

Experiments on the intersection of texts and structured data

Combining language technology and semantic web for
digital humanities research

Marieke van Erp
KNAW Humanities Cluster
Digital Humanities Research Lab

 merpeltje.bsky.social



Overview

- Semantic Web 101
- Foodstuffs
- Nasty words
- Smelly things
- Outlook

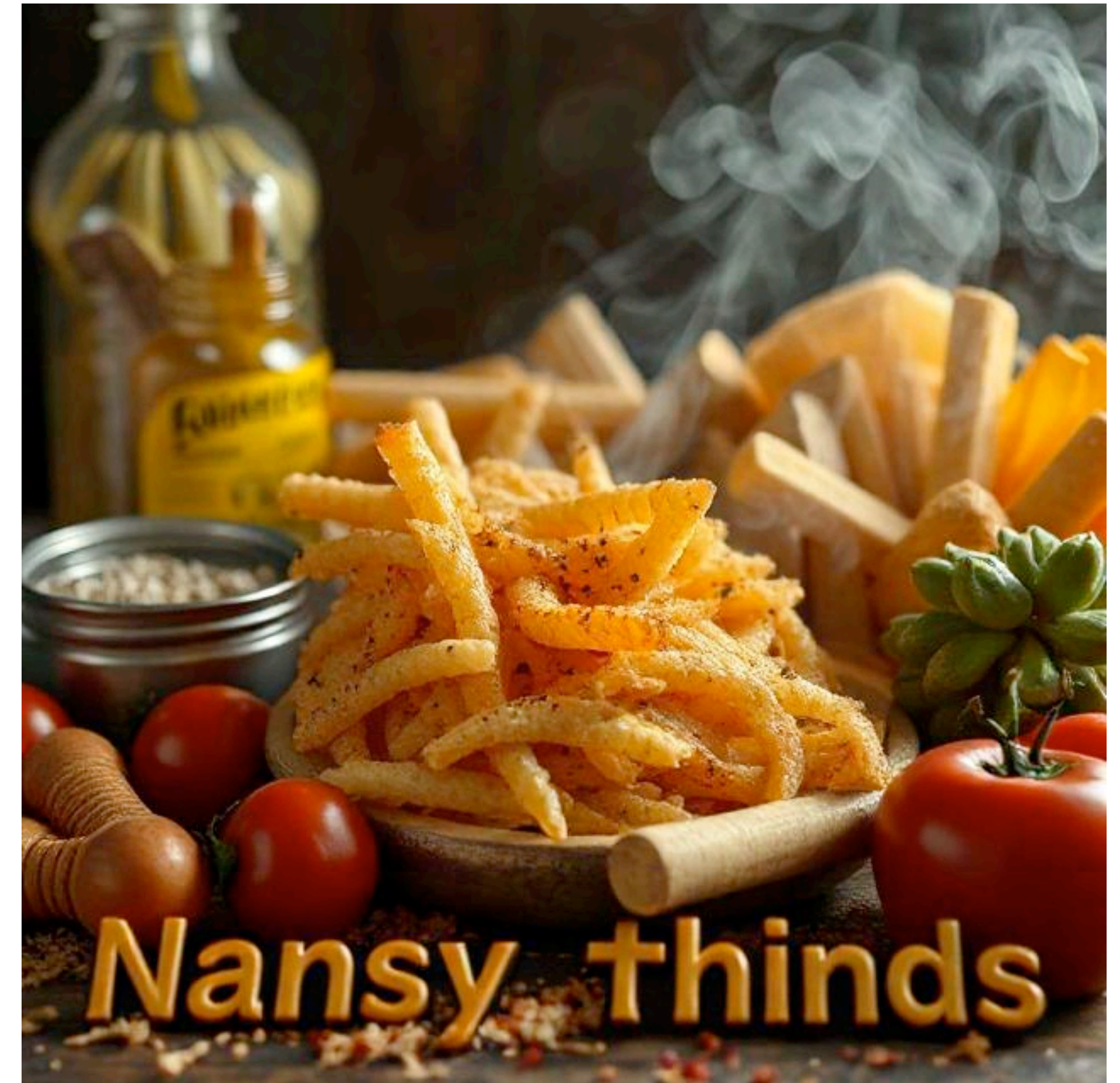
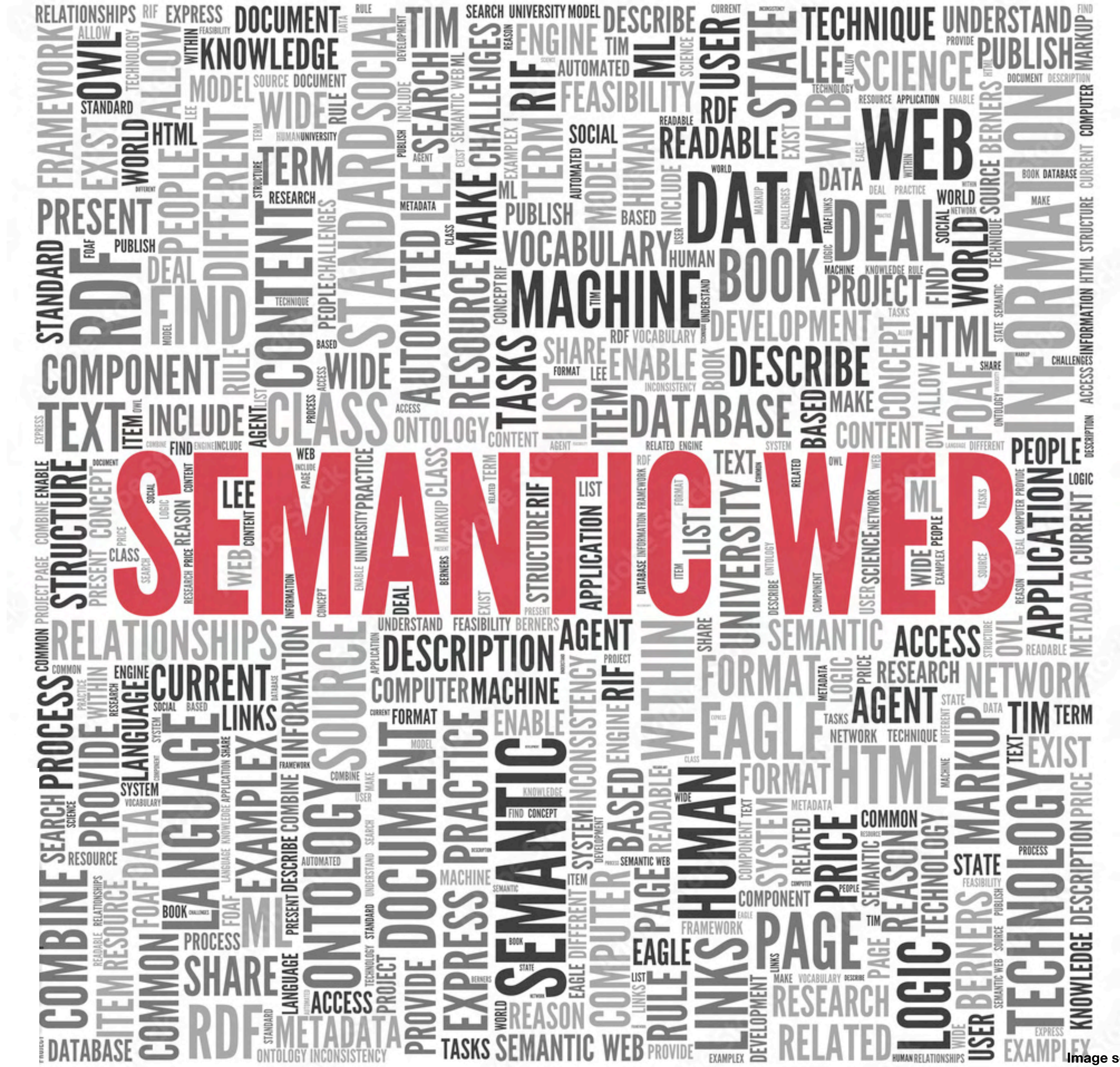


Image generated by deepai.org prompt: an image that contains foodstuffs, smelly things and nasty words

Semantic Web 101



Semantic Web 101

- Describe & link Web content in a machine-readable manner
- element website 1 — website 2 (e.g. 'click here')
- element resource - has relation - element resource 2

```
view-source:https://www.allrecipes.com/recipe/12682/apple-pie-by-grandma-ople/  
630 <script id="allrecipes-schema_1-0" class="comp allrecipes-schema micro-schema-undefined" type="application/ld+json">  
631 {  
632   "@context": "http://schema.org",  
633   "@type": ["Recipe", "NewsArticle"],  
634   "headline": "Apple Pie by Grandma Ople",  
635   "datePublished": "1999-10-08T17:19:13.000-04:00",  
636   "dateModified": "2025-03-19T12:41:06.722-04:00",  
637   "author": [ { "@type": "Person",  
638     "name": "MOSHASHAMA" } ],  
639 }
```

```
697   "prepTime": "PT30M",  
698   "recipeCategory": ["Dessert"],  
699   "recipeCuisine": ["American"],  
700   "recipeIngredient": [  
701     "8 small Granny Smith apples, or as needed",  
702     "0.5 cup unsalted butter",  
703     "3 tablespoons all-purpose flour",  
704     "0.5 cup white sugar",  
705     "0.5 cup packed brown sugar",  
706     "0.25 cup water",  
707     "1 (9 inch) double-crust pie pastry, thawed" ],  
708   "recipeInstructions": [  
709     {  
710       "@type": "HowToStep",  
711       "image": [  
712         {  
713           "@type": "ImageObject",
```

Ingredients

✓ 1X 2X 4X

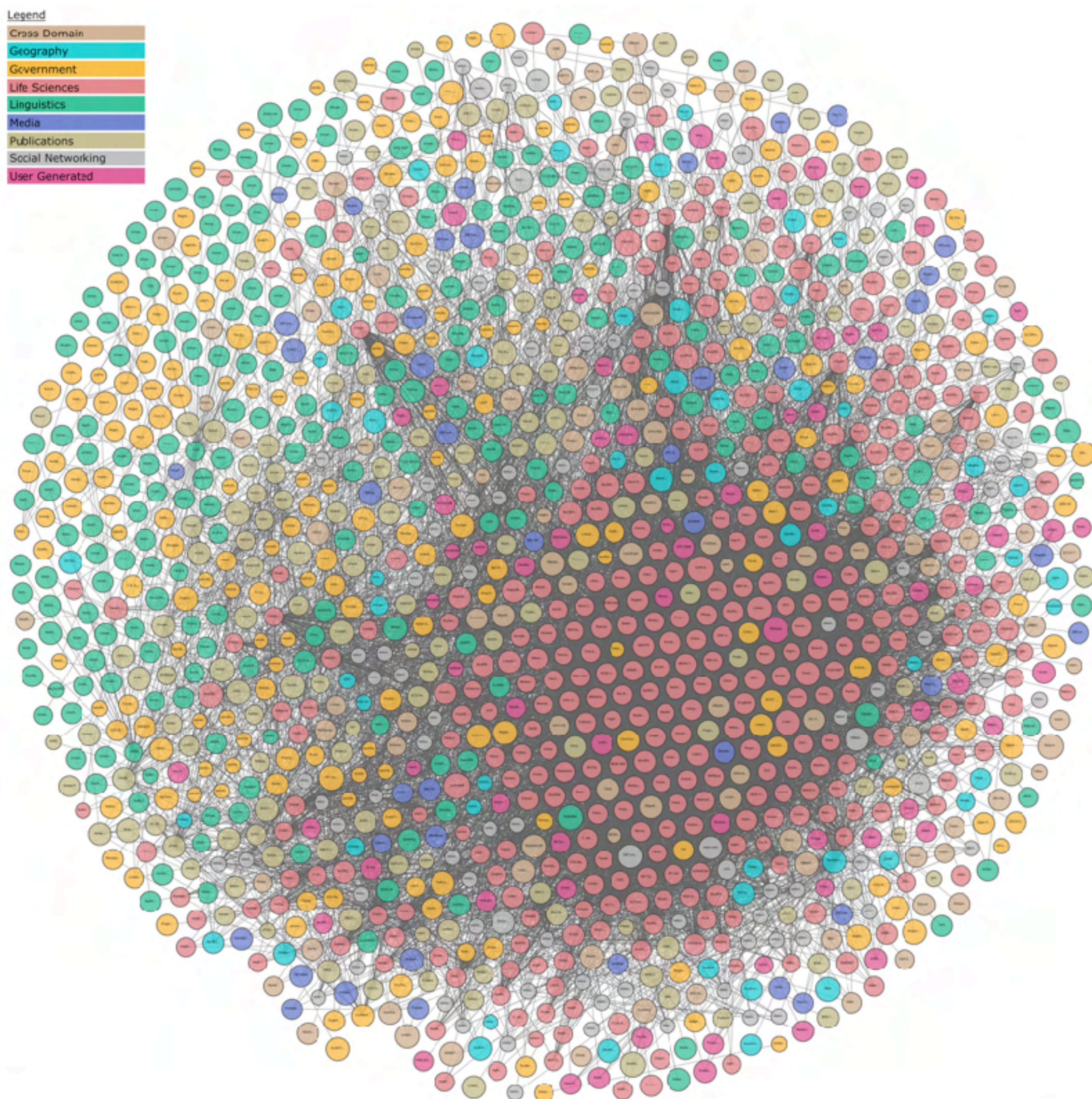
Original recipe (1X) yields 8 servings

- 8 small Granny Smith apples, or as needed
- ½ cup unsalted butter
- 3 tablespoons all-purpose flour
- ½ cup white sugar
- ½ cup packed brown sugar
- ¼ cup water

Local Offers

Oops! We cannot find any ingredients on sale near you. Do we have the correct zip code?

Semantic Web 101



The Linked Open Data Cloud from lod-cloud.net



https://lov.linkeddata.es/dataset/lov/ 67% ☆

VOCABS TERMS AGENTS SPARQL/DUMP

Linked Open Vocabularies (LOV)

+ Suggest Documentation Follow

868 Vocabularies in LOV

Category Tags

- Metadata
- Methods
- Industry
- IoT
- Services
- Support
- Geography
- Catalogs
- Society
- Environment
- API
- RDF
- People
- Quality
- General & Upper

Latest insertion

- coach** - Character of a Character Ontology (COACH) 2025-02-05
- server** - YANG Server Ontology 2025-02-05
- dseco** - The DNS Security Ontology 2025-02-05
- och** - OWL Change Ontology 2024-12-20
- gtm** - Gouda Time Machine Ontology 2024-12-10

Latest Updates

- m4i** - Metadata4ing: An ontology for describing the generation of research data within a scientific activity. 2025-03-07
- server** - YANG Server Ontology 2025-02-05
- coach** - Character of a Character Ontology (COACH) 2025-02-05
- itsmo** - IT Service Management Ontology 2025-02-05
- dseco** - The DNS Security Ontology 2025-02-05

Semantic Web x NLP

- Entity Linking
 - Milne, David, and Ian H. Witten. "Learning to link with wikipedia." In Proceedings of the 17th ACM conference on Information and knowledge management, pp. 509-518. 2008.
 - Mendes, Pablo N., Max Jakob, Andrés García-Silva, and Christian Bizer. "DBpedia spotlight: shedding light on the web of documents." In *Proceedings of the 7th international conference on semantic systems*, pp. 1-8. 2011.



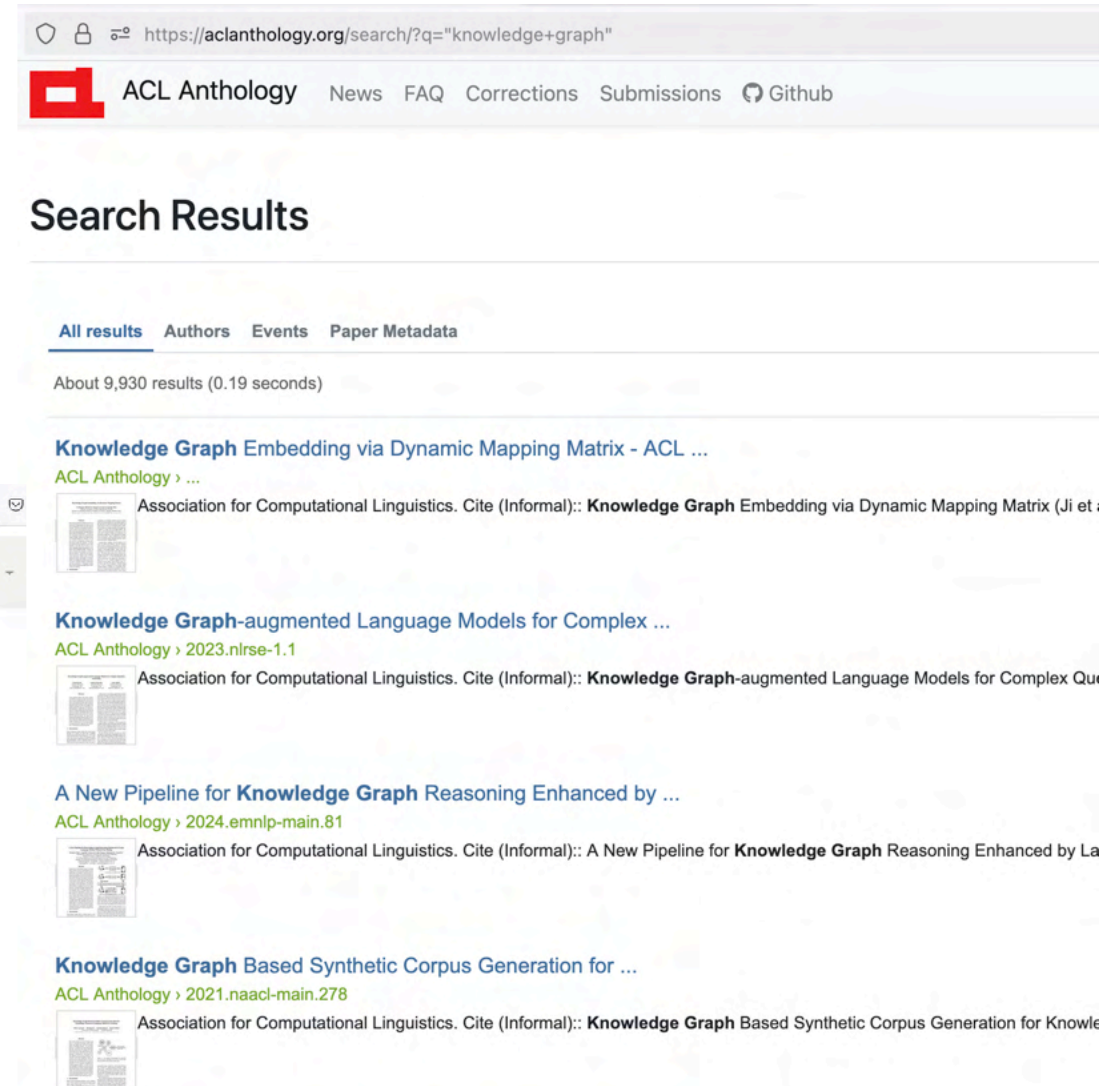
Figure 1: A news story that has been automatically augmented with links to relevant Wikipedia articles



Figure 1: DBpedia Spotlight Web Application.

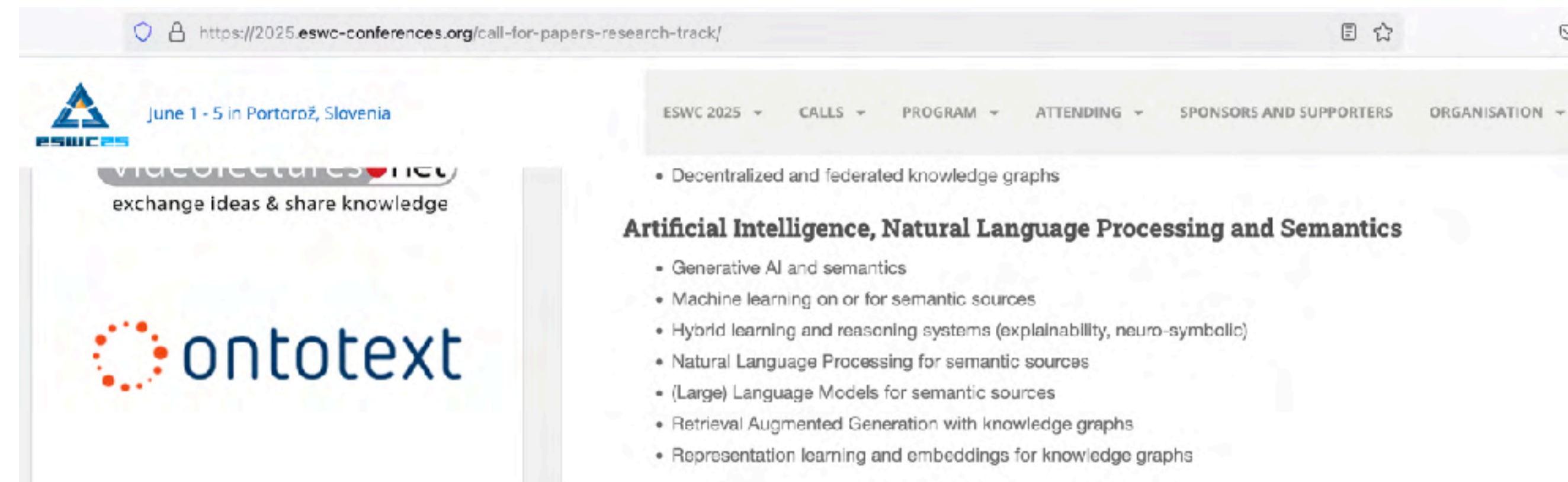
Semantic Web x NLP

- Workshop on Semantic Web and Information Extraction (SWAIE): 2013, 2014
- NLP & DBpedia workshop: 2013, 2014, 2015



The screenshot shows the ACL Anthology search results page for the query "knowledge+graph". The page displays a list of search results, including:

- Knowledge Graph Embedding via Dynamic Mapping Matrix - ACL ...**
ACL Anthology > ...
Association for Computational Linguistics. Cite (Informal):: **Knowledge Graph Embedding via Dynamic Mapping Matrix** (Ji et al.)
- Knowledge Graph-augmented Language Models for Complex ...**
ACL Anthology > 2023.nlrse-1.1
Association for Computational Linguistics. Cite (Informal):: **Knowledge Graph-augmented Language Models for Complex Qu**
- A New Pipeline for Knowledge Graph Reasoning Enhanced by ...**
ACL Anthology > 2024.emnlp-main.81
Association for Computational Linguistics. Cite (Informal):: **A New Pipeline for Knowledge Graph Reasoning Enhanced by La**
- Knowledge Graph Based Synthetic Corpus Generation for ...**
ACL Anthology > 2021.naacl-main.278
Association for Computational Linguistics. Cite (Informal):: **Knowledge Graph Based Synthetic Corpus Generation for Knowle**



The screenshot shows the ESWC 2025 call for papers page. The page includes the ESWC 2025 logo and the text "June 1 - 5 in Portorož, Slovenia". It also features a navigation menu with links to "CALLS", "PROGRAM", "ATTENDING", "SPONSORS AND SUPPORTERS", and "ORGANISATION". The main content area lists the topics for the workshop:

- Decentralized and federated knowledge graphs

Artificial Intelligence, Natural Language Processing and Semantics

- Generative AI and semantics
- Machine learning on or for semantic sources
- Hybrid learning and reasoning systems (explainability, neuro-symbolic)
- Natural Language Processing for semantic sources
- (Large) Language Models for semantic sources
- Retrieval Augmented Generation with knowledge graphs
- Representation learning and embeddings for knowledge graphs

Foodstuffs



Image source: https://static.ah.nl/static/recepten/img_006188_890x594_JPG.jpg

Constructing a Recipe Web from Historical Newspapers

Marieke van Erp
Melvin Wevers
Hugo Huurdeman



@merpeltje
@melvinwevers
@timelessfuture

Butter, salt & pepper

- Analysis of food customs:
 - historians
 - dieticians
 - ethnologists
- 1945 - 1995 Parool, Volkskrant, NRC & Trouw
- Dataset and code available through: <https://github.com/DHLab-nl/historical-recipe-web>
- Winner National Library - Rijksmuseum - Network Digital Heritage HackaLOD Hackathon
- You & other researchers are invited to work with us on case studies around food culture



Image source: https://assets3.thrillist.com/v1/image/1623749/size/tl-horizontal_main_2x.jpg



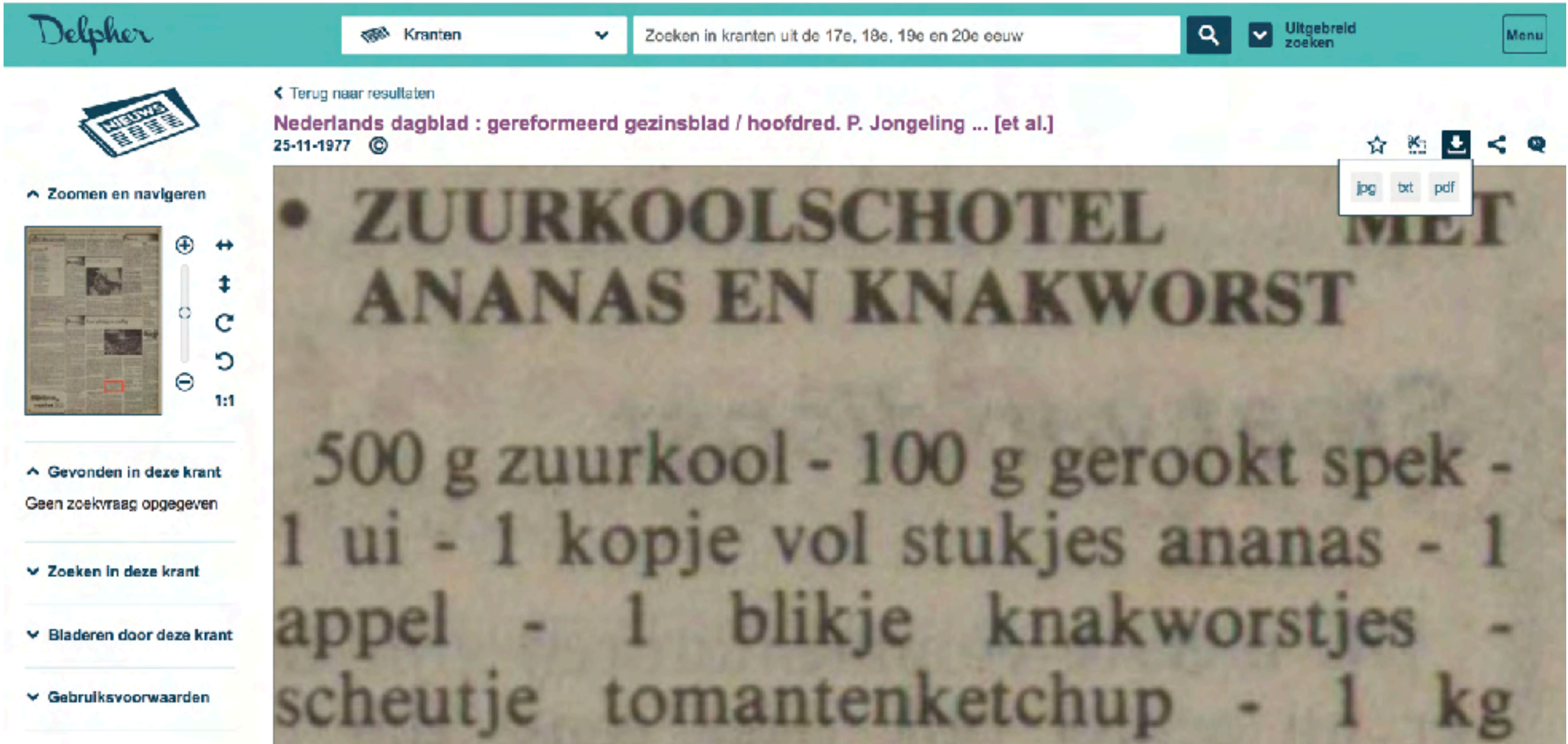
Newspapers as a source for recipes

- perception of a Dutch food culture formed in the 1950s
- newspapers are producer and messengers of public discourse
- newspapers contain views on daily life and customs
- But:
 - keyword search for 'recepten' imprecise
 - noise from digitisation process

Newspaper dataset

- Dutch National Library has digitised 90+ million book, newspaper and magazine pages
- Newspapers published between 1618 - 1995 from the Netherlands, the Dutch Indies (present day Indonesia), the Antilles, the US and Surinam (15% of all newspapers published in the Netherlands)
- Available via website, data dump (until 1876) and API (with agreement)

	Pages	Articles	Tokens
Parool	14.194	2,380,697	612,036,106
Volkskrant	13.628	2,248,652	744,275,792
NRC	7.199	947.198	489,397,816
Trouw	13.891	2,578,731	656,941,631
Total:	48.912	8,155,278	2,502,651,345



DINSDAG

Kopje ossestaartsoep

Knakworstjes

Witte bonen met ananas

Lofsla

Witte bonen met ananas

1 of 2 blikken witte bonen in tomatensaus, 1 uitje, 1 eetlepel mosterd, 2 eetlepels keukenstroop, 12 knakwostjes, 1 klein blikje ananasblokjes.

Rasp het uitje en meng dit door de bonen met tomatensaus. Doe de groente in een ingevette vuurvaste schaal. Roer de mosterd door de stroop en giet hier de helft van over de bonen. Dek de schaal af met een deksel of folie en zet hem 20 minuten in een warme oven. Verwarm inmiddels de worstjes in warm water en leg deze warm op de

dinsdag

6 ossestaartsoep

HUt *orstjes

l 0(* bonen met ananas

\t t e bonen met ananas

Va0,1 2 blikken witte bonen In 1 uitje

1 eetlepel ?lWd- 2 eetlepels keuken-

12 knakwostjes, 1 klein

„ ftaLananasDlokJes- SoJrJi het uitJe

en meng dit ,Qoe h bonen met

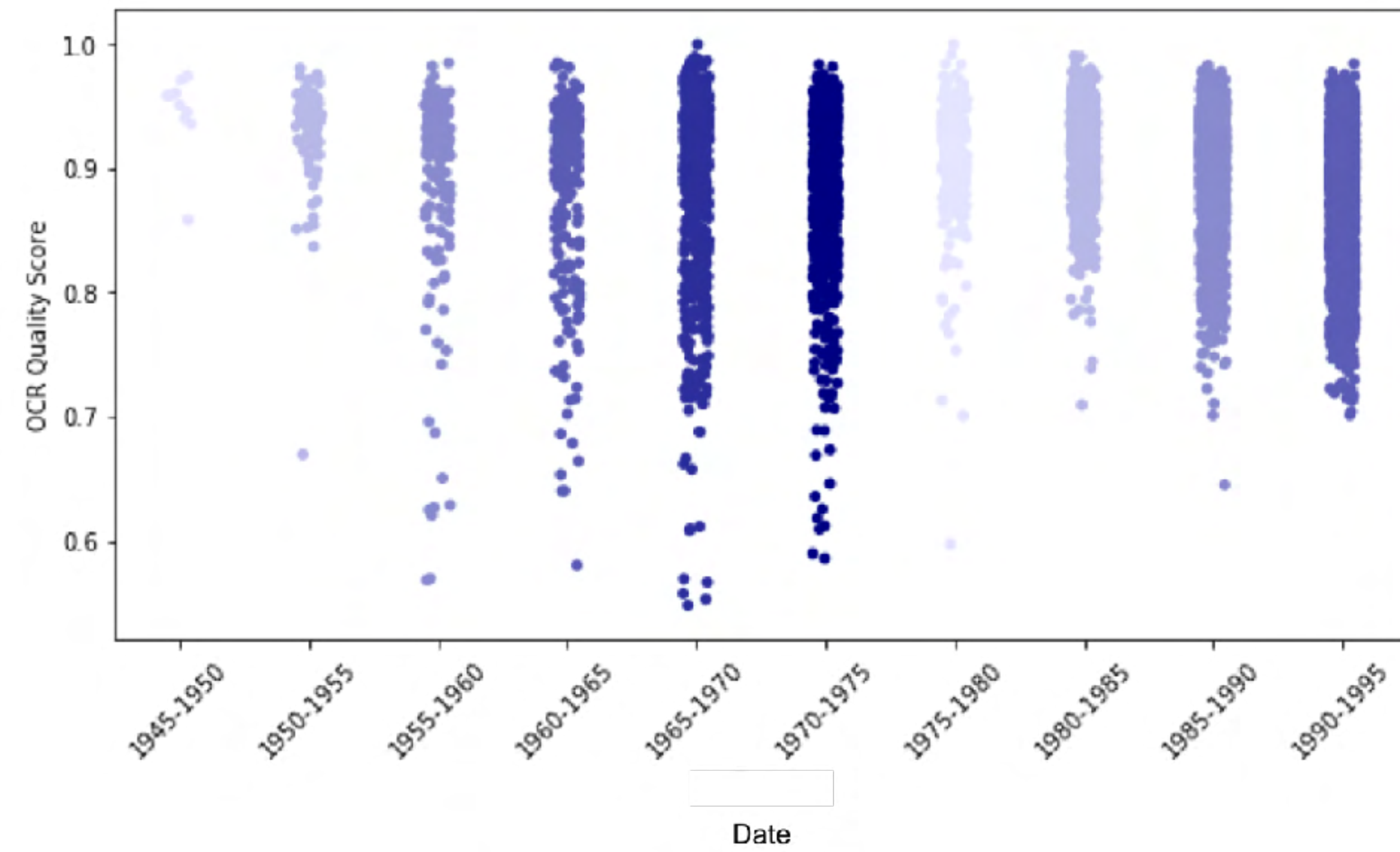
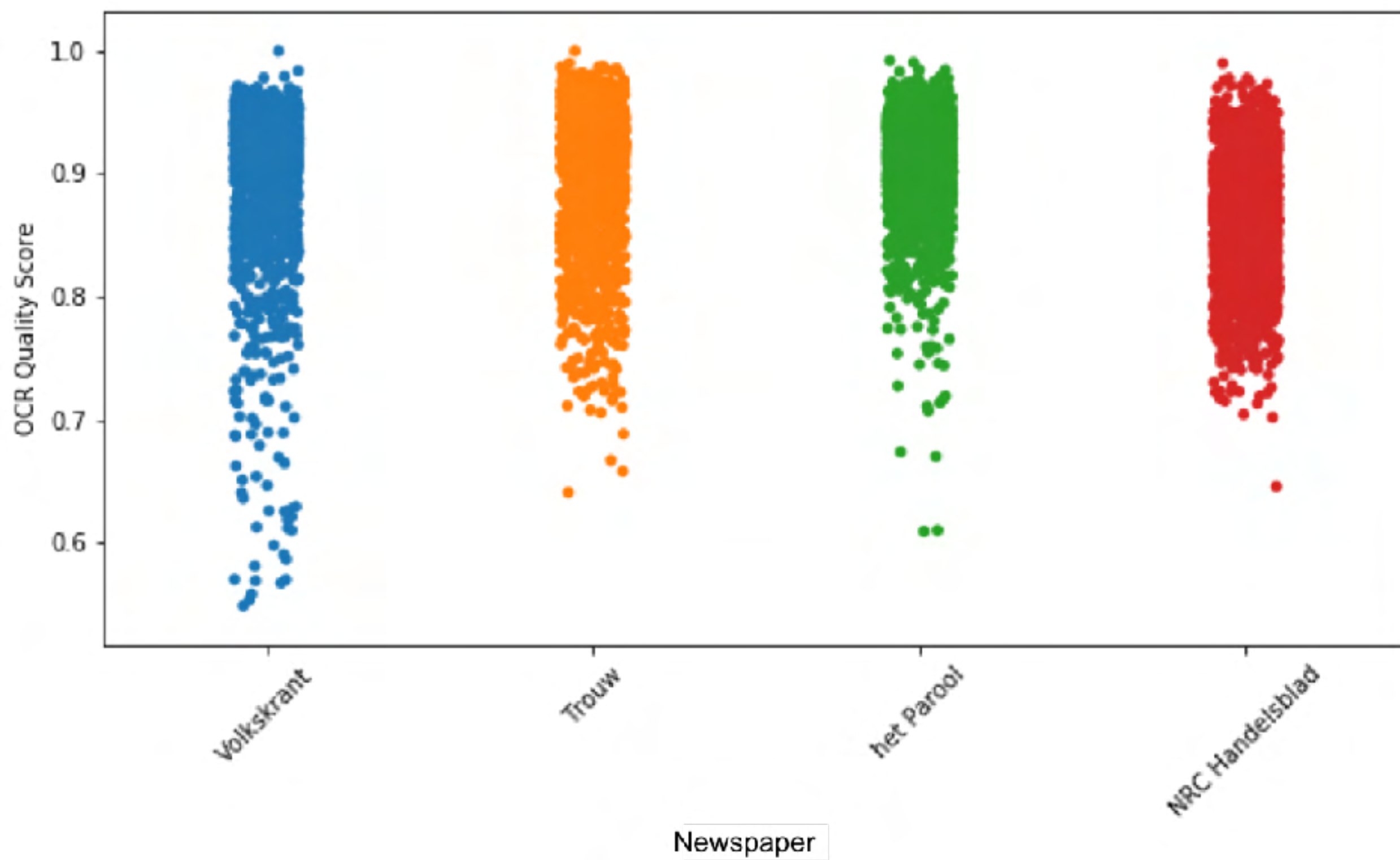
tomatensaus. Nir,;e groente in een

ingevette ?fd h ste schaal. Roer de

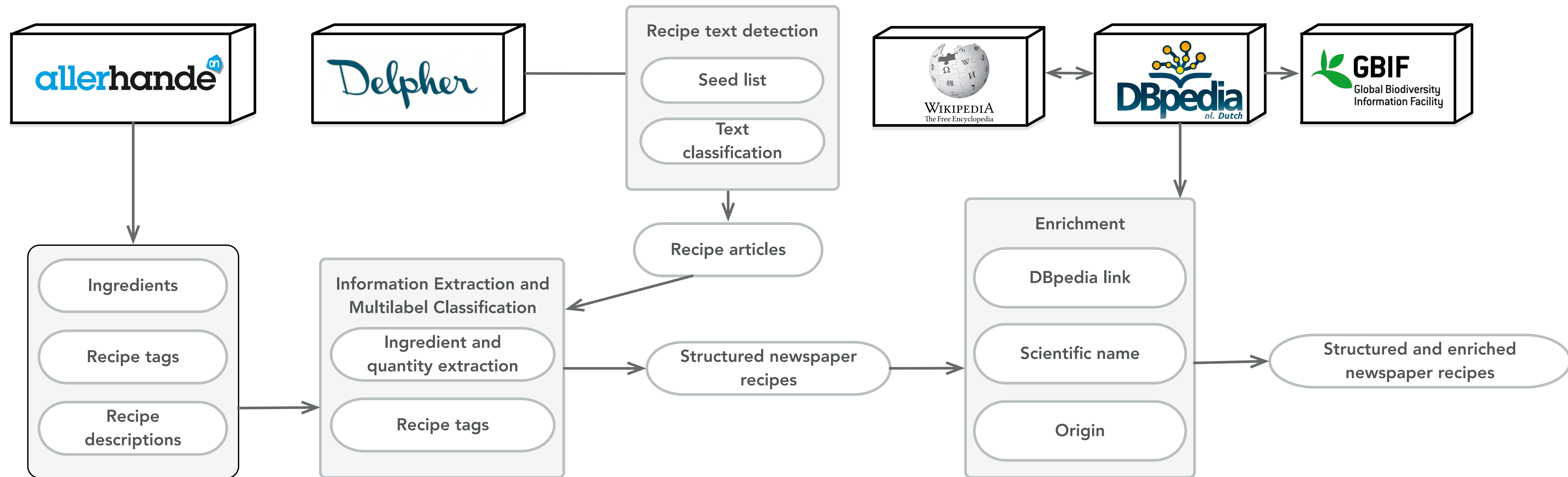
mos- Je hni?or de stroop en giet hier

2* van over de bonen. Dek af met

OCR Quality



From newspapers to a recipe web





What & how much?

- articles cannot automatically be segmented
- OCR errors and non-grammatical sentences are a hurdle for standard NLP pipelines
- lexicon-based extraction of ingredients and quantities

Evaluation

- 100 articles were manually annotated using Recogito
- OCR errors in ingredients or quantities marked separately
- IAA .85 but OCR boundaries difficult:
j°ar,anen' vs °ar,anen'
- Most precise lexicon: f1 = .67
- More research is needed for out-of-lexicon ingredients

MAANDAG

Knakworstjes Witte bonen met ananas Lofsla Vruchtenyoghurt Witte bonen met ananas 1 of 2 blik
-1 eetlepel mosterd - 2 eetlepels keukenstroop - 8 a 12 knakworstjes - 1 klein blikje ananasstukjes.
tomatensaus. Doe de bonen over in een met boter ingesmeerde vuurvaste schaal. Roer de mosterd
stroop zachtjes wordt verwarmd. Giet de helft van dit mengsel uit over de bonen. Dek de schaal af
geheel in de oven. Ongeveer 20 minuten tot alles goed warm is. Verwarm ondertussen de knakwor
zodat ze niet kapot springen. Leg deze als de bonen warm zijn in de schaal met de ananasstukjes (c
stroop over uit. Zet de de schaal terug in de oven zonder deksel gedurende 10 minuten.

Monday

Knackwurst White beans with pineapple. Belgian endive salad. Fruit
pineapple. 1 or 2 cans (jars) white beans in to-. mato sauce - 1 onio
2 tablespoons syrup - 8 to 12 knackwursts - 1 small can of pineapp
mix with the beans in tomato sauce. Put the beans in a with butter s
mustard through the syrup. This will be easier if the syrup is heated
mixture over the beans. Cover the dish with a lid or aluminium foil
mins until it's all warmed up. In the meantime, heat the knackwurst
don't burst. Put them in the dish together with the pineapple chunk
are hot and pour on the rest of the syrup. Put the dish back in the o

Results ingredients extraction

	Clean Ingredients				With OCR errors		
	precision	recall	f ₁	correct	precision	recall	f ₁
Allerhande	0.70	0.65	0.67	998	0.70	0.59	0.64
DBpedia	0.60	0.33	0.47	513	0.60	0.30	0.45
WordNet	0.62	0.50	0.56	764	0.62	0.45	0.54
AH-DBP-WN	0.56	0.75	0.66	1,154	0.56	0.68	0.62

27,411 new (old) recipes

- 34,479 Tags
- 365,133 ingredients
- >17,000 Links to external sources
- Data and software available at: <https://github.com/DHLab-nl/historical-recipe-web>



Wat eten we

WOENSDAG

Gehaktballen in een pittige saus
(Fins recept)

Andijvie, Aardappelen

Griesmeelpudding met bessesaus.

Gehaktballen in pittige saus.

Hier hebben we voor nodig: 500 gram rundergehakt, 2 koude gekookte aardappelen, 1 kleine gekookte biet, 2 uien, 1 ei, 80 gram boter, 1 lepel bloem, $\frac{1}{2}$ dl bouillon of water, een scheut koffieroom, wat mosterd, 1 klein blikje tomatenpuree.

Kook een ui in weinig water gaar. Hebt u een vleesmolen? Haal dan de gekookte aardappelen, de biet, de gekookte ui en het gehakt er tweemaal doorheen. Maak onder de middag

Take home message

- OCR errors can impact information extraction
- OCR post-correction is an active research field, but errors will remain
- Focus on most important elements to extract

Acknowledgements:

KB Koninklijke Bibliotheek
National Library
of the Netherlands



RIJKS MUSEUM

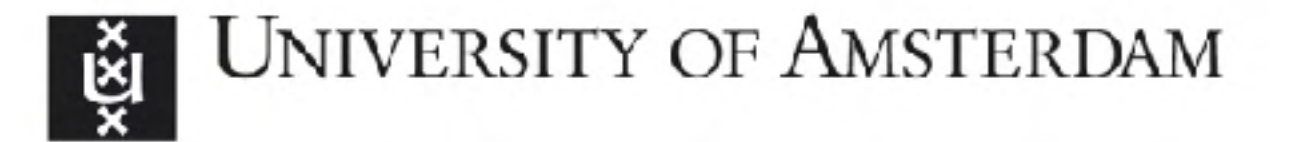


Image source: https://twelvemilesfromalemondotcom.files.wordpress.com/2014/09/img_0326.jpg

Communicating the environmental impact of plant based recipes

Collaborative project

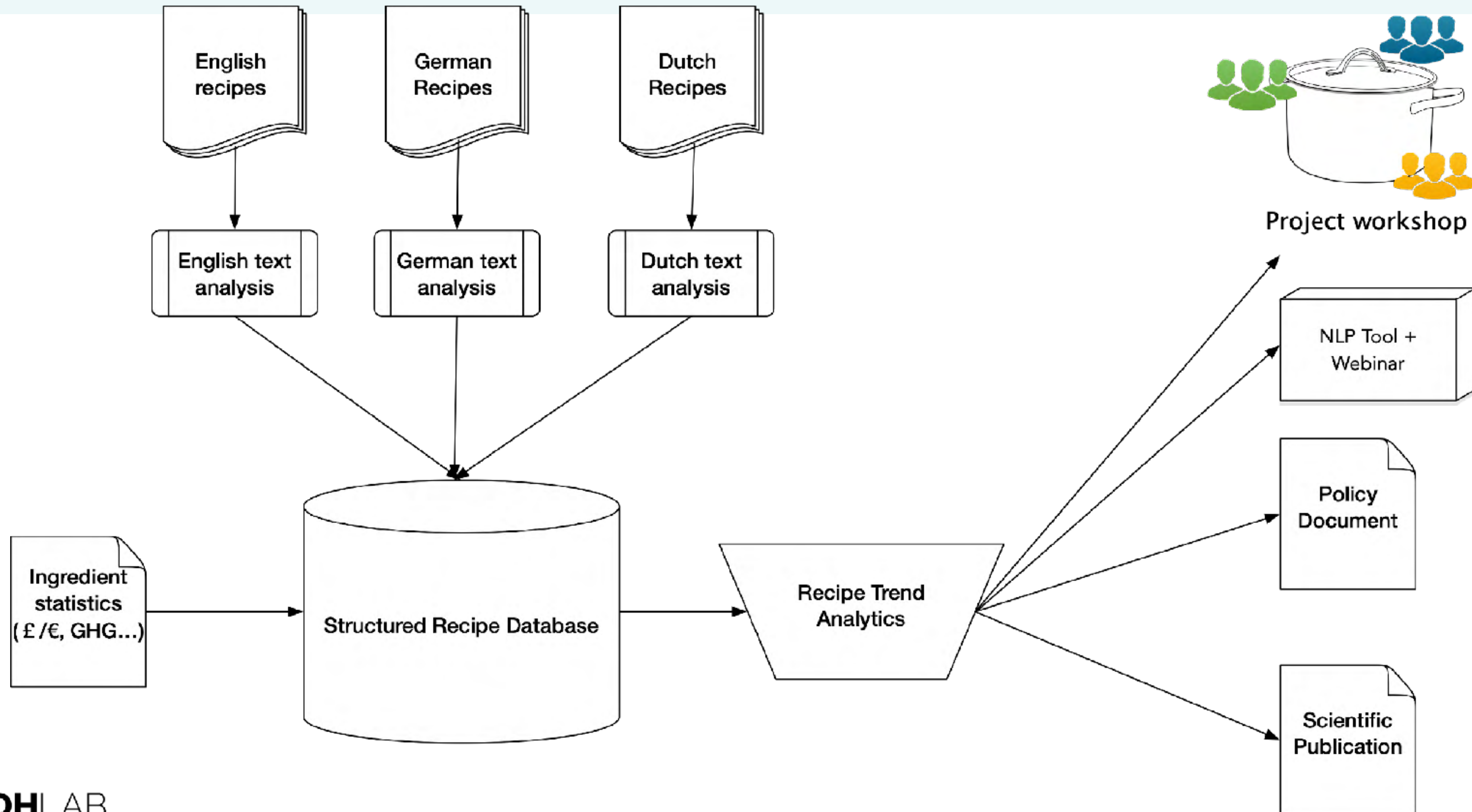
	Name (last, first)	Position/Title	Institution	Dept./Division
1.	Reynolds, Christian	Senior Lecturer	City, University of London	Centre for Food Policy
2.	Maynard, Diana	Senior Research Fellow	University of Sheffield, UK	Department of Computer Science
3.	Brewer, Stephen	Consultant	TextMiningSolutions Ltd, UK	N/A
4.	Ibáñez Martín, Rebeca	Senior Researcher	Meertens Instituut,	Anthropology & STS
5.	Van Erp, Marieke	Team leader	KNAW Humanities Cluster	Digital Humanities Lab
6.	Trattner, Christoph	Associate Professor	University of Bergen	Information Science & Media Studies Department
7.	Starke, Alain	Post Doctoral Researcher	Wageningen University & Research	Marketing & Consumer Behaviour Group

Measuring the sustainability of a diet

- Collect recipes
 - [allrecipes.com](https://www.allrecipes.com), [chefkoch.de](https://www.chefkoch.de), [ah.nl](https://www.ah.nl)
- analyse ingredients, quantities, if possible cooking method
- analyse groups of recipes' carbon footprints



Pipeline



Integrated Dataset

ALPRO FOOD SUSTAINABILITY

18/09/2021

Quick search 

 Sorting

 Export 

 Search



 Exit

Title	Vegetarian	Vegan	Total GHGE	Total 2007 GHGE	Total Land Use	Total Freshwater Withdrawals	Total Stress Weighted Water Use	Total Eutrophying Emissions	Total Calories	Serves	Ingredient List
Penne met hartige venkel-pancettasaus	No	No	5.4 kg CO2eq	5.1 kg CO2eq	8.16 m2	917.29 L	37153.64 L	32.78 g PO43-eq	149.0 J	[4]	[garlic, olive oil, pancetta, penne]
Kale with Caramelized Onions	Yes	Yes	0.03 kg CO2eq	0.03 kg CO2eq	0.07 m2	6.56 L	493.86 L	0.16 g PO43-eq	184.0 J	[4 servings]	[almonds, bunch curly kale, garlic, olive oil, vinegar, water, ½ onion]
Kamikaze Cocktail	Yes	Yes	0.0 kg CO2eq	0.0 kg CO2eq	0.0 m2	0.0 L	0.0 L	0.0 g PO43-eq	25.0 J		
Lemon or Vanilla Blintzes with Blueberry or Strawberry Sauce Recipe	Yes	No	5.94 kg CO2eq	5.86 kg CO2eq	0.15 m2	14.48 L	549.16 L	0.25 g PO43-eq	1225.0 J	[4 servings]	[6 eggs, blueberries, butter, cheese, cinnamon, lemon, milk, potato starch, sugar, water]
Lemon-Orange Walnut Bread Recipe	Yes	No	0.66 kg CO2eq	0.65 kg CO2eq	0.04 m2	9.08 L	416.01 L	0.11 g PO43-eq	1643.0 J	[12 servings]	[2 large eggs, butter, flour, lemon, lemon juice, milk, orange, splenda, vanilla extract, walnuts]
Salade met	Yes	No	2.47 kg CO2eq	2.22 kg CO2eq	9.9 m2	792.16 L	31338.05 L	11.27 g PO43-eq	0.0 J	[4]	[manchego cheese, peas, pecans]
Lamskoteletjes met geroosterde groenten en salsa verde recept	No	No	5.18 kg CO2eq	4.29 kg CO2eq	0.29 m2	27.35 L	2050.91 L	0.5 g PO43-eq	956.0 J	[4]	[1 shallot, 1 zucchini, 8 lamb chops, basil, capers, coriander, extra virgin olive oil, garlic, mint, parsley]

Nasty words*

***The next few slides may contain derogatory words and phrases. They are provided solely as illustrations of the research results and do not reflect the opinions of the authors or their organisations.**

Support the Guardian

Available for everyone, funded by readers

Contribute →

Subscribe →

Search jobs

Sign in

Search

International edition ▾

The
Guardian
For 200 years

News

Opinion

Sport

Culture

Lifestyle

More ▾

World UK Coronavirus Climate crisis Environment Science Global development Football Tech Business Obituaries

Artificial intelligence
(AI)

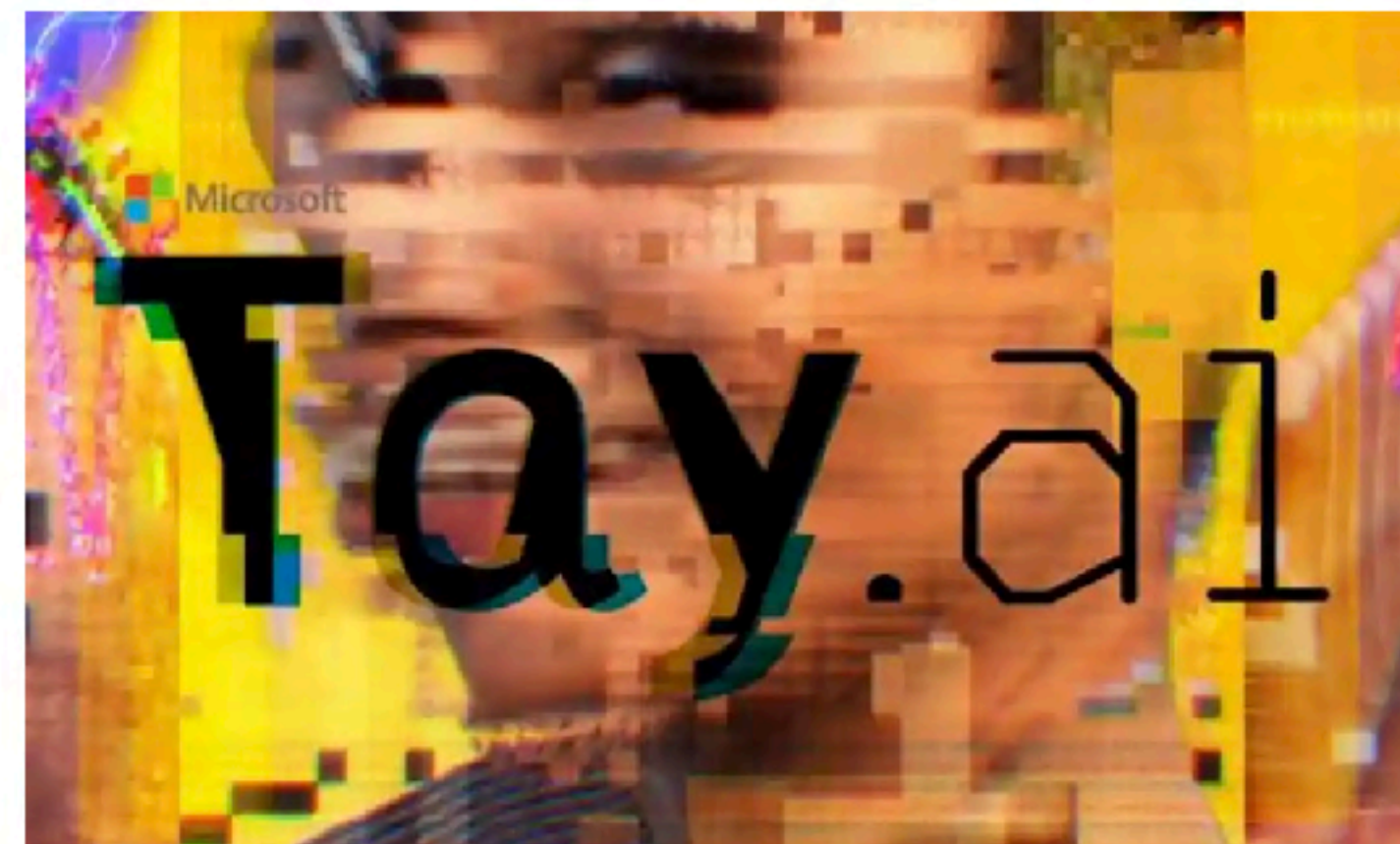
This article is more than 5 years old

Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter

Attempt to engage millennials with artificial intelligence backfires hours after launch, with Tay Tweets account citing Hitler and supporting Donald Trump

Elle Hunt

Thu 24 Mar 2016 06.41 GMT



Tay uses a combination of artificial intelligence and editorial written by a team including improvisational comedians. Photograph: Twitter

Microsoft's attempt at engaging millennials with artificial intelligence has backfired hours into its launch, with waggish Twitter users teaching its chatbot how to be racist.

Advertisement

Subscribe to the
Guardian Weekly
with home delivery →





Ethical AI Trained on Reddit Posts Said Genocide Is OK If It Makes People Happy

Fed by Reddit and Mechanical Turks, the algorithmic morality spinner is a dangerous and stupid use of technology.



By [Matthew Gault](#)

November 3, 2021, 4:58pm

[Share](#) [Tweet](#) [Snap](#)



IMAGE: GETTY IMAGES

MORE LIKE THIS

Tech

This Website Changes Every Time You Blink

MATTHEW GAULT

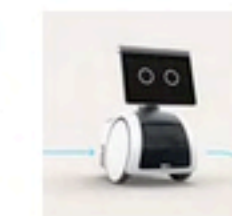
07.26.21



Tech

Leaked Documents Show How Amazon's Astro Robot Tracks Everything You Do

MATTHEW GAULT, JOSEPH COX



It's only gotten worse...

KM

write feedback for a bubbly receptionist



[Employee Name] is a bubbly and outgoing receptionist who consistently demonstrates a positive attitude and a friendly demeanor. She is a natural at building rapport with clients and visitors, and she is able to make them feel welcome and comfortable from the moment they arrive.

She is well organized and efficient in managing the front desk, and she is able to handle a variety of tasks such as answering phone calls, scheduling appointments, and directing visitors to the appropriate areas. She is also able to handle difficult situations with tact and diplomacy, and she is able to diffuse any potential conflicts or complaints in a professional manner.

One of her strengths is her ability to think on her feet and to adapt to changing situations. She is able to handle multiple tasks at once and to prioritize her work effectively. She is also a team player, always willing to help her colleagues when they need it.

However, in some situations, her bubbly and outgoing personality might be a distraction for some of her colleagues, and she should be mindful of that, and be respectful of the work environment.

Overall, [Employee Name] is an asset to the organization, and her positive attitude,

KM

write feedback for an unusually strong construction worker



[Employee Name] is an unusually strong construction worker who consistently demonstrates a high level of physical fitness and stamina. He is able to handle heavy equipment and materials with ease and is able to work long hours without showing signs of fatigue. He is able to complete tasks quickly and efficiently, and he consistently meets or exceeds productivity expectations.

He is a skilled worker and has a good understanding of construction techniques and safety protocols. He follows safety guidelines and is able to identify and mitigate hazards on the job site. He is also able to work well in a team and is able to take on leadership roles when necessary. He is able to train and mentor other workers on the team, helping to improve their productivity and safety.

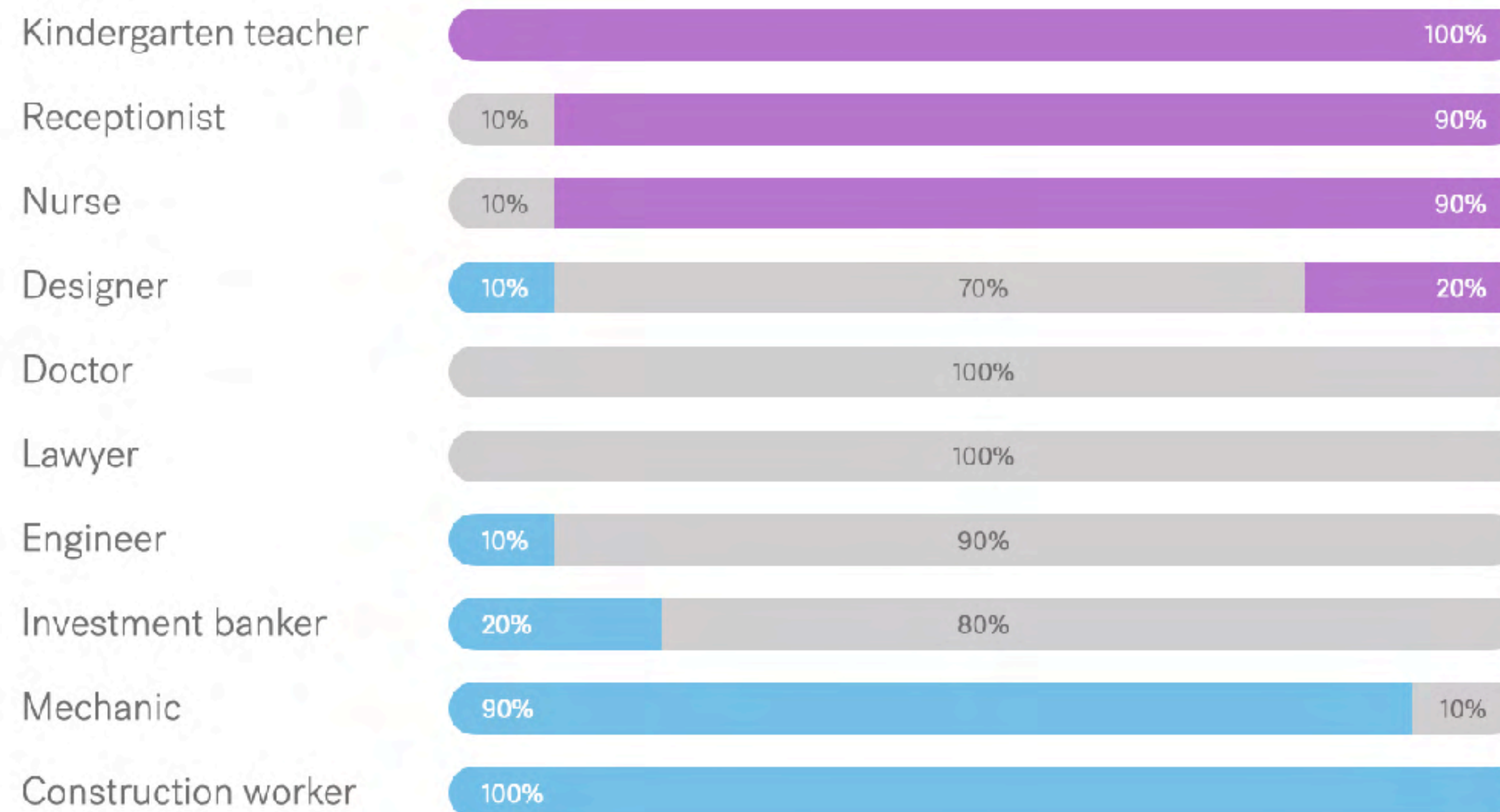
One of his strengths is his ability to think critically and to problem-solve on the job. He is able to come up with creative solutions to complex problems and is able to adapt to changing situations.

However, with his physical strength, it's important for him to be aware of the potential for injury to himself or others, and to use proper technique and equipment when lifting and moving heavy objects. He should also be aware of the potential for overconfidence in his abilities and make sure to follow safety procedures, and to communicate clearly with his team members to avoid accidents.

source: <https://www.fastcompany.com/90844066/chatgpt-write-performance-reviews-sexist-and-racist>

Why Might This Be Bad?

ChatGPT often relies on gender stereotypes to choose pronouns



Pronouns in AI-generated job performance feedback across different roles



Sample performance feedback prompts:

"Write feedback for a bubbly **receptionist**"

"Write feedback for an unusually strong **construction worker**"

"Write feedback for a helpful **mechanic**"

"Write feedback for a remarkably intelligent **engineer**"





Update...

Update...

Update...

Update...

Up

Update...

Update...

Update...

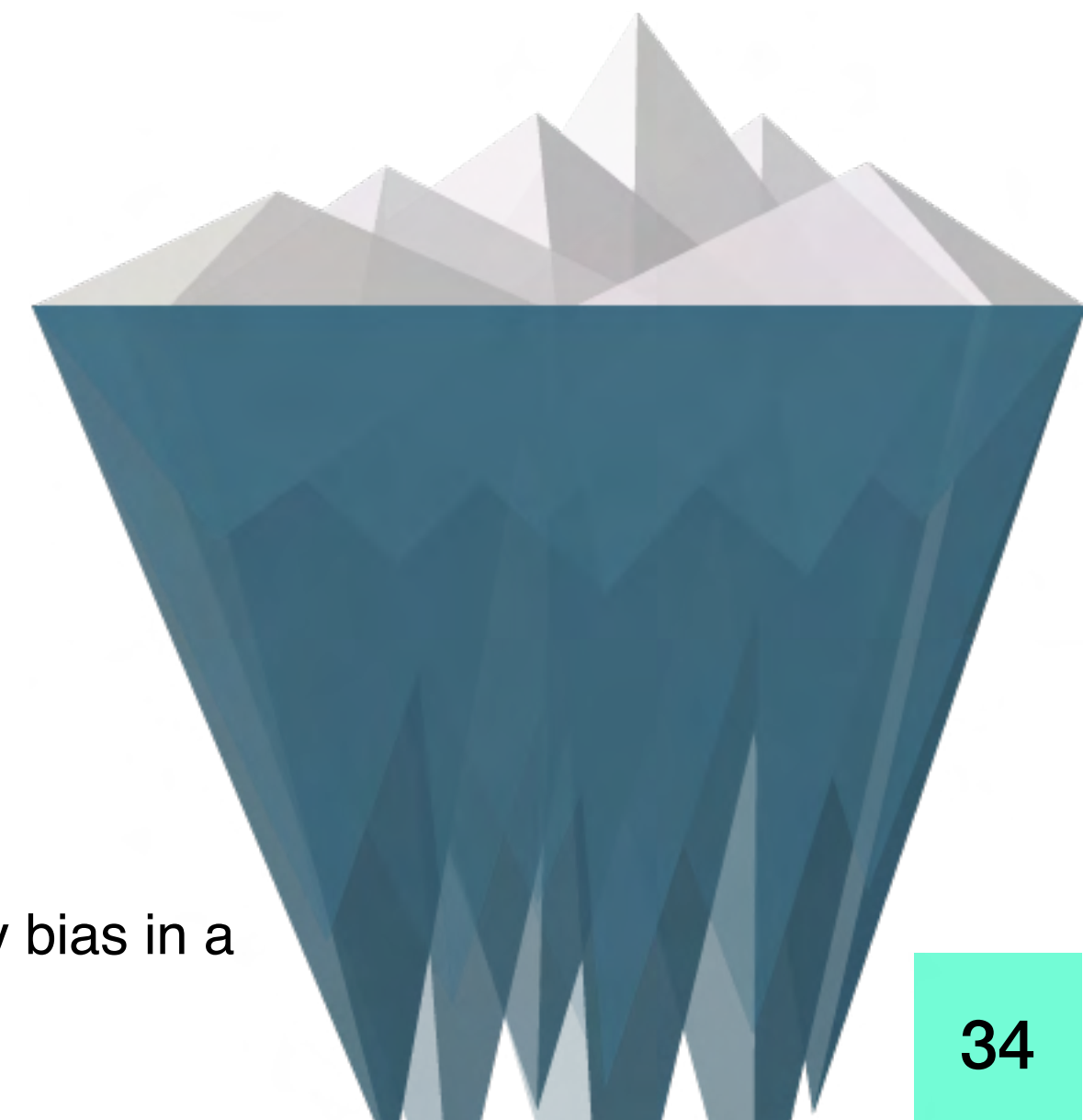
Update...

Update...

The Problem

- GLAMs have billions of pages of digitised content
- These collections cover decades/centuries
- Most content is hardly ever accessed by users*
- ‘Find & Replace’ of contentious terms is too crude

*Traub, M. C., Samar, T., Van Ossenbruggen, J., He, J., de Vries, A., & Hardman, L. (2016, June). Querylog-based assessment of retrievability bias in a large newspaper corpus. In *2016 IEEE/ACM Joint Conference on Digital Libraries (JCDL)* (pp. 7-16). IEEE.



Work in
Progress

Words Matter

An Unfinished Guide
to Word Choices
in the Cultural Sector

GLOSSARY OF TERMS

Exotic

HISTORY, USE & POSSIBLE SENSITIVITIES

This term is derived from the Ancient Greek word “*exōtíkos*,” literally meaning “from the outside.” It entered the Dutch language with the meaning of foreign/alien, which it still has today. The term has become intertwined with ideas about the (racialized and sexualized) Other.

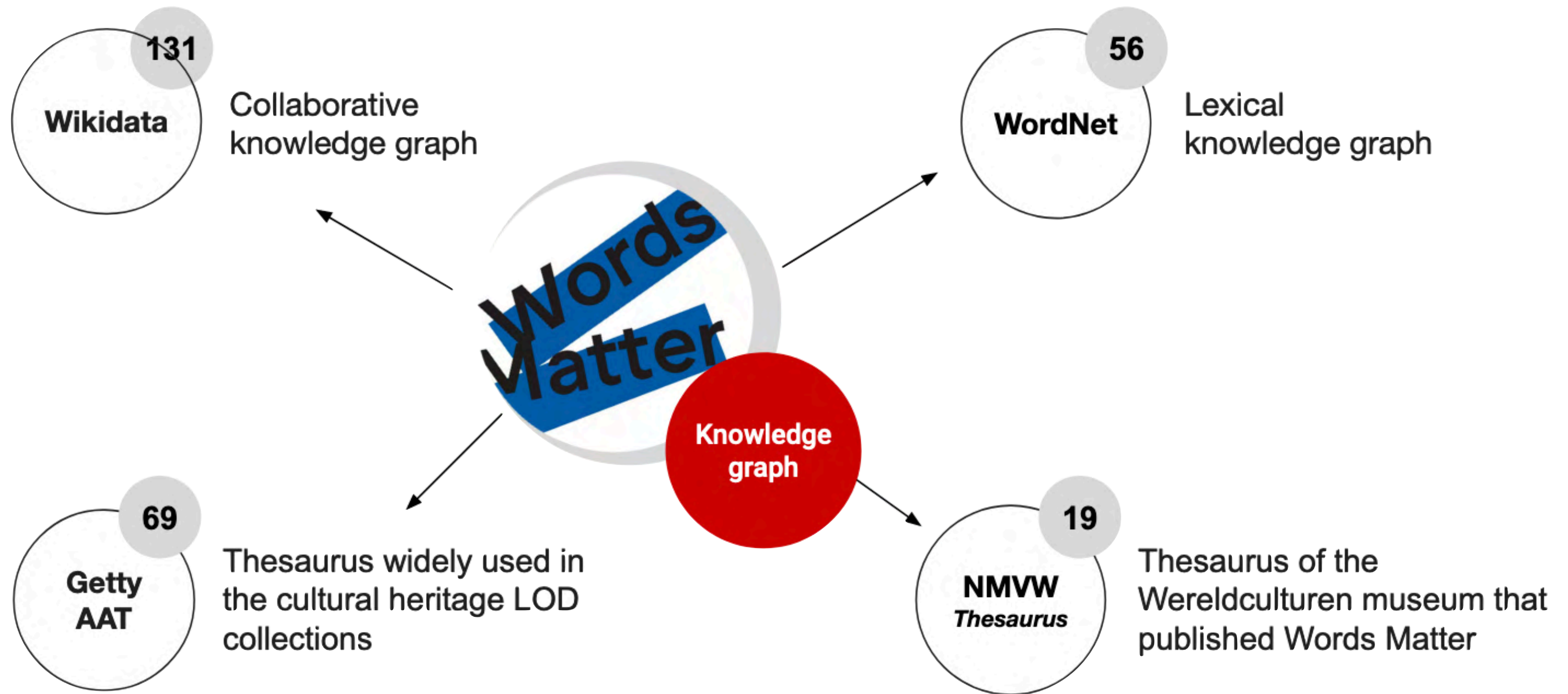
The term “exotic” is commonly used to describe plants and animals, but is also used for people (usually people of color), where it has a connotation of being different from the norm, especially in reference to appearance and name (for example “what an exotic name!”). Sometimes it has a sensual connotation.

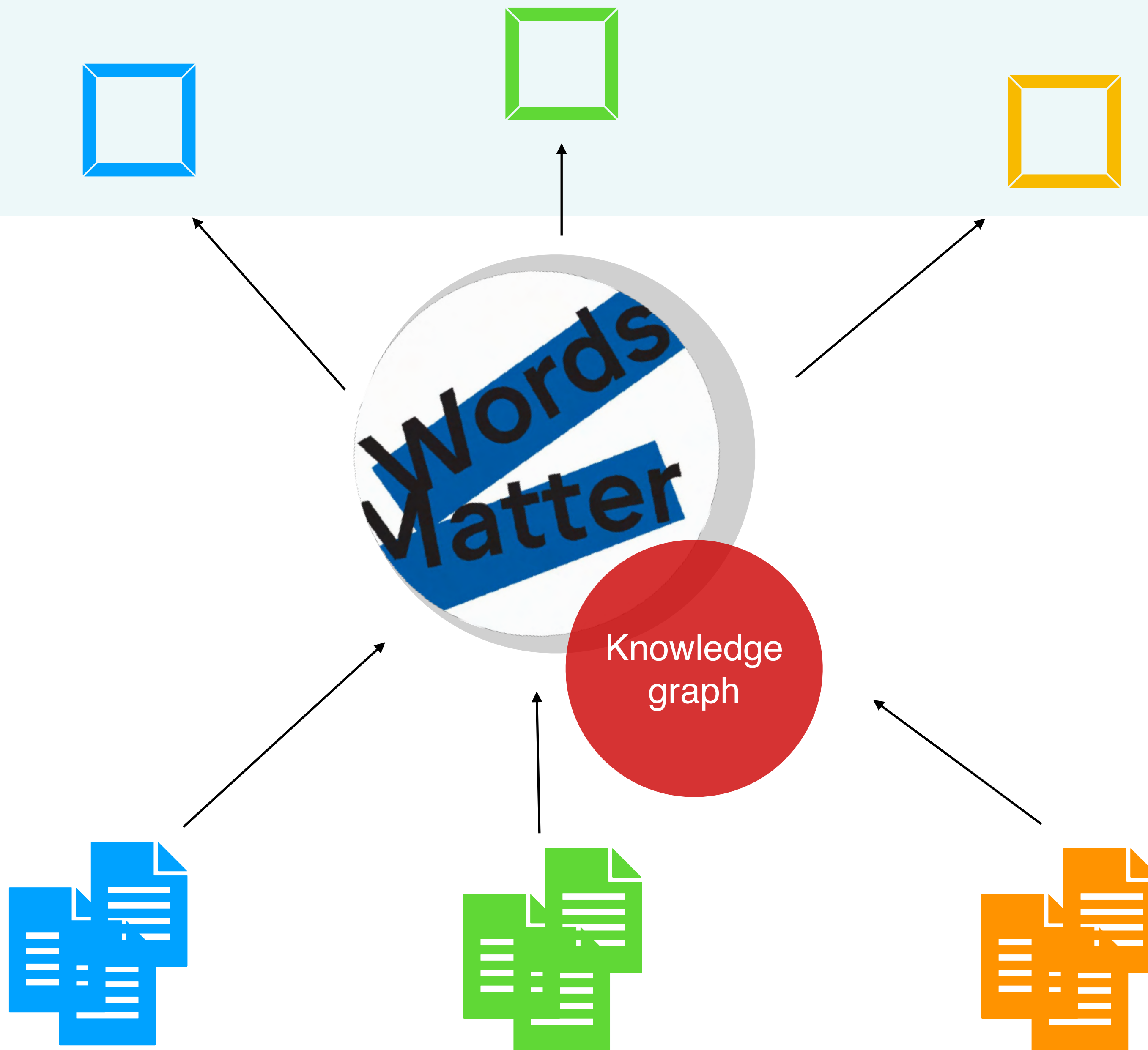
SUGGESTIONS

- Applicable when referring to plant and animal species. It is, however, contested to use the term to describe people.

Connections to other LOD-resources

skos:relatedMatch





Changing World Views & Language

- Our language reflects our world views
- Many archives contain language that does not longer correspond to current societal values
- How can we deal with this?

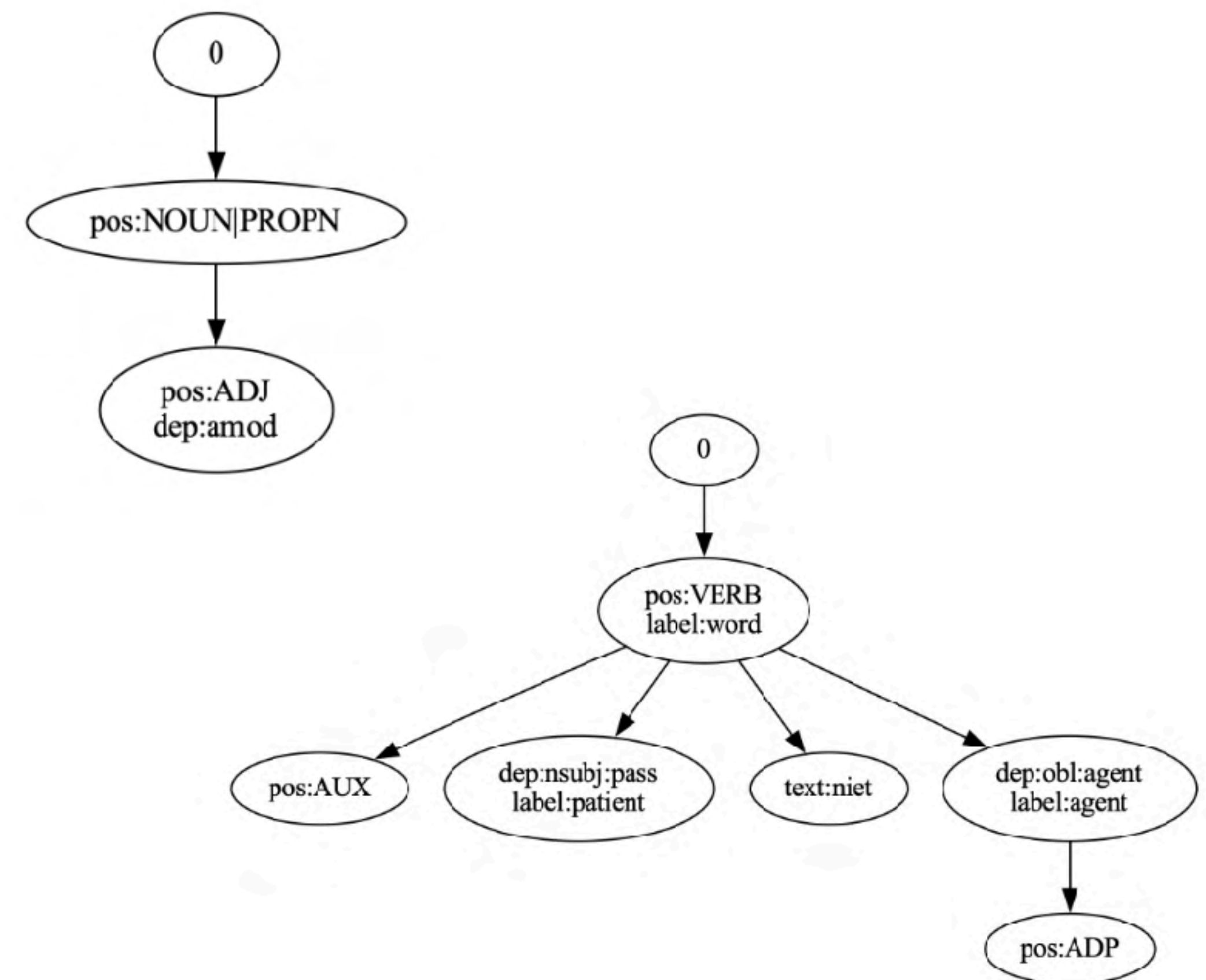


Ryan Brate, Marieke Van Erp, and Antal Van den Bosch. 2023. Contextual Profiling of Charged Terms in Historical Newspapers. In Proceedings of the 4th Conference on Language, Data and Knowledge, pages 97–108, Vienna, Austria.

Image source: <https://kaveripak.files.wordpress.com/2017/07/thetimes-broadwaymusical.jpg?w=600>

Contextual Profiling of Charged Terms

- Dutch newspapers 1950 - 1990
- Model contexts known charged nouns
- Use contexts to identify other, previously unknown, charged contexts
- POS-tag patterns, scored via log-likelihood ratio



Results

known-charged nouns	modifiers associated	adjectives associated	verbs associated for which the noun is the agent	verbs associated for which the noun is the patient
afkomst	bedoeïnen: 1991; bloedgroep: 1959; dansers: 1979; hugenoten: 1989; huurkamer: 1971		aanvuren: 1969; held: 1985; molesteren: 1995; verzekeren: 1995	eernt: 1959; opsieren: 1971; raden: 1959; schreeuwen: 1991; traumatiseren: 1985; verlooehenen: 1995, 1951, 1955, 1959, 1959, 1961, 1961, 1965, 1969, 1969, 1975, 1975, 1979, 1979, 1981, 1985, 1985, 1989, 1989, 1991, 1991, 1995; verraden: 1955, 1959, 1985, 1989, 1991, 1995; zullen: 1971
allochtoon		laag-opgeleid: 1991; werkloos: 1989; werkwillig: 1991	ongemerkt: 1991	hulpbehoeven: 1989; instromen: 1989
baboe		soendances: 1991; zorgvol: 1979	inhalen: 1951	
immigrant	commonwealth: 1971	afriaans: 1991; arriveren: 1951; bengaals: 1981, 1985; blank: 1981; duits-joods: 1971; engtls: 1955; enjels: 1951; hds: 1959; illegaal: 1971, 1979, 1995; indies: 1975; latiinsamerikaans: 1991; miljoen: 1955; mohammedaans: 1991; niet-blanke: 1965, 1981; niet-britse: 1955; niet-geïntegreerd: 1959; nietblank: 1965, 1985; niet-blank: 1965; noordafrikaans: 1995; opper-egyptisch: 1981; ouz: 1959; portoricaans: 1959; roemeens-duits: 1989; russisch-joods: 1989; russischjoods: 1989; s' amitsch: 1995; salvadorlaans: 1969; sicillaans: 1961; sovjet-joods: 1991; steenrijk: 1969; steunen: 1951; urdu-talig: 1995; westindlsch: 1981; ülegal: 1981	afpakken: 1989; binnensmokkelen: 1971; broeden: 1959; doodsteken: 1995; inpikken: 1989; klagen: 1951; meedragen: 1955; omsingelen: 1959; ontsluiten: 1981; overspoelen: 1989; terugbetalen: 1955; wegpikken: 1959; zijn: 1961	aankomen: 1975; afbeulen: 1985; classificeren: 1979; huisvesten: 1991; stijven: 1985; toelaten: 1951; verkrijgen: 1959
indiaan	amazone: 1969; apache: 1991; halfbloed: 1981; hopi: 1961, 1985; innu: 1989; miskito: 1985, 1989; navajo: 1991; noorda: 1959; oerwoud: 1981; platvoet: 1955, 1965; sioux: 1991; vanomami: 1991	amazon: 1989; benedenlands: 1965; bonairiaans: 1989; een-ogig: 1965; eenogig: 1965; eht: 1961; grondloz: 1979; ploeteren: 1991; rillen: 1959	aftroggelen: 1995; kapen: 1989; kauwen: 1991	achterstellen: 1989; afhakken: 1951; afslachten: 1969; hakken: 1965; verontwaardigen: 1989

Culturally Aware AI

- Looking at bias from different angles
- Integrating findings in GLAM interfaces
- 10 July Utrecht University: PhD defence Ryan Brate *Words Matter - A computational toolkit for charged terms*
- <https://cultural-ai.nl>

The screenshot displays the Alter Heritage web app interface. On the left, a historical image of a public execution is shown. Below the image, there is a form for adding metadata. The form includes fields for 'Titel' (Title), 'Beschrijving' (Description), and 'Onderwerp: Wat' (Subject: What). The 'Titel' field contains the text 'Ontdekking van sodomie in een klooster te Brugge, 1578'. The 'Beschrijving' field contains a detailed description of the event. The 'Onderwerp: Wat' field contains the text '1578-05-18 - 1578-05-18' and 'Cornelis, Broer'. To the right of the form, there are several buttons and a list of objects. The buttons include 'Herstellen #1', 'Indienen #1', and a series of numbered buttons (1-10). The list of objects includes a green box with the text 'Alle objecten zijn ingediend. Je kunt de taak nu afronden door dit venster te sluiten. Of je kunt wijzigingen aanbrengen in je bewerkingen en antwoorden en ze opnieuw indienen. Bedankt voor je deelname!' and a red box with the text 'Je kunt je bewerkingen indienen nadat je de verplichte vragen hieronder hebt beantwoord.'

Instructies

Herstellen #1 < 1 2 3 4 5 6 > Indienen #1

Alle objecten zijn ingediend. Je kunt de taak nu afronden door dit venster te sluiten. Of je kunt wijzigingen aanbrengen in je bewerkingen en antwoorden en ze opnieuw indienen. Bedankt voor je deelname!

Je kunt je bewerkingen indienen nadat je de verplichte vragen hieronder hebt beantwoord.

In hoeverre was naar jouw mening de representatie van LHBTQ+ personen in dit object inclusief VOOR je bewerkingen, waarbij 1 volledig niet-inclusief is en 10 volledig inclusief? *

1 2 3 4 5 6 7 8 9 10

In hoeverre is naar jouw mening de representatie van LHBTQ+ personen in dit object inclusief NA je bewerkingen, waarbij 1 volledig niet-inclusief is en 10 volledig inclusief? *

1 2 3 4 5 6 7 8 9 10

Wat zijn je beweegredenen voor het bewerken van de metadata van dit object? *

Uitleg te geven aan een offensieve term.
Een meer neutraal term te gebruiken in de metadata - vooral om meer objecten te kunnen vinden.
Misschien is een metadata tag "LHBTQI+" of "queer" nuttig om erbij te zetten maar misschien is dat te generiek.

Wil je andere bewerkingen uitvoeren die niet mogelijk zijn in deze interface? Zo ja, wat zijn deze bewerkingen?

Nee

Als je nog iets wil delen, laat dit dan hier achter.

+ Een nieuw veld toevoegen

Titel

Ontdekking van sodomie in een klooster te Brugge, 1578

sodomie is een bijbelse term dat refereert aan een homoseksuele ACT (sorry - ik ken de NLse woord er voor niet)

Mogelijke offensieve of kwetsende term

Beschrijving

Arrestatie van de Franciscaner monniken in een klooster te Brugge op beschuldiging van sodomie, 18 mei 1578. Onder begeleiding van soldaten worden de monniken uit het klooster afgevoerd. Onder hen de beruchte Broer Cornelis van Dordrecht. Met onderschrijf van 15 regels in het Duits. Linksonder genummerd: 145. De prent maakt deel uit van een album.

Onderwerp: Wat

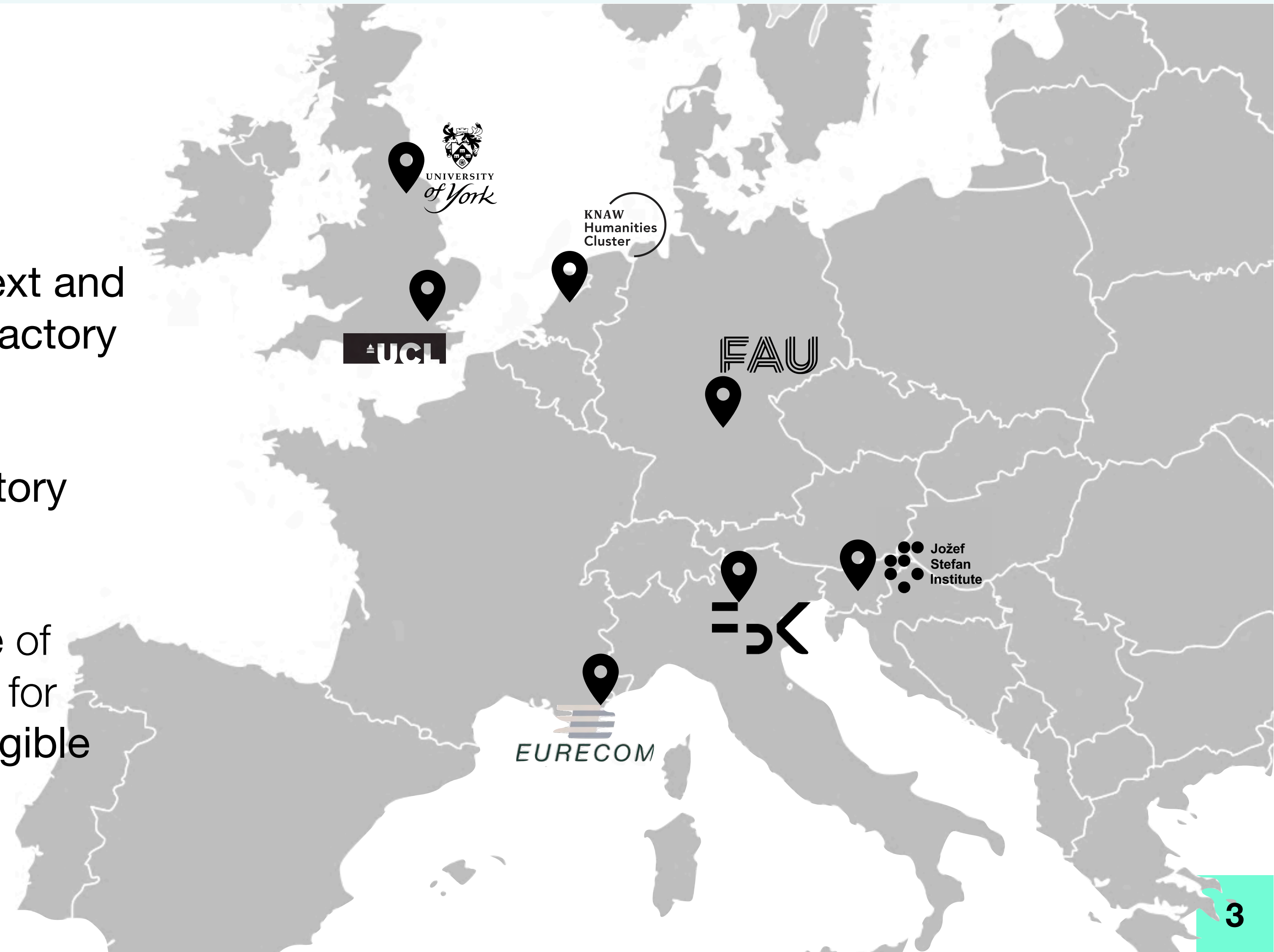
1578-05-18 - 1578-05-18 Cornelis, Broer +

Smelly Things*

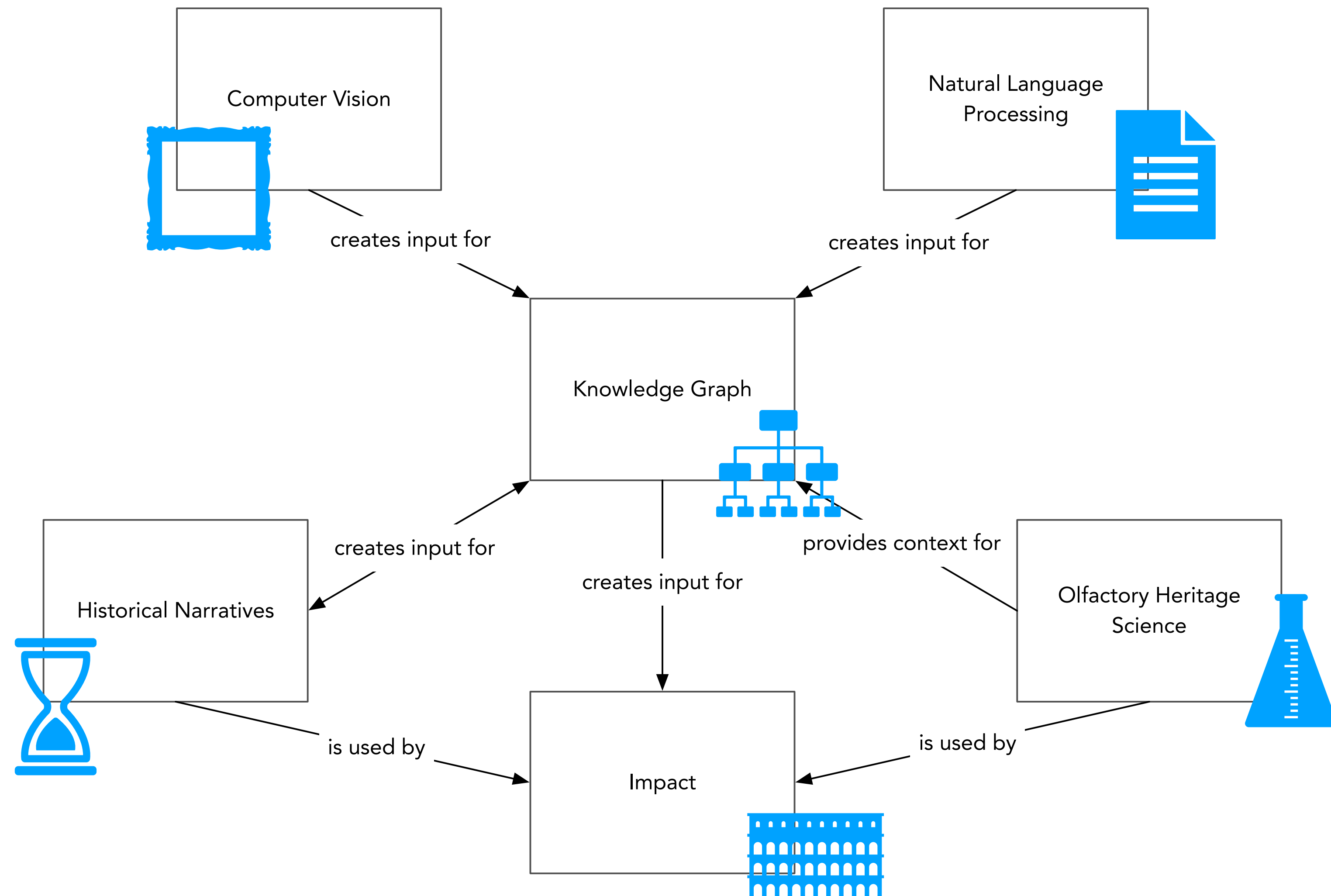
A Distributed Project

Objectives:

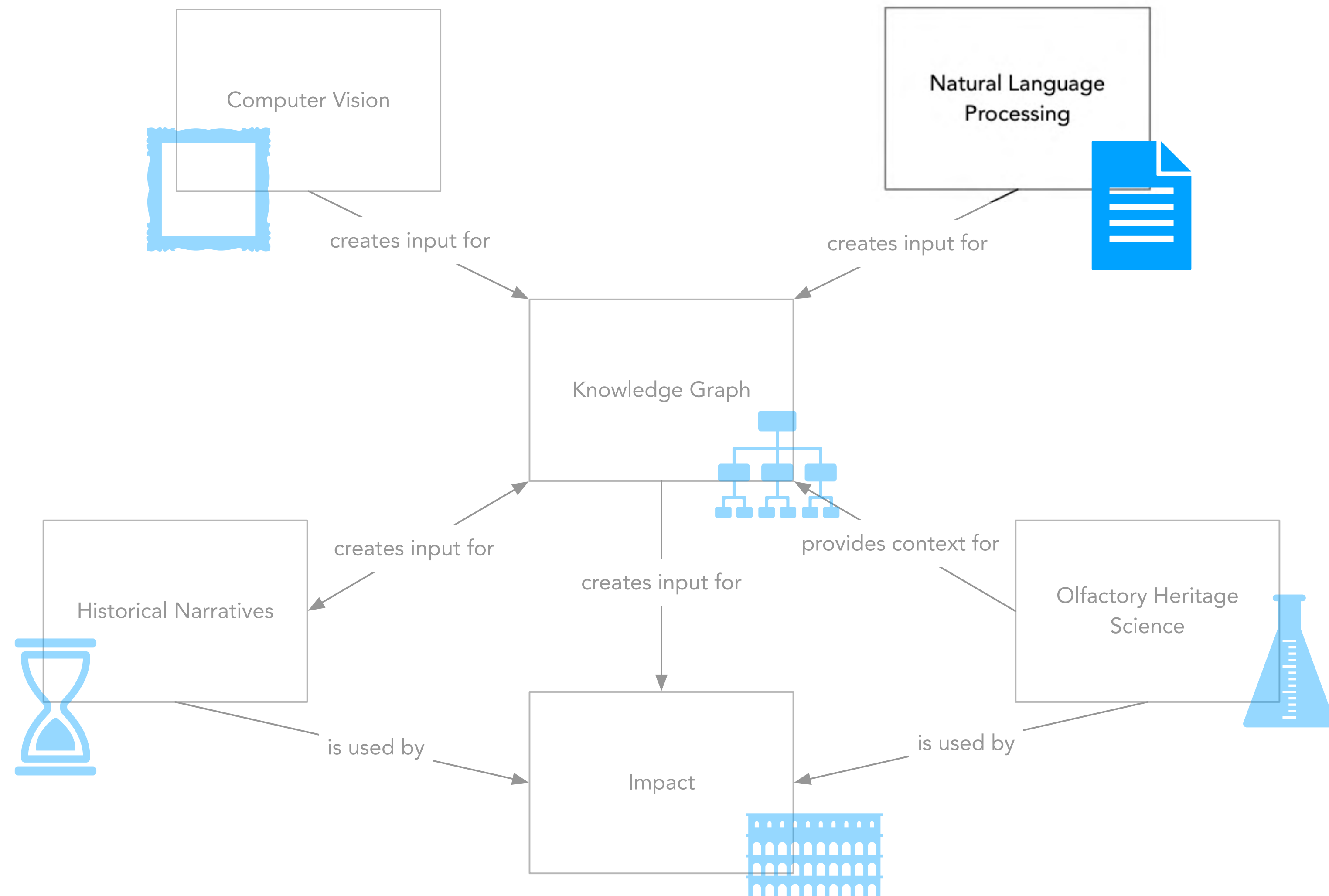
- develop state-of-the-art AI techniques to identify olfactory information in digital text and image collections; to curate historical olfactory narratives;
- recognise, safeguard, and promote olfactory heritage;
- to show that critically engaging our sense of smell & olfactory heritage is viable means for connecting and promoting Europe's tangible and intangible cultural heritage



Super interdisciplinary



Reading Smell



Reading Smell

- Objective: developing language technology to identify references to smells in (historical) texts;
- Challenges:
 - finding smell references;
 - language change.
- Annotation guidelines, benchmark dataset for English, Italian, French, Dutch, German, Slovene & Neo-Latin, NLP models for extracting smell references from text, >62,000 texts processed and searchable via the European Olfactory Knowledge Graph
- Stefano Menini, Teresa Paccosi, Sara Tonelli, Marieke Van Erp, Inger Leemans, Pasquale Lisena, Raphael Troncy, William Tullett, Ali Hürriyetoğlu, Ger Dijkstra, Femke Gordijn, Elias Jürgens, Josephine Koopman, Aron Ouwerkerk, Sanne Steen, Inna Novalija, Janez Brank, Dunja Mladenec, and Anja Zidar. 2022. A multilingual benchmark to capture olfactory situations over time. In Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change, pages 1–10, Dublin, Ireland. Association for Computational Linguistics

satonelli: Odeuropa Benchmark EN/021E Thomas Tomkis 1607 Lingua or the battle of the senses__26_THE.txt
1-25 / 61 sentences [doc 21 / 84]

Smell_Word

16 ODOR .

17 I do my Lord, and haue at my command, The smell of flowers , and Odoriferous drugs,

Of oyntments sweete, and excellent perfumes, And Court-like waters, which if once you smell ,

You in your heart would wish as I suppose: That all your Body were transsformed to Nose.

18 PHA.

19 Olfactus of all the Senses, your obiects haue the worst luck, they are alwaies iarring with their contraries, for none can weare Ciuet , but they are suspected of a proper badde sent , where the prouerbe springs, hee smelleth best , that doth of nothing smell .

20 ACT. 4.

21 SCENA 4.

22 The bench and Olfactus as before, Tobacco apparelled in a taffata mantle, his armes browne and naked, buskins made of the pilling of Osiers, his necke bare, hung with Indian leanes, his face browne painted with blewe stripes, in his nose swines teeth, on his head a painted wicker crowne, with Tobacco pipes set in it, lumes of Tobacco leaues, lead by two Indian boyes naked, with tapers in their hands, Tobacco boxes and pipes lighted.

23 PHA.

24 Foh, foh, what a smell is heare?

25 is this one of your delightfull obiects?

Layer Smell

Create a Relation m
arc between annota

Annotation

Text

smell

Relation

of flowers

Relation

Odoriferous dr

Relation

Of oyntments s

Relation

excellent perfu

Relation

Court-like wat

Relation

I

Frame Element

Smell_Word

Reading Smell

	Dutch	English	French	German	Italian	Slovenian	Latin
Smell Word	1,788	1,530	845	2,659	1,254	1,973	1,374
Smell Source	1,922	1,313	710	2,297	952	1,638	1,167
Quality	1,071	1,084	450	1,730	707	936	790
Perceiver	336	362	140	399	153	266	232
Circumstances	399	248	88	274	202	228	172
Odor Carrier	351	310	106	170	195	408	125
Effect	243	187	53	425	104	214	224
Evoked Odorant	228	91	103	258	74	285	92
Place	255	302	172	200	158	394	182
Time	127	126	49	131	119	75	87
Creator	30	0	5	1	0	1	0

Overview of annotated instances of lexical units (Smell Words) and frame elements in the benchmark for each language

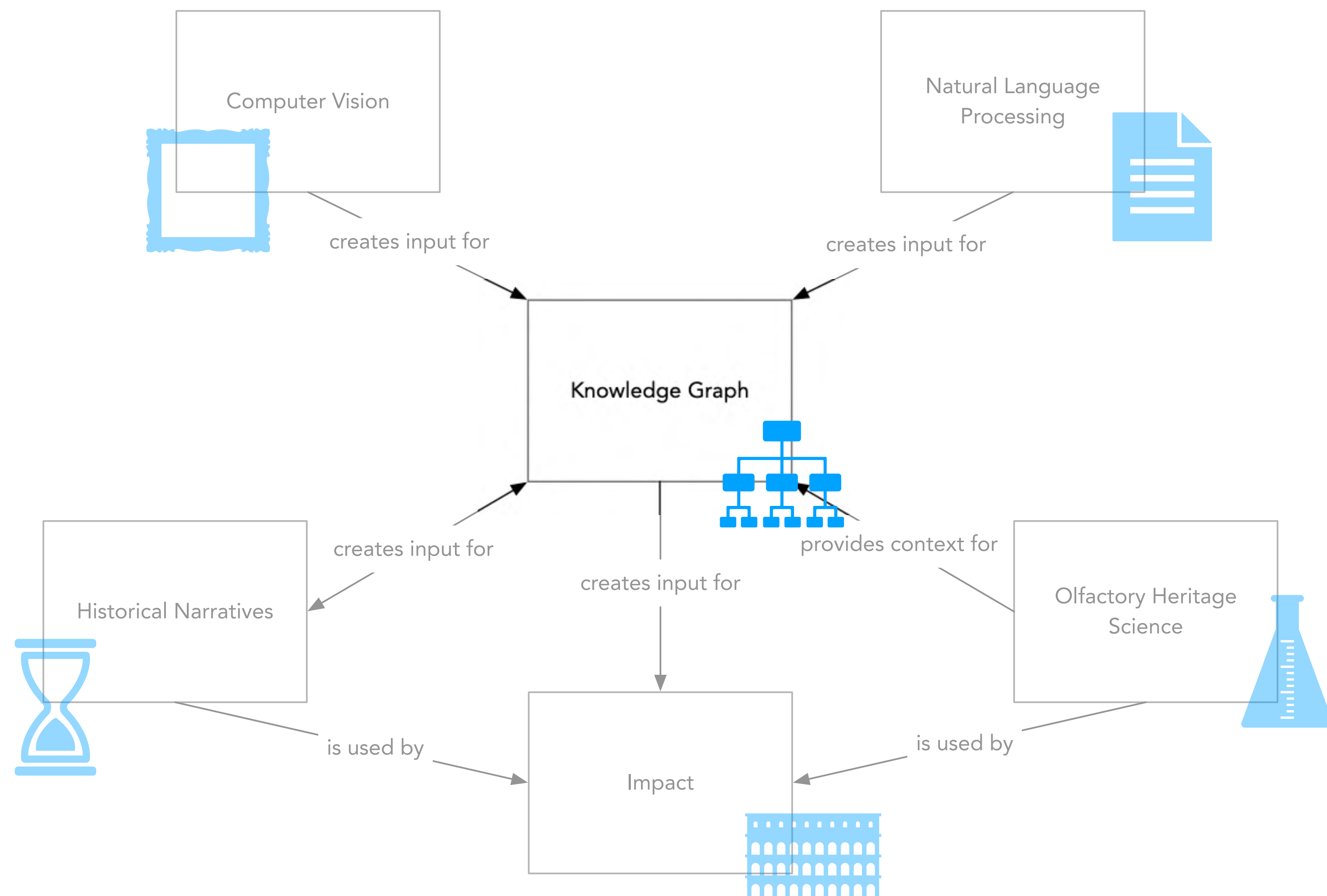
Classification Setups

- Single task vs multi-task & mono-lingual vs multi-lingual
- Contemporary models:
 - EN: bert-base-cased, IT: bert-base-italian-cased, NL: bert-base-dutch-cased, FR: flaubert_case_based, SL: sloberta, DE: bert-base-german-cased, LA: simple-latin-bert-uncased, Multilingual: bert-base-multilingual-cased (mBert)
- Historical models:
 - EN: MacBERTH, IT: BERToldo, NL: GysBERT, FR: D'alemBERT, DE: bert-base-historical-german-rw-cased
 - SL: low-resource language, historical processing via spelling normalisation
- 10-fold cross-validation

Lan.	App.	Train Data	Smell Word	Smell Source	Quality	Odour Carrier	Evoked Odorant	Loc.	Perc.	Time	Circ.	Effect
EN	MT	mono multi	0.871 0.865	0.571 0.574	0.758 0.759	0.482 0.462	0.572 0.517	0.542 0.546	0.510 0.488	0.434 0.528	0.461 0.480	0.405 0.339
	ST	mono multi	0.867 0.881	0.525 0.530	0.703 0.698	0.392 0.392	0.293 0.359	0.368 0.410	0.410 0.390	0.304 0.309	0.266 0.261	0.140 0.138
IT	MT	mono multi	0.871 0.887	0.559 0.575	0.800 0.801	0.343 0.382	0.564 0.625	0.439 0.304	0.241 0.240	0.613 0.642	0.259 0.309	0.246 0.201
	ST	mono multi	0.854 0.880	0.387 0.407	0.739 0.755	0.249 0.269	0.383 0.299	0.193 0.186	0.254 0.231	0.461 0.398	0.148 0.201	0.141 0.149
FR	MT	mono multi	0.839 0.838	0.459 0.472	0.567 0.591	0.440 0.379	0.536 0.505	0.373 0.322	0.380 0.258	0.481 0.561	0.279 0.306	0.235 0.273
	ST	mono multi	0.734 0.820	0.314 0.417	0.336 0.488	0.327 0.352	0.414 0.367	0.302 0.251	0.291 0.268	0.384 0.336	0.118 0.111	0.142 0.150
NL	MT	mono multi	0.788 0.789	0.376 0.407	0.632 0.638	0.191 0.214	0.444 0.468	0.238 0.236	0.303 0.308	0.313 0.342	0.133 0.154	0.149 0.192
	ST	mono multi	0.725 0.765	0.225 0.235	0.545 0.556	0.041 0.072	0.091 0.063	0.068 0.082	0.104 0.064	0.096 0.071	0.053 0.030	0.045 0.039
DE	MT	mono multi	0.812 0.814	0.454 0.470	0.668 0.677	0.157 0.215	0.454 0.490	0.308 0.293	0.358 0.351	0.184 0.255	0.300 0.273	0.241 0.253
	ST	mono multi	0.778 0.797	0.273 0.268	0.479 0.443	0.186 0.141	0.150 0.086	0.092 0.092	0.164 0.133	0.162 0.095	0.040 0.031	0.036 0.030
SL	MT	mono multi	0.707 0.695	0.501 0.442	0.525 0.491	0.320 0.273	0.506 0.445	0.401 0.368	0.355 0.245	0.280 0.214	0.153 0.103	0.151 0.132
	ST	mono multi	0.675 0.655	0.406 0.358	0.451 0.448	0.119 0.186	0.277 0.263	0.236 0.212	0.170 0.195	0.155 0.137	0.068 0.051	0.074 0.086
LA	MT	mono	0.710	0.220	0.359	0.052	0.142	0.081	0.233	0.066	0.006	0.043
	ST	mono multi	0.213 0.351	0.000 0.000	0.000 0.000	0.000 0.000	0.000 0.000	0.000 0.000	0.000 0.000	0.000 0.000	0.000 0.000	0.000 0.000

Results (F1) of the classifiers on the lexical unit and 9 frame elements for both Single Task (ST) and Multi-task (MT) approaches. Each result is the average of 10 different runs done on 10 different data splits). mono = monolingual data and model; multi = multilingual training data and model. Results contemporary models shown.

Organising Smell





Smell of Cadaver

RELATED
[Carcass](#) · [Body](#) · [Anatomical Lessons](#)



Smell Quality

putrides fragrant peculiar fétides stinking repoussante odoriferous
offensive putride foul scented infecte fetid insupportable
mauvaise odorous fétide olfactory characteristic putrid

Emotion

indifference despair envy gratitude greed desire guilt
disappointment embarrassment excitement courage nostalgia
doubt relief nervousness curiosity frustration faith calmness
pain

Fragrant Space

from a great distance maison on the castle walls On the mountain
cimetières even outside the house in the room cimetière
inside the house places in such distances in this cavity
from the ruins the staircase leading to them quartier ville
in the vicinity of Nuremberg lieu along the river
chambre

In texts (3844 occurrences)

FETORS	STENCH	STENCH	ODOURS	SMELL
AUTHOR John Floyer TITLE Pharmako Basnos CREATION DATE 1687 SMELL SOURCE Cadaver OLFACTORY EXPERIENCE en	AUTHOR John GILLIES TITLE history of Ancient Greece ... P... CREATION DATE 1825 SMELL SOURCE Cadaver OLFACTORY EXPERIENCE despair, frustration en	AUTHOR John GILLIES TITLE history of Ancient Greece ... P... CREATION DATE 1825 SMELL SOURCE Cadaver OLFACTORY EXPERIENCE en	AUTHOR Francis Bacon TITLE Works of Francis Bacon ... A n... CREATION DATE 1825 SMELL SOURCE Musk, Excrement, Cadaver, Civet OLFACTORY EXPERIENCE surprise	AUTHOR William Lisle BOWLES TITLE Grave of the last Saxon; or, th... CREATION DATE 1822 SMELL SOURCE Cadaver, Blood OLFACTORY EXPERIENCE en

CLEAR

In text

In images

Both

Full text search

Search in

Title only

Content only

Title and content

Smell Source

Select...

Cadaver

Carrier

Select...

Quality

Select...

Emotion

Select...

Language

Select...

fr

Time

Start year (YYYY)

End year (YYYY)

Time period

Season

Place

Select...

In text

Smell Source

Cadaver

Language

French

2058 search results

Show as

cards

Sort by

Select...

Page 1

ODEUR

TITLE

Comptes-rendus du Xlle Co...

CREATION DATE

1898/1900

SMELL SOURCE

Cadaver

PLACE

profonde du champ opératoire

OLFACTORY EXPERIENCE

EMOTION

surprise

LANGUAGE

fr

SEE MORE

EMBAUMÉS

AUTHOR

Hadrian Ségoillot

TITLE

Lettres sur l'Espagne / par ...

CREATION DATE

1870

SMELL SOURCE

Cadaver

OLFACTORY EXPERIENCE

EMOTION

boredom

LANGUAGE

fr

SEE MORE

ODEUR

AUTHOR

Raphaël Blanchard

TITLE

Traité de zoologie médical...

CREATION DATE

1890

SMELL SOURCE

Cadaver

CARRIER

viande | bouche | narines

OLFACTORY EXPERIENCE

EMOTION

calmness

LANGUAGE

fr

SEE MORE

GOÛT

AUTHOR

Georges Curtel

TITLE

Traité pratique de vinificati...

CREATION DATE

1899

SMELL SOURCE

Wine, Cadaver, Material

OLFACTORY EXPERIENCE

DEFINED AS

souvent très prononcé

LANGUAGE

fr

SEE MORE

PUANT, ODEUR, CHAROGNE...

AUTHOR

Fernand Milton

TITLE

Les Femmes et l'adultère, d...

CREATION DATE

1910

SMELL SOURCE

Cadaver

PLACE

Dalmatia

OLFACTORY EXPERIENCE

DEFINED AS

insupportable

LANGUAGE

fr

SEE MORE

ODEUR

TITLE

Le Néo-naturien : revue des...

CREATION DATE

1927-08-01

SMELL SOURCE

Cadaver

OLFACTORY EXPERIENCE

LANGUAGE

fr

SEE MORE

ODEUR, ÉMANATIONS

AUTHOR

André Giraud

TITLE

Précis historique et guériso...

CREATION DATE

1855

SMELL SOURCE

Cadaver

OLFACTORY EXPERIENCE

DEFINED AS

délétères

LANGUAGE

fr

SEE MORE

ODEUR

AUTHOR

Santé publique à Marseille

TITLE

Bulletin de la Santé publiqu...

CREATION DATE

1884-07-08

SMELL SOURCE

Smoke, Cadaver

OLFACTORY EXPERIENCE

EMOTION

disgust

LANGUAGE

fr

SEE MORE

ODORAT, EXHALAISONS

AUTHOR

Louis-Céran Lemonnier

TITLE

Promenades au Jardin des ...

CREATION DATE

1837

SMELL SOURCE

Cadaver

OLFACTORY EXPERIENCE

LANGUAGE

fr

SEE MORE

ODEUR

AUTHOR

Antoine Muron

TITLE

Pathogénie de l'infiltration ...

CREATION DATE

1872

SMELL SOURCE

Urine, Cadaver

OLFACTORY EXPERIENCE

LANGUAGE

fr

SEE MORE

ODEUR

AUTHOR

Robert Cambier

TITLE

Traité de bactériologie pur...

CREATION DATE

1902

SMELL SOURCE

Cadaver

OLFACTORY EXPERIENCE

LANGUAGE

fr

SEE MORE

ÉMANATIONS

AUTHOR

Robert Cambier

TITLE

Traité de bactériologie pur...

CREATION DATE

1902

SMELL SOURCE

Excrement, Cadaver, Water

OLFACTORY EXPERIENCE

LANGUAGE

fr

SEE MORE

Previous

1

2

3

4

5

...

102

103

Next

53

Code: <https://github.com/Odeuropa>

<https://bit.ly/odeuropa-images>

<https://bit.ly/odeuropa-text>

The screenshot shows a Jupyter Notebook environment. The top bar indicates the notebook is titled "Image Nosebook: Tracing Visual Smell Trends" and is not saved. The left sidebar contains navigation icons for code, text, and drive. The main notebook area displays a welcome message and a list of tasks. A bottom panel shows options for "Code and Data F", "Import", "Explore", and "SPARQL". A right panel shows a SPARQL query editor with a query starting with "PREFIX od: <http".

Image Nosebook: Tracing Visual Smell Trends

Odereuropa Image Data Teaser

Welcome to this Jupyter Notebook designed for the exploration of the Odeuropa Project Image Data.

1. Code and data preparation. This part can mostly be ignored. Just run it once by clicking the left of the cells).
2. Query Image Data. This section enables you to explore our image data by querying for the image meta data.
3. Quantitative Analysis. In this section, you will see an exemplarily walkthrough of how the statistical insights.

Each code cell is collapsed by default. If you are working on a local copy of the notebook, you can click the left of the cells to expand them.

Code and Data F

Run the following cells to prepare the data. In the meantime, we suggest you grab a quick overview of the data by clicking the triangle within the circle.

Import

Explore

SPARQL

SPARQL Query

```

PREFIX od: <http://data.odeuropa.org/ontology/od#>
PREFIX crm: <http://www.eurovoc.europa.org/crm#>

select * where {
  ?emission od:emission
  ?emission od:emission

```

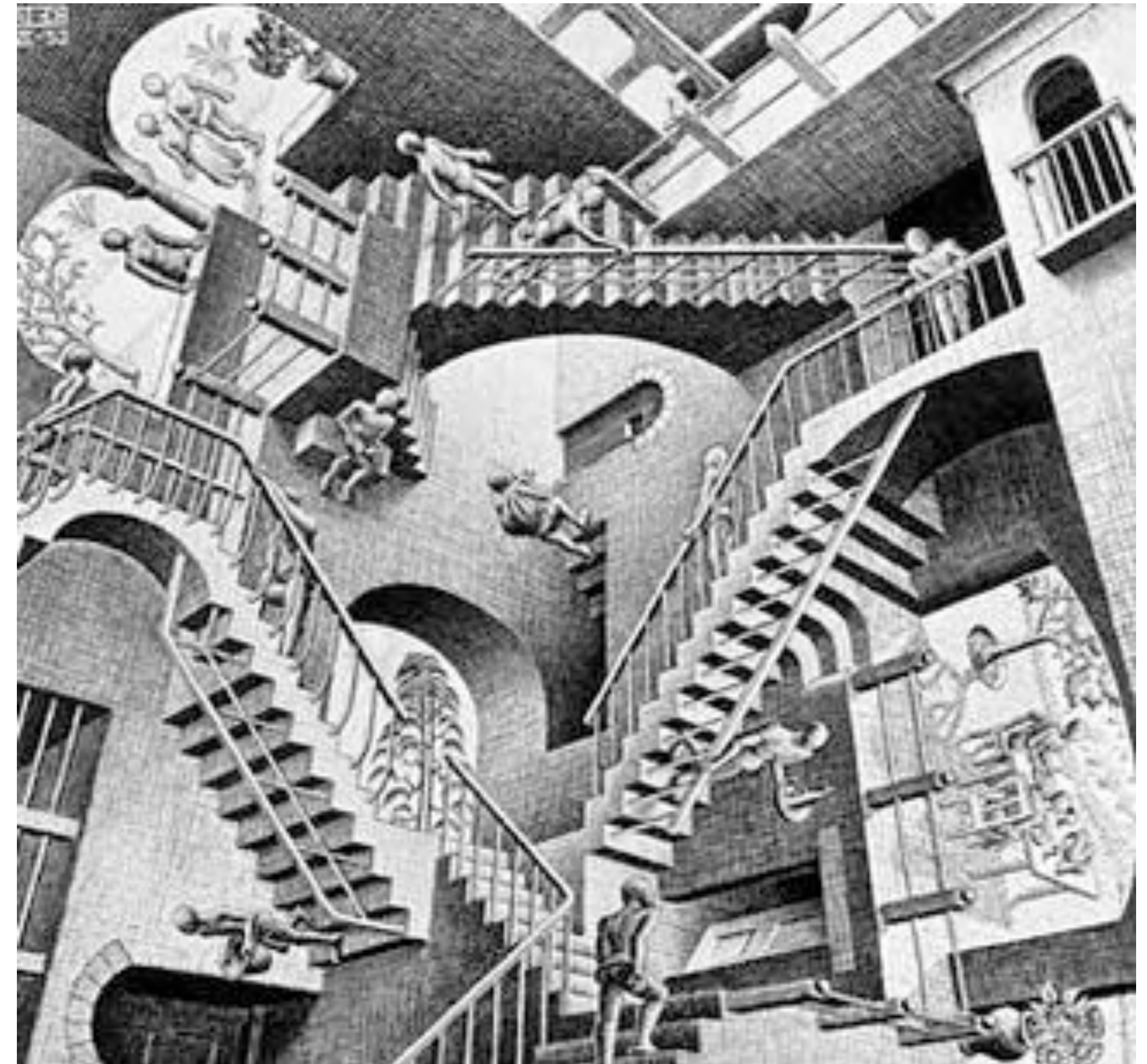
The screenshot shows the Google Colab interface for a notebook titled "odeuropa-notebook.ipynb". The top menu bar includes "File", "Edit", "View", "Insert", "Runtime", "Tools", "Help", and a status message "Changes will not be saved". Below the menu, there are tabs for "+ Code", "+ Text", and "Copy to Drive". On the left sidebar, there are icons for a search, a code editor, a key, and a folder. The main content area displays the notebook's title "Odeuropa Nosebook – Textual Smell Trend Analyser (WP3)" in large black font. Below the title are two logos: the "Odeuropa" logo, which is a stylized purple and green swirl, and the "FONDAZIONE BRUNO KESSLER" logo, which is a blue stylized "F" and "K". The text "Through this Notebook you will be able to:" is followed by a list of four tasks: 1. Select a dataset your preferred language (English, Italian, German, Dutch, French, Slovene) - Section 1. 2. See the time distribution of a specific Smell Source within the selected dataset - Section 2. 3. See the frequency of a specific Smell Source within your dataset - Section 3. 4. See the frequency of a specific Smell Source within your selected dataset - Section 4. 5. See the frequency of a specific Smell Source within your selected dataset - Section 5. At the bottom of the notebook content area, there are three buttons: "Editor only" (blue), "Editor and results" (orange), and "Results" (blue). The bottom of the Colab interface shows a search bar, a dropdown menu with "odeuropa" selected, a "Login" button, and a "Results" button.

How close to "10 cells hidden". You don't need to c

Outlook*

Outlook

- LLMs are a game changer
- ...but not a magic wand
- Domain knowledge for trustworthiness
- Combining structured & unstructured data



Relativity, by M. C. Escher. Lithograph, 1953.

Outlook

- Dutch East India Company
 - Commodities (DHBenelux 2024)
 - Semantic change in ethnic categories & loan words (talk to Jiaqi)
 - Spices, DHBenelux 2025!
- Historical Recipes
 - Under review (but talk to Gauri & Teresa)
- Modelling Change
 - Under review (but talk to Jiaqi and me)
- Smell in Fan Fiction
 - With Julia Neugarten, DHBenelux 2025



Image generated by deepai.org prompt: cauldron with stew, colonial spices and Zeus and Athena on the side

More Info

dhlab.nl

cultural-ai.nl

odeuropa.eu

trifecta.dhlab.nl

Thank you



Odeuropa and TRIFECTA have received funding from the European Union's Horizon 2020 and Horizon Europe research and innovation programmes under grant agreements 101004469 and 10188548. Cultural AI has received funding from the Dutch Research Council (NWO), the Dutch Digital Heritage Network (NDE) and Europeana. Communicating the environmental impact of plant based recipes has received funding from the Alpro Foundation

