Die Werkbänke der Digital Humanities

Zur Rolle von Tools und Software für die Forschungsarbeit

Frank Fischer[1 3] Manuel Burghardt[2] Jan Luhmann[2]
Laure Barbot[3] Yoann Moranville[3] Alireza Zarei[4]

[1]Higher School of Economics, Moskau

[2]Universität Leipzig

[3]DARIAH-EU

[4]GWDG Göttingen

vDHd2021, 26. März 2021, 14:00 – 15:00

https://vdhd2021.hypotheses.org/135

Ablauf

1. Dauer des Panels: 60 Minuten.
2. Vorstellungsrunde und Präsentation in 2 Teilen à 20 Minuten.
3. Fragen können im Chat gesammelt werden.
4. Danach 20 Minuten moderierte Diskussionsrunde.
5. Folien werden auf Zenodo veröffentlicht.

“Reading Tools” (DH2007)

G. Rockwell: Why do we care that our tools are used?

S. Sinclair: Because that is what we do as academics – we care for others in different ways through teaching and through sharing our research. Would we publish if we didn't think what we said would be a welcome gesture that someone else might respond to?

“Reading Tools – Dialogue with Stéfan Sinclair and Geoffrey Rockwell” (DH2007, University of Illinois, Urbana-Champaign)

doi:10.7939/r3-9vkj-k220

Warum überhaupt Tool-Monitoring?

1. Praxeologische Aspekte.
2. Häufig genutzte Tools können bei der Finanzierung nachhaltiger Lösungen priorisiert werden.
3. Häufig genutzte Tools können für Trainingseinheiten bevorzugt werden (etwa auf Plattformen wie Programming Historian, #dariahTeach oder forTEXT).
4. Ein umfassender Tool-Überblick ermöglicht auch die Analyse von Lücken in der Toolchain der DH.

Rückblick: Tool Directories

Einerseits: “Directory Paradox” (Quinn Dombrowski)

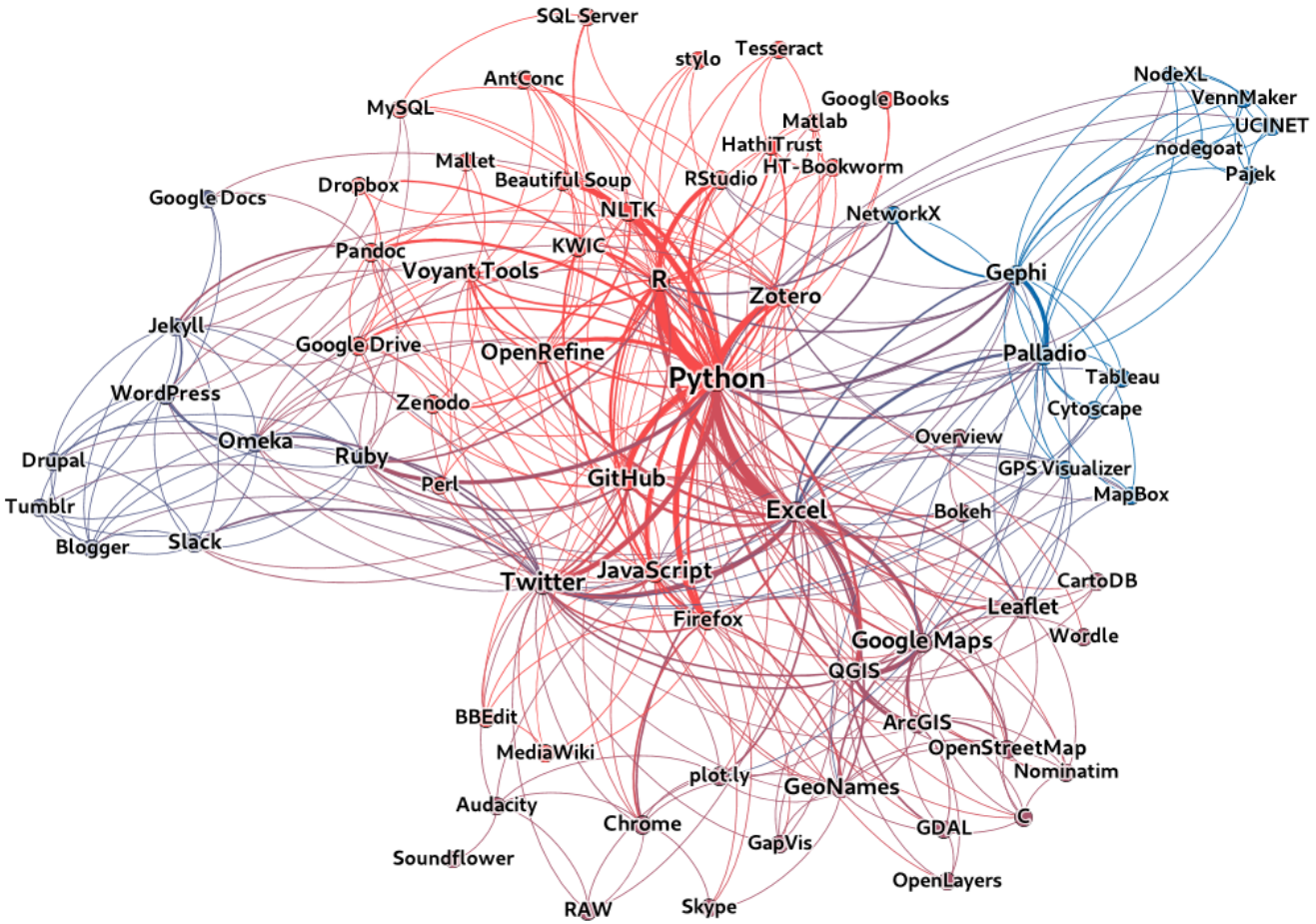
Andererseits:

- DiRT Directory
- TERESAH
- TAPoR
- DARIAH-DE Dienste und Werkzeuge
- SSH Open Marketplace

ToolXtractor

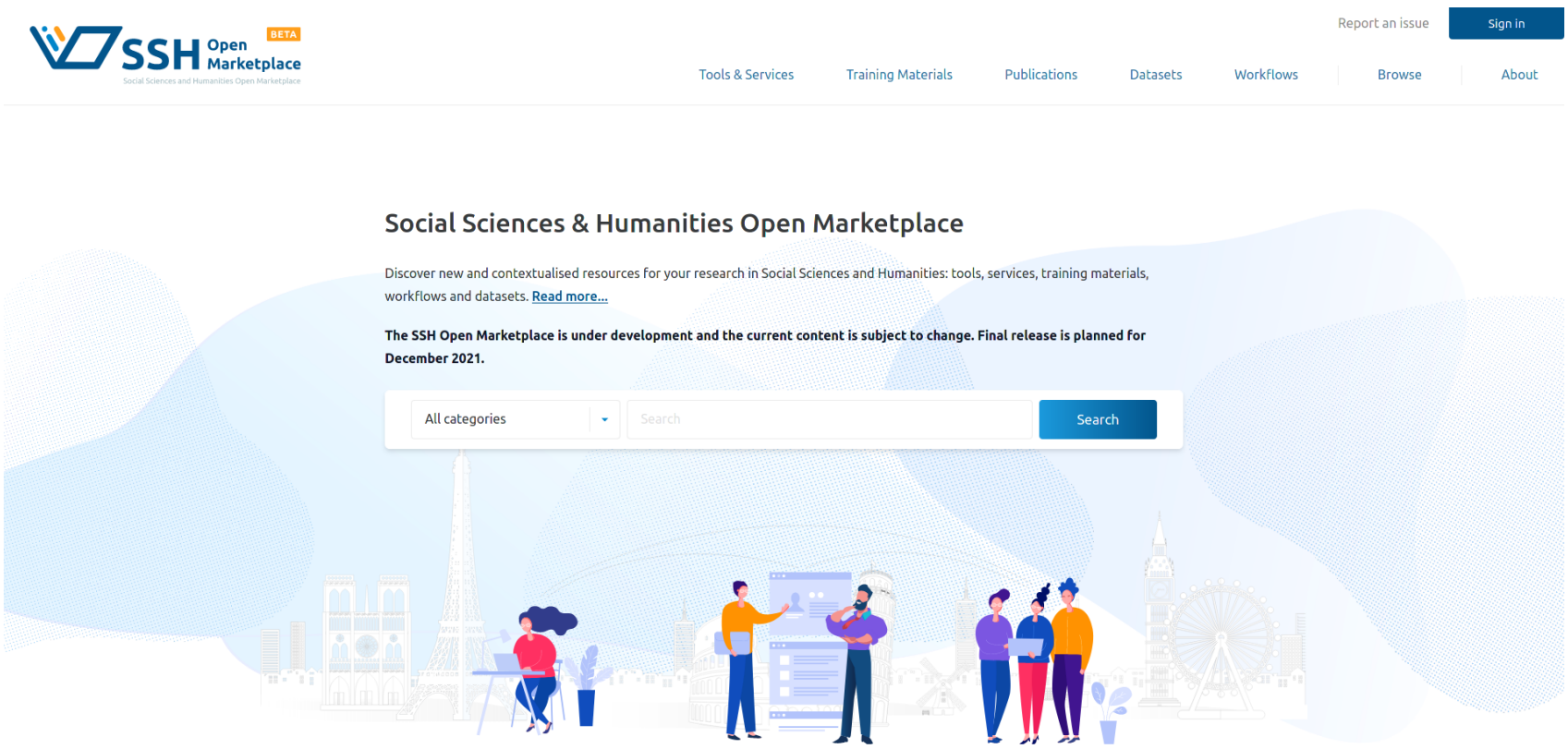
- Kommandozeilentool (geschrieben in Java)
- Ansatz: String Matching, Positivliste via TAPoR
- erste quantitative Analysen
- am häufigsten erwähnte Tools in den Abstracts der Konferenzen DH2015 – DH2019: Gephi, Omeka, stylo, MALLET, Excel, D3.js, NLTK, WordPress, Drupal, TextGrid, CollateX, GeoNames, TXM, Apache Solr, Voyant Tools
- Blogpost (2019): `https://weltliteratur.net/dh-tools-used-in-research/`

Netzwerkvisualisierung von Tool-Okkurrenzen in Lerneinheiten des “Programming Historian”



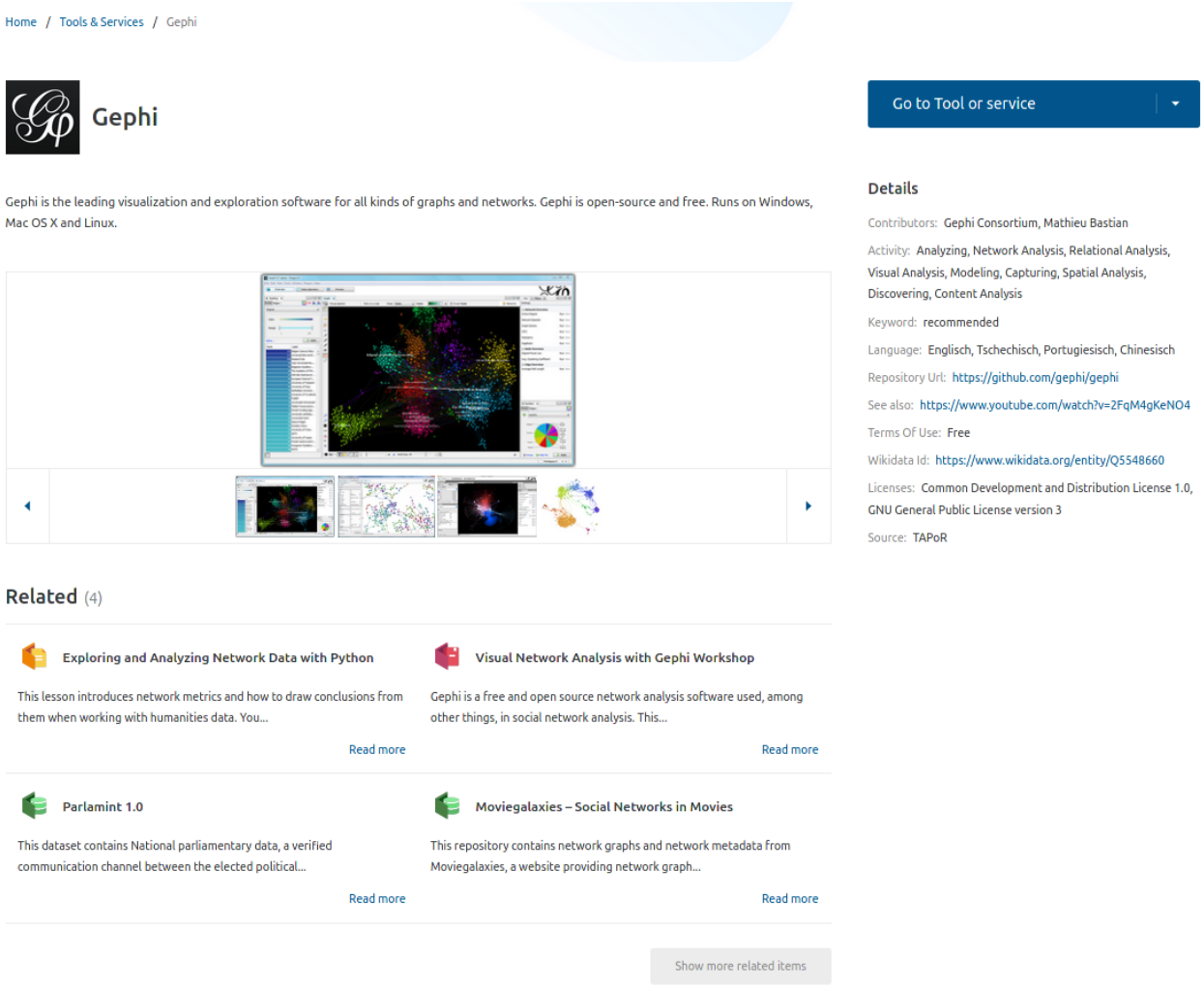
Frank Fischer, Yoann Moranville (2020):
`https://weltliteratur.net/dh-tools-programming-historian/`

SSH Open Marketplace



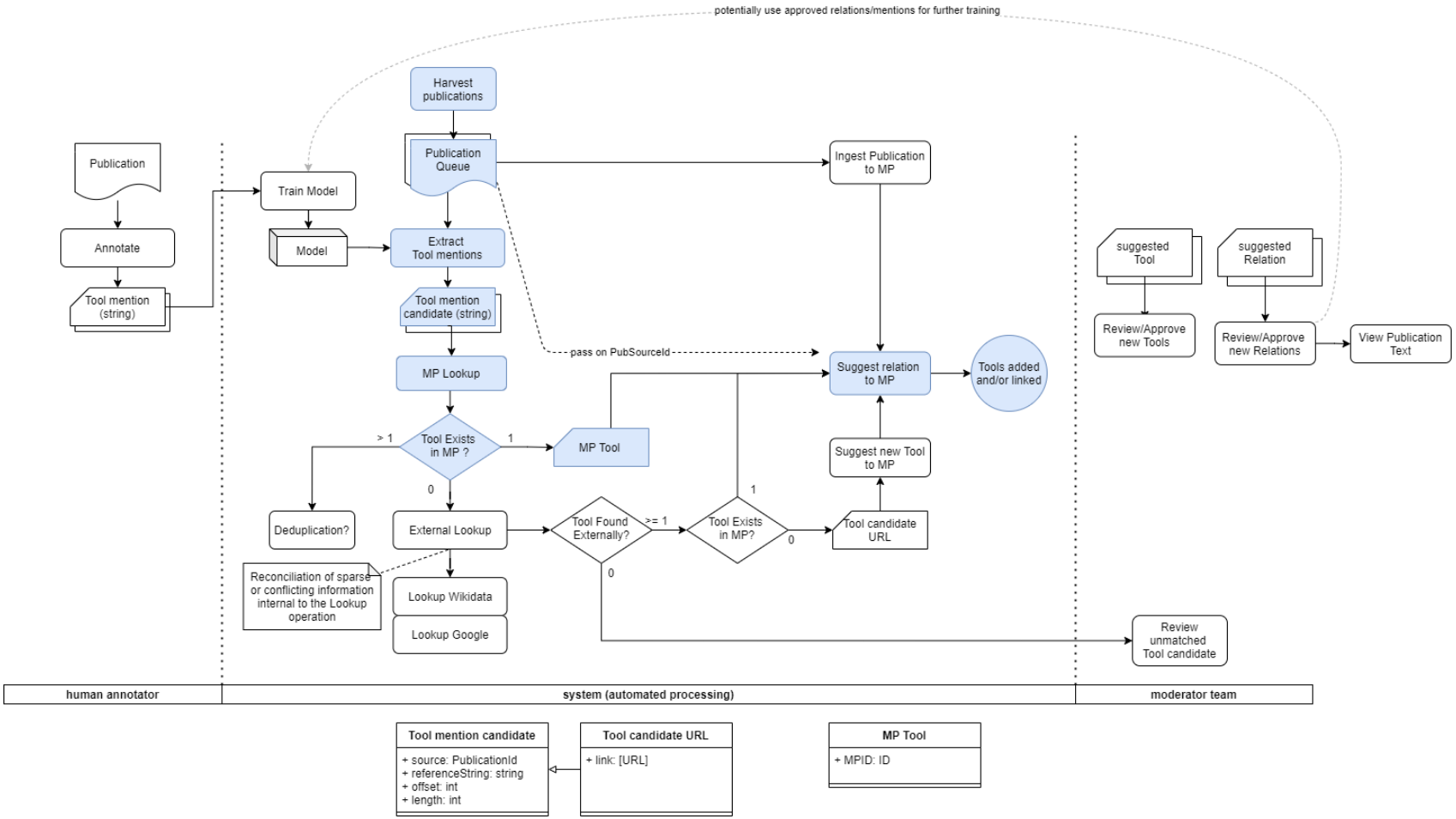
`https://marketplace.sshopencloud.eu/` (Beta-Version)

SSH Open Marketplace



Tooleintrag zu ‘Gephi’ inkl. Metadaten, Bildmaterial und Kontextualisierung
(= ähnliche Tools, wiss. Publikationen, Trainingseinheiten, Datensets, Workflows)

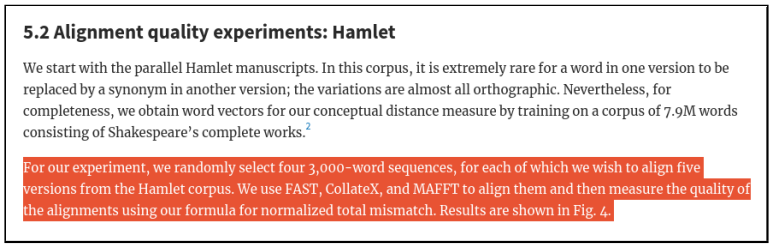
Tool Extraction Workflow für den Marketplace



Teil der Systemspezifikation (Dank an Matej Ďurčo).

Tool Extraction Demonstrator

Ausgangspunkt:



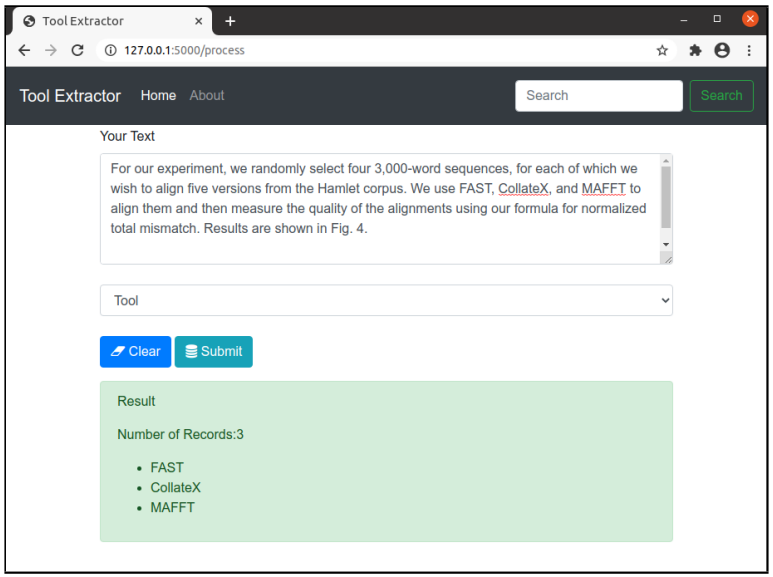
5.2 Alignment quality experiments: Hamlet

We start with the parallel Hamlet manuscripts. In this corpus, it is extremely rare for a word in one version to be replaced by a synonym in another version; the variations are almost all orthographic. Nevertheless, for completeness, we obtain word vectors for our conceptual distance measure by training on a corpus of 7.9M words consisting of Shakespeare's complete works.[2]

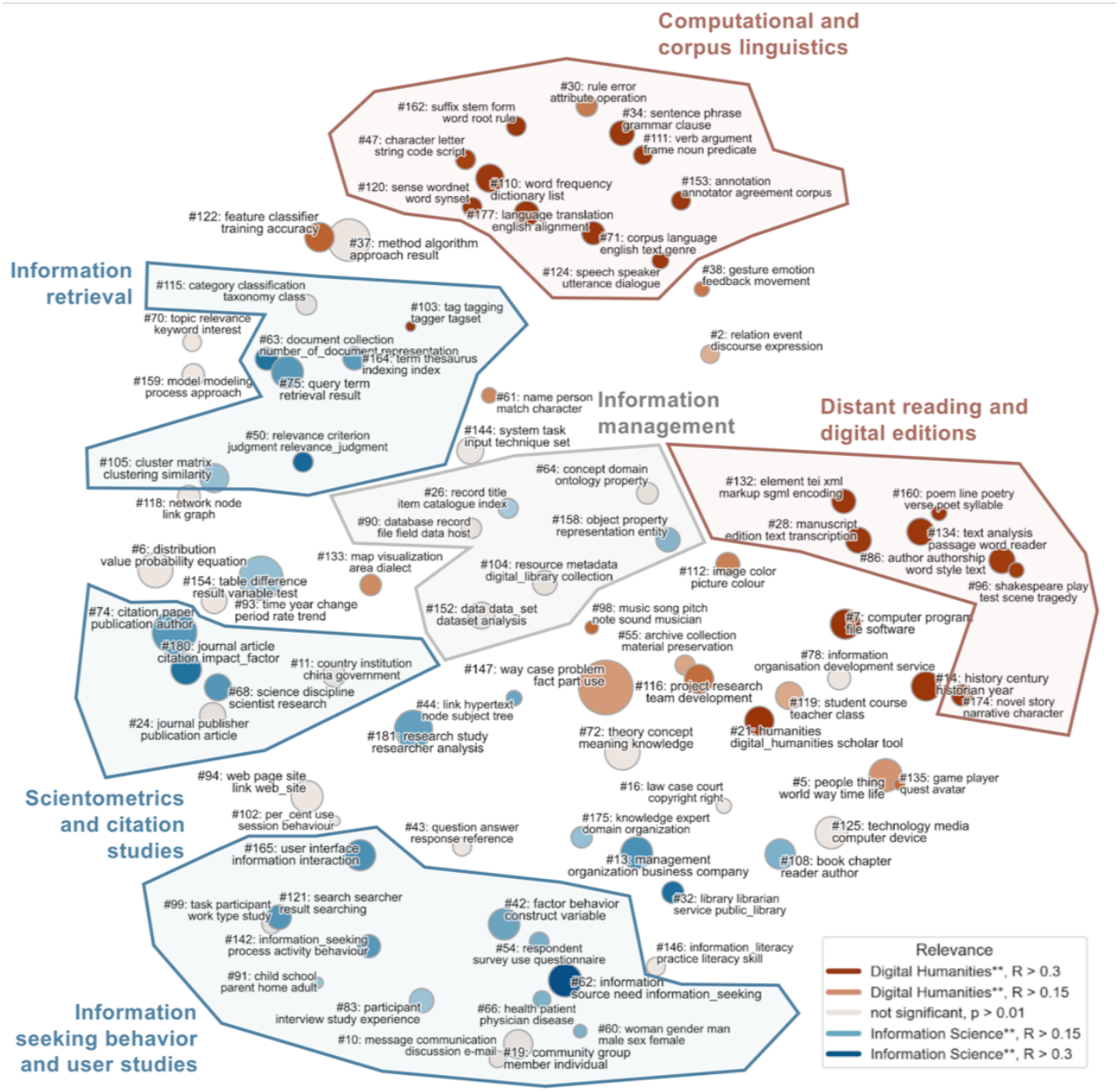
For our experiment, we randomly select four 3,000-word sequences, for each of which we wish to align five versions from the Hamlet corpus. We use FAST, CollateX, and MAFFT to align them and then measure the quality of the alignments using our formula for normalized total mismatch. Results are shown in Fig. 4.

Beispielartikel aus DSH: doi:10.1093/llc/fqz029.

Tool-Extraktion:



Meanwhile in Leipzig: Analyse von DH-Journalen



“unser Schreibzeug arbeitet mit
an unseren Gedanken”

– Friedrich Nietzsche (KSB 6, S. 172)

Korpus

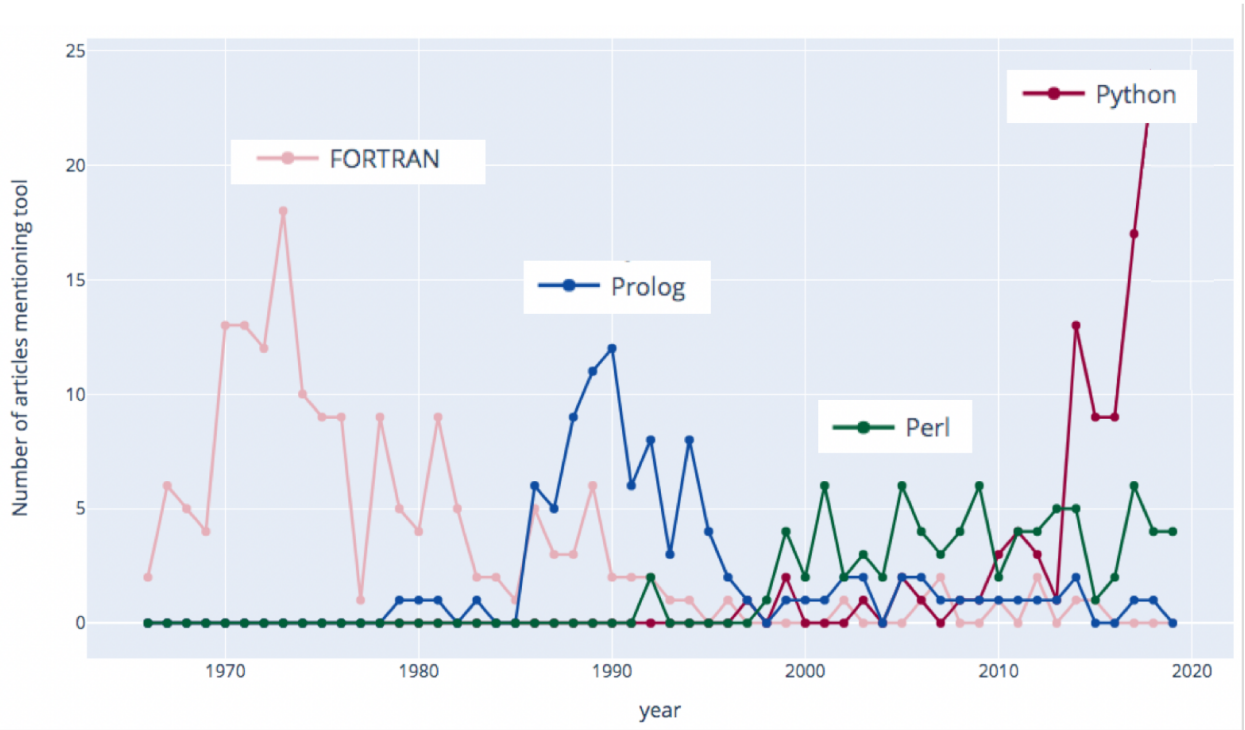
Journal	Zeitraum	Anzahl Artikel	Anzahl Tokens
Computers and the Humanities	1966 – 2004	1560	5,8 Mio.
Literary and Linguistic Computing	1986 – 2014	1454	5,9 Mio.
Digital Scholarship in the Humanities	2015 – 2019	305	1,8 Mio.
Digital Humanities Quarterly	2007 – 2019	418	2,9 Mio.
		3737	16,4 Mio.

Tool-Frequenzen

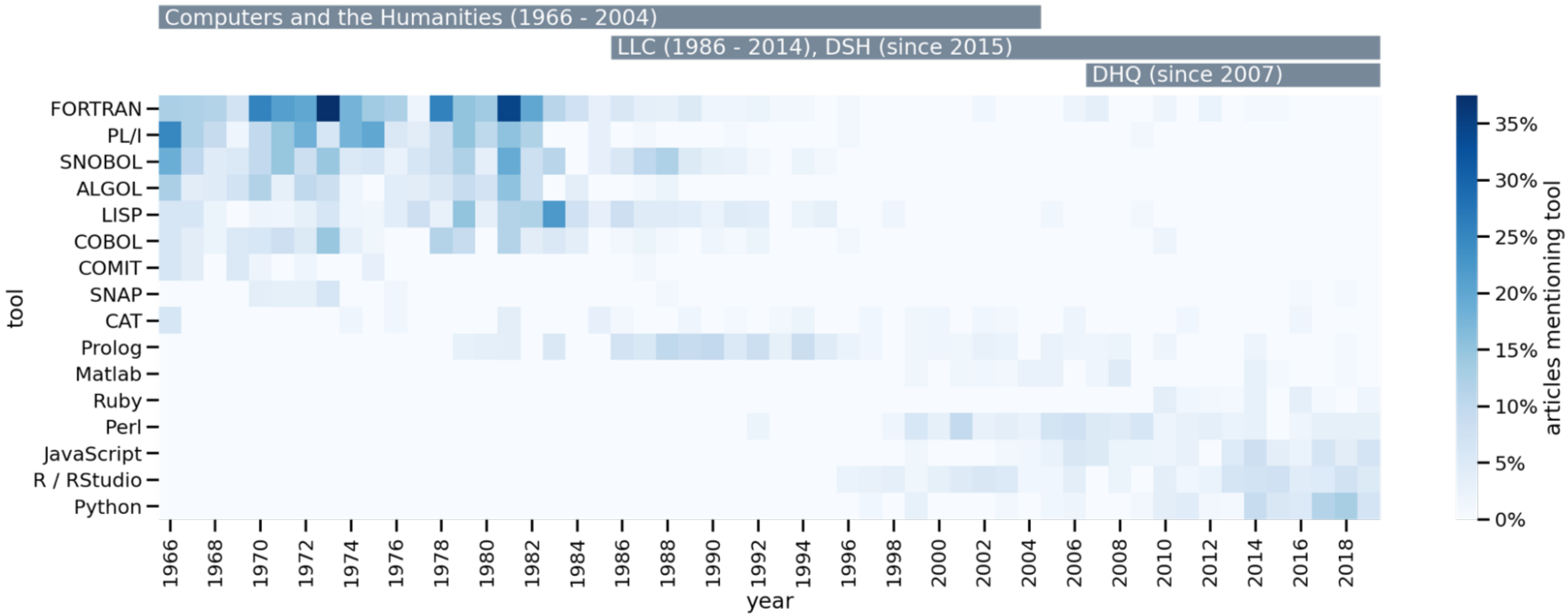
tool	number of articles
FORTRAN	174
R / RStudio	119
SNOBOL	111
Twitter	110
SPSS	96
Prolog	95
PL/I	95
OCP	88
LISP	81
Python	69
Perl	66
Excel	65
TACT	55
JavaScript	55
COCOA	54
COBOL	52
ALGOL	51
WordCruncher	44
SPITBOL	43
Nota Bene	39

Tool-Frequenzen

tool	number of articles
FORTRAN	174
R / RStudio	119
SNOBOL	111
Twitter	110
SPSS	96
Prolog	95
PL/I	95
OCP	88
LISP	81
Python	69
Perl	66
Excel	65
TACT	55
JavaScript	55
COCOA	54
COBOL	52
ALGOL	51
WordCruncher	44
SPITBOL	43
Nota Bene	39

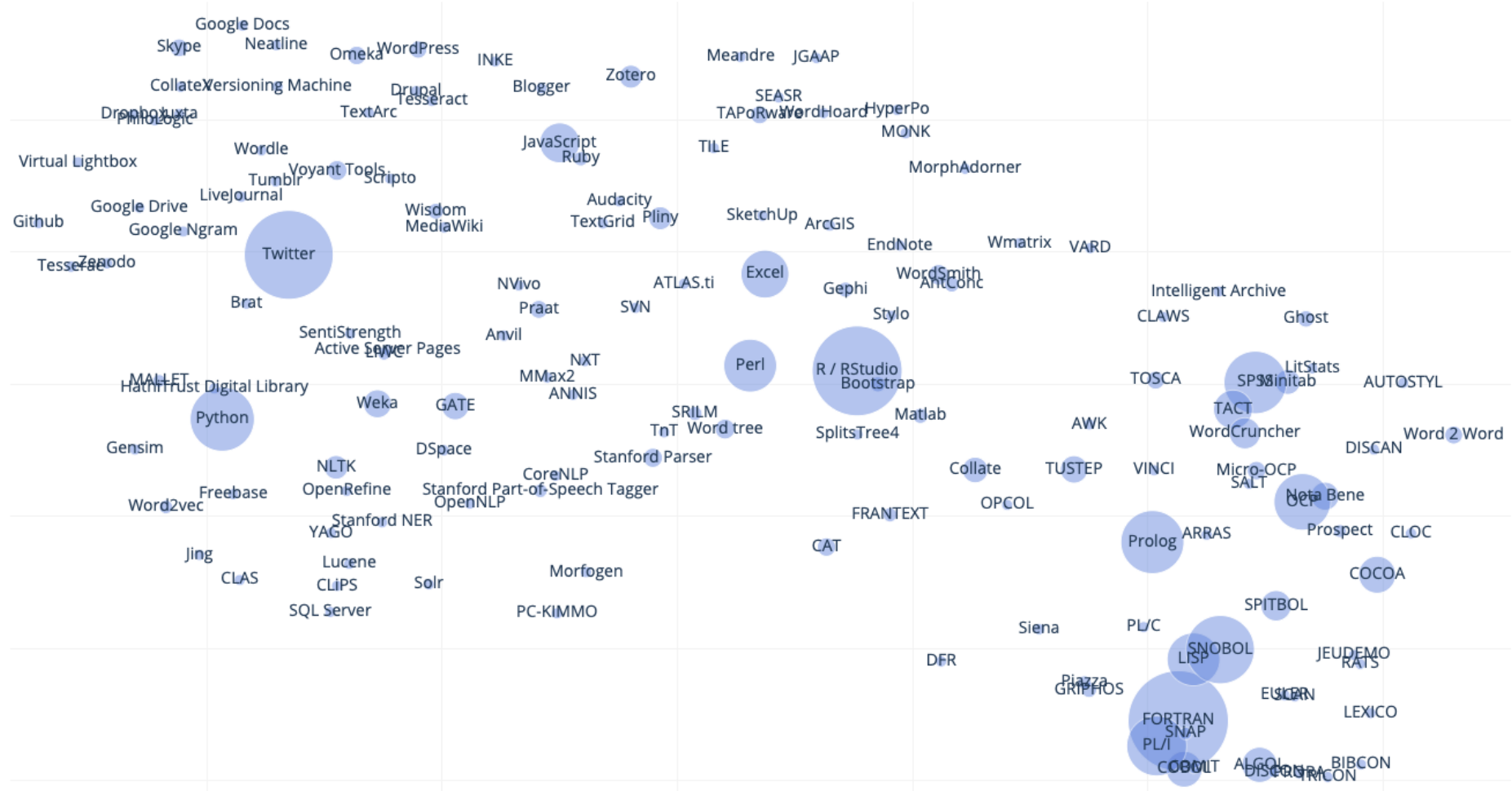


Programmiersprachen in den DH



Tool-Kookkurrenzen

Top 150 most-occurring tools:
UMAP 2D projection of nPMI co-occurrence scores // point size by number of occurrences



Tools & Topics

	Tools mit höchster Korrelation
Topic 13: "Network Analysis" network, node, tree, relationship, graph, edge, link, path, social_network, et_al	NETMET (0.95) GapVis (0.95) Cytoscape (0.87) Gephi (0.86) Prefuse (0.86)
Topic 20: "Image Annotation" image, color, object, colour, photograph, painting, resolution, picture, art, technique	ImageJ (0.98) GIMP (0.98) Viewshare (0.97) Virtual Lightbox (0.96) Annozilla (0.89)
Topic 52: "Document Markup" element, tei, document, xml, sgml, markup, tag, encoding, attribute, structure	TATOE (0.91) W3C (0.9) Graphviz (0.9) Oxygen XML Editor (0.88) T-PEN (0.86)

Rückbindung der Softwarenutzung in den DH an eine allgemeiner ausgerichtete “Tool Science”

1. *generic development tools*: Programmiersprachen, Libraries, Frameworks, IDEs etc.
2. *standard commercial off-the-shelf (COTS) software:* Microsoft Office, Abbyy OCR etc.
3. *software infrastructures*: VREs wie CLARIN oder TextGRID; aber auch SPSS oder Atlas.TI
4. *software specifically designed for specific research projects or research questions*: CATMA? AntConc? . . .

Vgl. Christian Wolff: The Case for Teaching “Tool Science” (2015) – doi:10.1109/EDUCON.2015.7096085

Nächste Schritte

- Verbesserung der Named Entity Recognition zur Erkennung von Software und Tools (jenseits der statischen TAPoR-Liste)
  - Trainingsdaten: String-Matching (TAPoR, ToolXtractor), manuelle Korrektur
- Weiterführende Perspektive: Vergleich mit anderen nahegelegenen Disziplinen (Informationswissenschaft, Computerlinguistik, Informatik etc.)
- Nachhaltige Integration der identifizierten Tools in den SSH Open Marketplace

Diskussion

- **Rückfragen**, **Anregungen**, **Kommentare**?