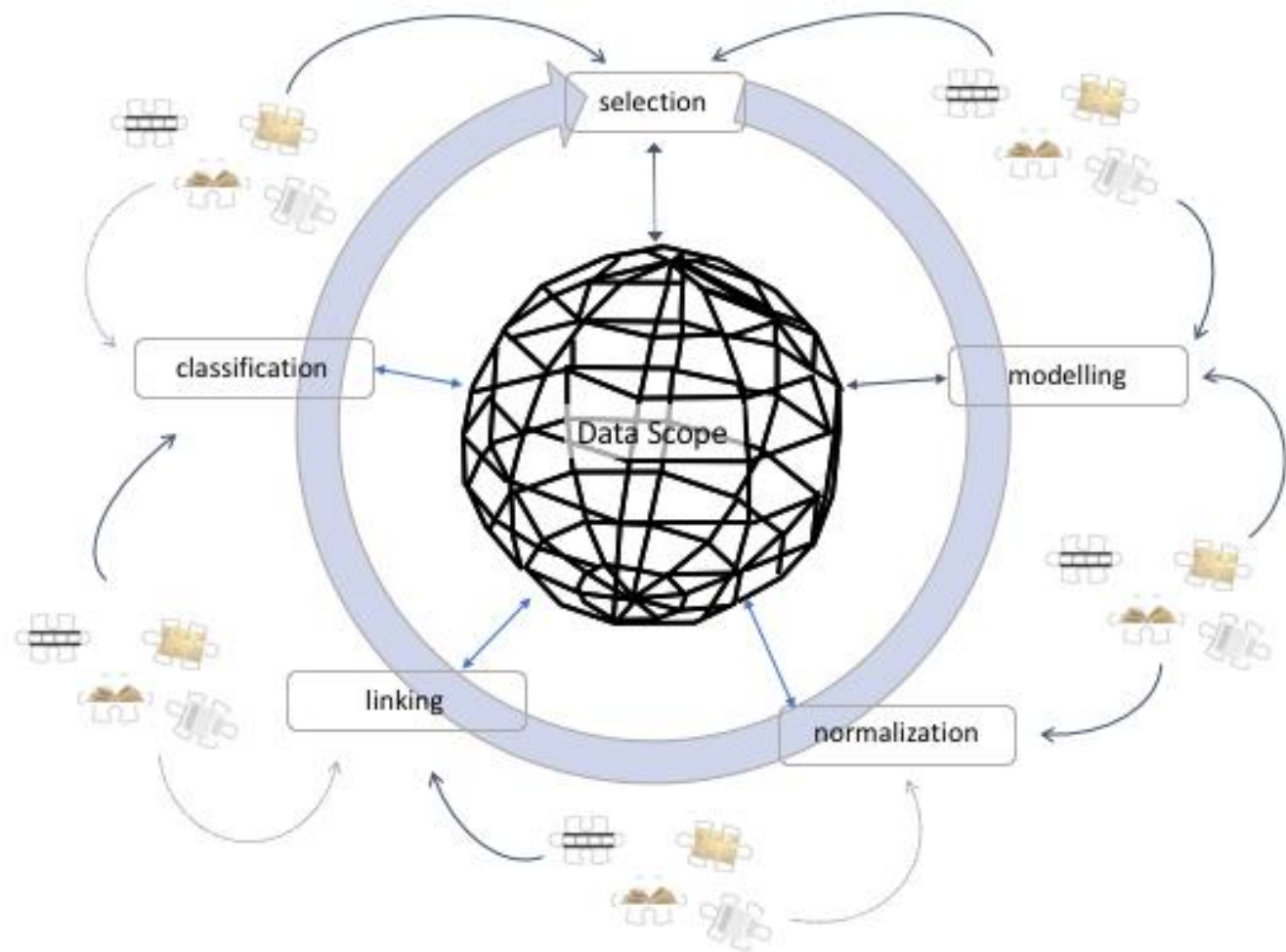
Motivation

- Making data work for research requires:
 - Technical know-how of how digital tools handle data
 - Intimate knowledge of the domain and subject of source materials
- But also:
 - Reflection on how choices are informed by prior knowledge and experience
 - Reflection on how choices put emphasis on some aspects, while pushing back others
 - Reflection of the transformation of data in the research process
- Often requires collaboration...
 - How to organise that
- ... and lots of discussion
 - Choices that one collaborator makes should be visible to the rest

Progressive steps of data transformation

Data processing: steps (incomplete are far more linear than reality)



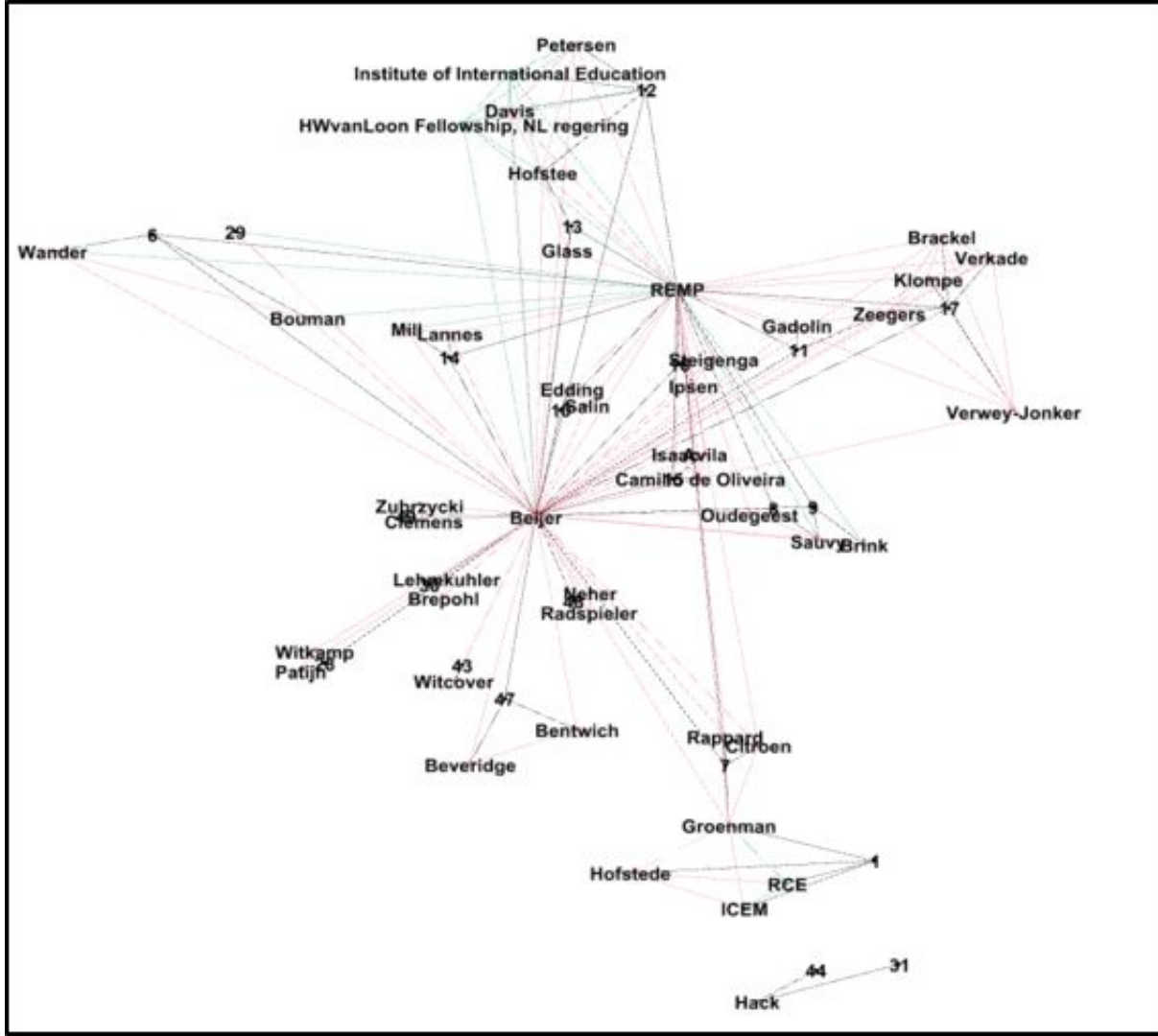


Creating a Data Scope

- Research plan:
 - Analyse network of experts involved in discourse on migration
- Research process:
 - Translate plan into sequence of data selections and transformations
 - Dataset 1: publications of Research group on European Migration Problems (REMP)
 - Transform actors involved in into a social network
 - Dataset 2: articles in journal International Migration (IM)
 - Transform article titles into word clouds
 - Cycle of interpretations, decisions and actions
- Research description (van Faassen & Hoekstra 2017):
 - *“To find out exactly how these experts were connected to key actors from the political sphere, [...], we went through the prefaces of the publications. We modelled the different roles of the key actors based on issues such as: who were writing forewords, prefaces or introductions to each other’s work; Who ordered the research? Who financed it? Etc.”*

Selecting

- Which materials do I include? Which do I exclude and why?
 - How important are completeness and representativeness?
 - Potentially **huge** impact on network analysis
- Algorithmic selection:
 - Everything matching a (set of) keyword(s)
 - Documents by type, creator, title, size, ...
 - How does technology allow and limit selection?
- What are consequences of these selections?

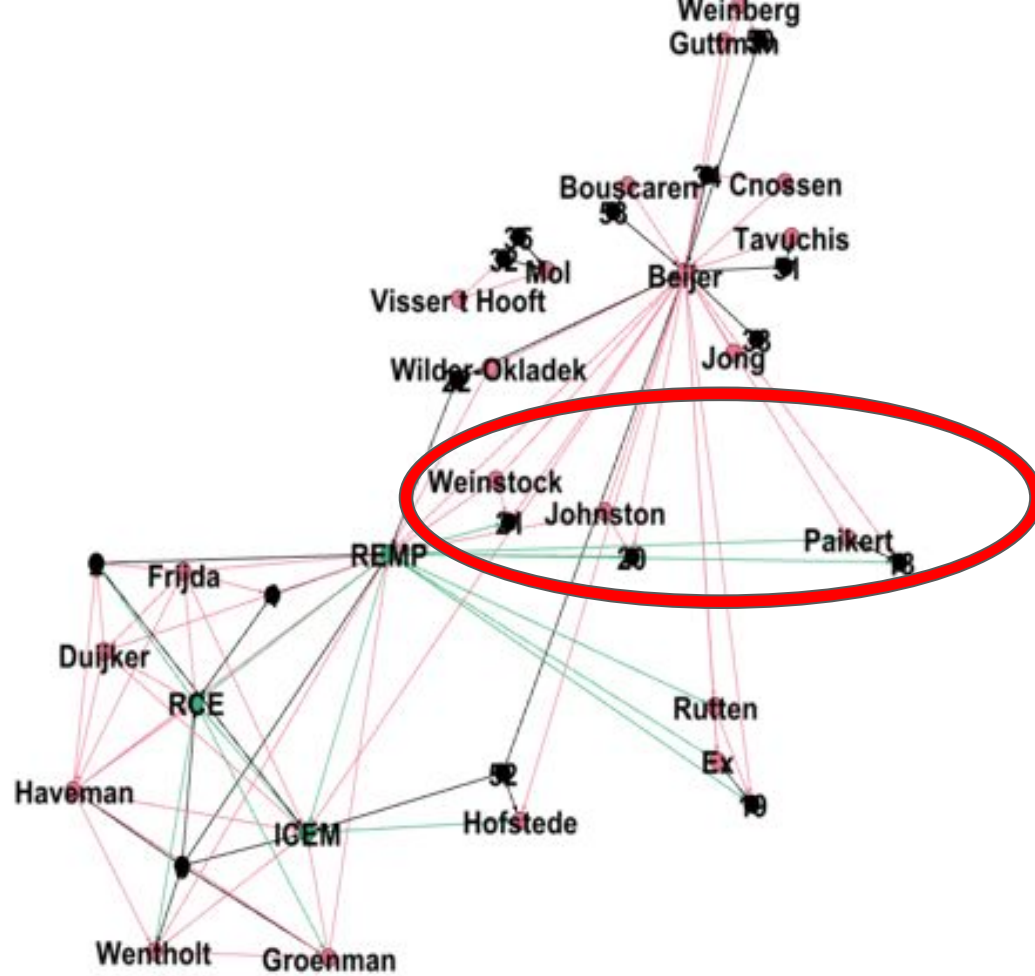


Publications
1950-1959

Modelling

- Computational approach requires modelling data (McCarty 2004)
- What aspects/elements of data to focus on and what to leave out (why?)
 - Coalition actors:
 - i. People and organizations involved in publications on international migration
 - ii. Authors, editors, commissioners, sponsors
 - iii. Change in coalitions from 1950s in 10 year periods
 - Coalition topics:
 - i. Content words in titles of articles published in journals on International Migration
- Structures data in sources around research focus
 - Transforms data, affects interpretation!

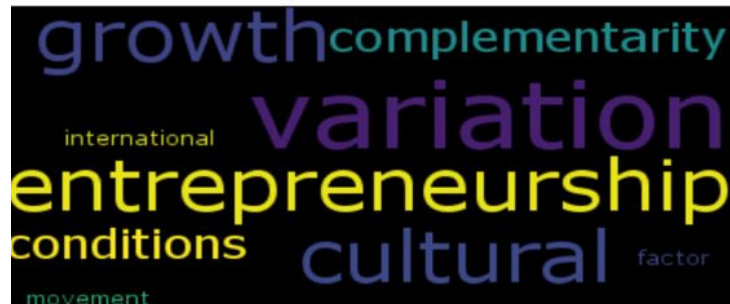
	A	B	C	D	E	F	G	H	J	K
1	Organisation	Period_start	Last_known_d	Prs_id	Prs_surname	Prs_inf	Prs_initials	Prs_function	Is_academic	Is_public_administ
2	REMP	1952	1983	1	Beijer		G.	demographer, The Hague	yes	
3	REMP	1952	1969	2	Groenman		Sj.	sociologist, Leiden	1947	1943-1950
4	REMP	1952	1969	3	Zeegers		G.H.L.	economist, sociologist, Nijmegen	yes	1941-1950
5	REMP	1952	1969	4	Hofstee		E.W.	sociologist, Wageningen	yes	yes, advisor 5 minis
6	REMP	1952	1969	5	Bouman		P.J.	sociologist, Groningen	yes	
7	REMP	1952	1969	6	Oldendorff		A.	sociologist Tilburg		1946-1947
8	REMP	1952	1954	7	Gelissen		H.	chemist		1935-1937
9	REMP	1952		8	Schokking		J.J.	lawyer + sociologist(?), Cologne	yes	
10	REMP	1952	1969	9	Sauvy		A.	demographer, UN_comission for	1945	1939
11	REMP	1952	1969	10	Gottmann		J.	geographer, Paris, Princeton		
12	REMP	1952	1969	11	Lacroix		M.	New York		
13	REMP	1952	1957	12	Jacobson		M.P.	lawyer, linguist, Geneva	yes	1952-1957
14	REMP	1952	1969	13	Winkler		W.	Statistician, Vienna		
15	REMP	1952		14	Janne		H.	classicist, sociologist, Brussels, Brugge		
16	REMP	1952	1969	15	Mertens de Wilmard		J.	lawyer, political scienctist Louvain		
17	REMP	1952	1969	16	Baade		F.	Economist, Kiel		
18	REMP	1952	1954	17	Mackenroth		G.	sociologist, statistician, demographer, Kiel		
19	REMP	1952	1969	18	Ritschl		H.	economist, Hamburg		
20	REMP	1952	1969	19	Hoffmann		W.	sociologist, economist, (?) Munster i.W.		
21	REMP	1952	1969	20	Neundorffer		L.	Frankfurt a.M.		
22	REMP	1952		21	Gadolin	de	A.	political economist, Helsingfors		
23	REMP	1952	1954	22	Vito		F.	economist, Milano		
24	REMP	1952	1954	23	Livi		L.	statistician, Florence		
25	REMP	1952	1969	24	Parenti		G.	Roma		
26	REMP	1952	1969	25	Vergottini		M.	statistician, mathematician, demographer, (?), Catania		
27	REMP	1952		26	Vampa		D	Paris		
28	REMP	1952	1969	27	Hyrenius		H.	Gothenburg		
29	REMP	1952	1969	28	Salin		E.W.	Basel		



Publications
1960-1969

Normalizing

- Bring surface forms expressed in data to underlying standard form
 - Is [G. Beijer](#) in article X the same as [G.O.K. Beyer](#) in article Y?
- Map variation onto a single representation:
 - Linguistic, geographical, spatial, temporal, structural
 - E.g. [entrepreneur](#), [entrepreneurs](#), [entrepreneurship](#)
 - Important consequences whenever you count frequencies or analyse networks
- What is irrelevant variation?
 - Is the distinction between [entrepreneur](#), and [entrepreneurship](#) important for research focus?
 - Uncertainty: Do [New York](#) and [NYC](#) refer to the same thing?
- Essential for next step: linking



Country Names and Nationalities

Country	# of titles
Australia	80
United States	56
Canada	46
Israel	40
Germany	19
Sweden	18
India	17
Latin America	15
Mexico	14
Japan	14

Country Names and Nationalities

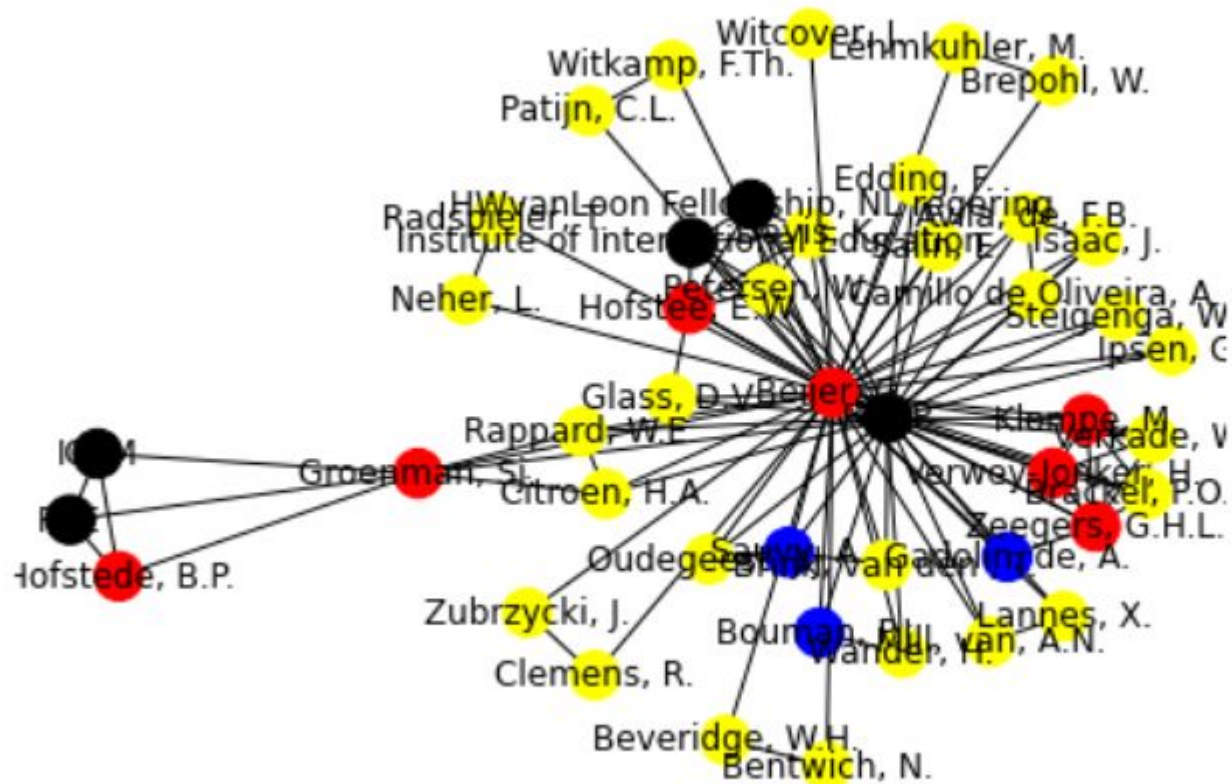
Country	# of titles	Country + Nationality	# of titles
Australia	80	Australia	80
United States	56	United States	75
Canada	46	Canada	65
Israel	40	Israel	40
Germany	19	China	29
Sweden	18	Mexico	25
India	17	Germany	23
Latin America	15	Greece	23
Mexico	14	Turkey	21
Japan	14	Macau	21

Linking

- Establishing explicit connections between objects in data sources
 - Within a dataset: relations between people, organizations
 - i.
 - Across datasets: e.g. mentions of same person, location, date, ...
 - i. Can bring together disparate data about single entity from different sources
- What counts as a link?
 - Editor - Main author
 - Preface author - Main author
 - Commissioner - Main author
 - Commissioner - Sponsor

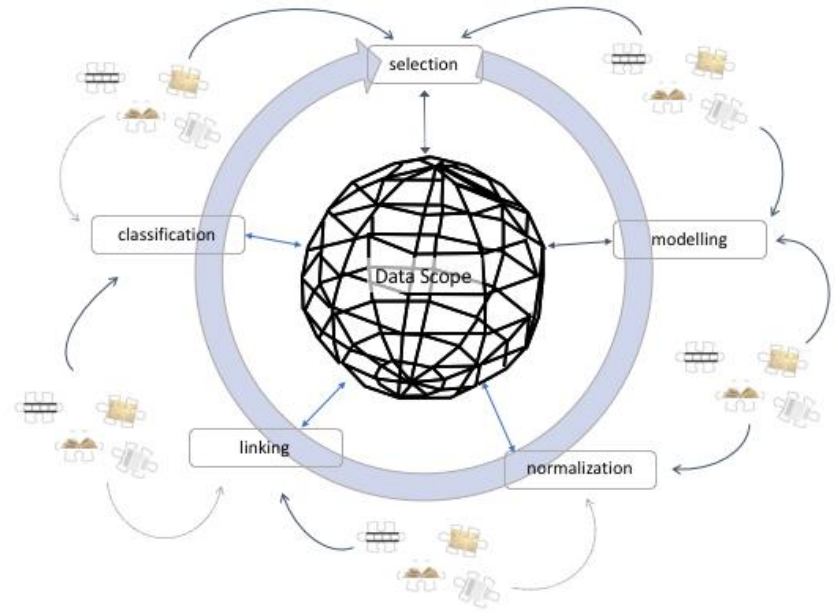
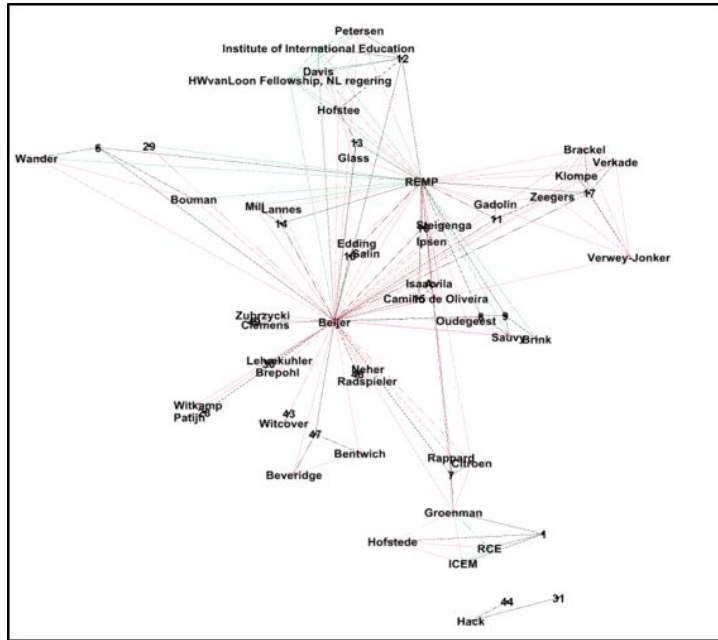
Classifying

- Reduction of complexity by grouping (data) objects into predefined categories, or classes
 - Bringing together objects with similar properties
 - Separating objects with dissimilar properties
 - E.g. categorise coalition actors as 'academic' vs. 'public administration'
- Adds new layers of structure and interpretation to data
 - Especially useful for low-frequency items
 - Many data dimensions have “long tails” which are hard to structure
- Deciding on classification dimensions and classes is part of modelling



Actors in REMP publications

Organisation, Public administrator, academic (confirmed), academic (assumed)



Understanding data scope affects interpretation of network visualization!

Decontextualisation - Recontextualisation

Delpher

FavorietenBekeken objectenInstellingenHandleiding

Kranten

migratie

Zoeken

Uitgebreid zoeken

☐ Illustratie met onderschrift (34)

Verspreidingsgebied

☐ Landelijk (9846)

☐ Nederlands-Indië / Indonesië (294)

☐ Nederlandse Antillen (954)

☐ Regionaal/lokaal (4777)

☐ Suriname (349)

☐ Verenigde Staten (2)

Krantentitel

Kies krantentitel...


Plaats van uitgave

Kies plaats van uitgave...


Herkomst

Kies herkomst...


Toevoegingen in Delpher



Extra woningen voor W.-Brabant
migratie Putte 12 (migratie), Etten-Leur 9 (industrie) Ossenheim 8 (migratie), Woensdrecht 5 (migratie) en 3 (Industrie), Oudenbosch 7 [industrie], Steenberg 7 (industrie), Zevenbergen (S (ind
Krantentitel Nieuwsblad van het Zuiden
Datum 11-11-1955
[Meer details](#)



Migratie-verlies in Zuid-Afrika
Migratie-verlies in Zuid-Afrika JQHANNESBURG (I Pil - steeds meer blanken verlaten Zuid-Afrika en het land toont nu regelmatig een maandelijks „migratie-tekort”. Politici van de oppositie gaven
Krantentitel Nederlands dagblad : gereformeerd gezinsblad / hoofdred. P. Jongeling ... [et al.]
Datum 08-10-1977
[Meer details](#)



Grote migratie van Antillianen te verwachten
e in 1967: 714 personen bedroeg in 1968: 636 personen (het staartje van het ontbreken van een echte migratie-traditie) in 1969: 1132 personen (volgens drs Van Amersfoort is dit cijfer, om het met een
Krantentitel Amigoe di Curacao : weekblad voor de Curacaosche eilanden
Datum 28-01-1972
[Meer details](#)

Search Engine as Mediator

- For many online materials access is limited to search interface
 - Browsing is guided by available structure
 - Drill down via facets
 - Navigate via metadata fields (if enabled)
 - Without (relevant) structure, direct search is only practical alternative
- Searching as exploration
 - How does search engine provide overview?
 - How big is collection?
 - How is collection structure communicated?
 - What (meta)data is available?
 - How are search characteristics explained?
 - How are search results summarised?

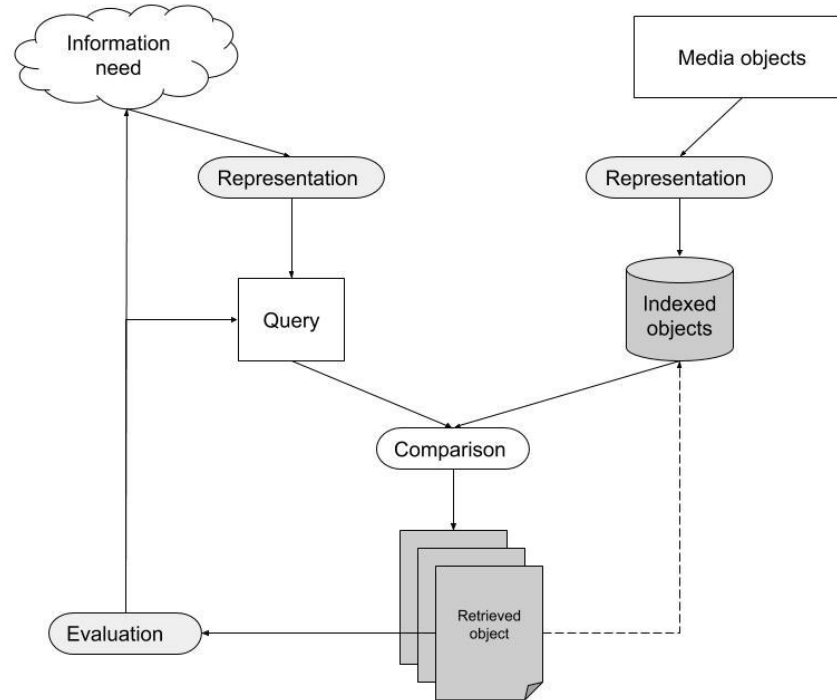
Digital Source and Data Criticism

- Power of the archive
 - Problem of perspective (from archive-as-source to archive-as-subject, Stoler 2002)
- History of the archive
 - Collections created over decades often go through changes in
 - selection criteria, cataloguers (human or algorithm),
 - cataloguing budgets, policies, rules, practice and vocabularies,
 - software (migrations and updates), hardware,
 - institutional mission, societal attitudes, ...
 - Most of these aspects remain undocumented or partially documented
- Consequences
 - Almost inherently **incomplete**, **inconsistent** and sometimes necessarily **incorrect**
 - After many years, it's hard to retrace what happened
 - and how it affects access, selection and analysis

Search and Accountability

- What should scholars account for?
 - Aspects of sources, tools and process
- Digital source criticism
 - How to evaluate digital sources (Fickers 2012)
 - Who made digital source, when, why, what for, how?
- Digital tool criticism
 - How to evaluate impact of digital tools (Koolen et al. 2018)
 - Reflection-in-action, experimentation
- Data Scopes
 - How to communicate research process to others (Hoekstra & Koolen 2018)
 - Discuss process of *selection, modelling, normalization, linking, classification*

Anatomy of Retrieval Process



Retrieval - Matching and Similarity

- Matching based on user query
 - Query: free text, controlled facet, example (doc, AV or text)
 - Matching docs returned in certain order (non-matching are not retrieved)
 - How does search engine perform matching (esp. for free text and example)?
 - Potentially many objects match query: does order matter?
- Similarity
 - Degree of matching: some match better than others (notion of similarity)
 - Retrieve most similar documents first (ranking)
 - Similar how? Does interface explain?
- Retrieval and ranking
 - Retrieval: which matching documents are returned to the user as results?
 - Ranking: in which order are the results returned?

Opacity of Interfaces and Experimentation

- Experiment to understand search functionalities
 - How can you find out if multiple search terms are treated with Boolean AND or OR operators?
 - How can you find out if terms are stemmed/normalized?
- Phrase search:
 - What happens when you use quotation marks to group terms into a phrase?
 - How do the results compare to those using no quotation marks?
- Proximity search:
 - Can you specify that terms should be near each other?
- Fuzzy search: wildcard and edit distance searches
 - Controlling lexical variation vs. uncontrolled wildcard search
 - voetbal+voetballen vs. voetbal* (matches voetbalvereniging, voetbalveld, ...)

Exercise

- Experiment with the search engine you're using
 - Find out if stopwords are removed
 - Find out if words are stemmed/normalized
 - Find out how multi-word queries are interpreted, i.e. as AND or OR
 - Find out how standard search operators work
 - Boolean AND, OR and NOT
 - Quotation marks for phrases

Search in Research

- More elaborate slide presentation:
 - Search in Research - Let's make it more complex
 - <https://www.slideshare.net/MarijnKoolen/search-in-research-lets-make-it-more-complex>
 - Some focus on audiovisual archives, but most points are more generic