Integrating NetworKit into a web-based Environment for Network Analysis and Exploration

Jörn Kreutel, Beuth University of Applied Sciences Berlin
@NetworKit Day, October 15, 2020

Contents

- Background and Motivation
- Functionality
- Architecture
- Current State and Open Issues
- Demo

Background and Motivation

Domain Background and Motivation

- Research on **bio-bibliographical networks** as a **work in progress**

- **Partners**: Humboldt-University / German Literature Studies, Berlin-Brandenburg Academy of Sciences and Humanities

- Three relatively large bibliographical datasets
  - *Bibliographical Yearbooks*: bibliography of all **fictional literature publications** in SOZ+GDR (1945-1990),
    - appr. 50.000 entries

  - *Bibliography of German Literature and Language Studies* **and** *International Bibliography of German Literature and Literary Studies*: international bibliographies of **scientific publications**
    - appr. 320.000 - 350.000 entries

  - **Enhancement** of bibliographical data with **biographical information** on authors (e.g. date of birth, affiliation with institutions, places of living etc.), mainly taken from **GND authority files**

- Since October 2019, creation of a dataset of **detailed bibliographical data** on GDR authors (*Research platform „Literary field GDR“: authors, works, networks*, funded by DFG)

- For many aspects of the talk, see Kreutel 2019

Objective, Method, Tools

- Separate subjects of analysis, **common objectives** and analysis **methods**
  - main focus: **copublication networks** based on contributions to collections and anthologies
  - Maximal network sizes of about 15.000-20.000 nodes and 300.000-600.000 edges, based on 3000-10.000 collections as network constituting entities

- **Objective**: Insight into a discipline‘s / a field‘s historical evolution, main agents and groups of agents, publication types, relevance of institutions, patterns of careers, contributions of private/professional networks etc. (following, e.g., Bourdieu 1999; De Nooy 2003; Bottero and Crossley 2011)

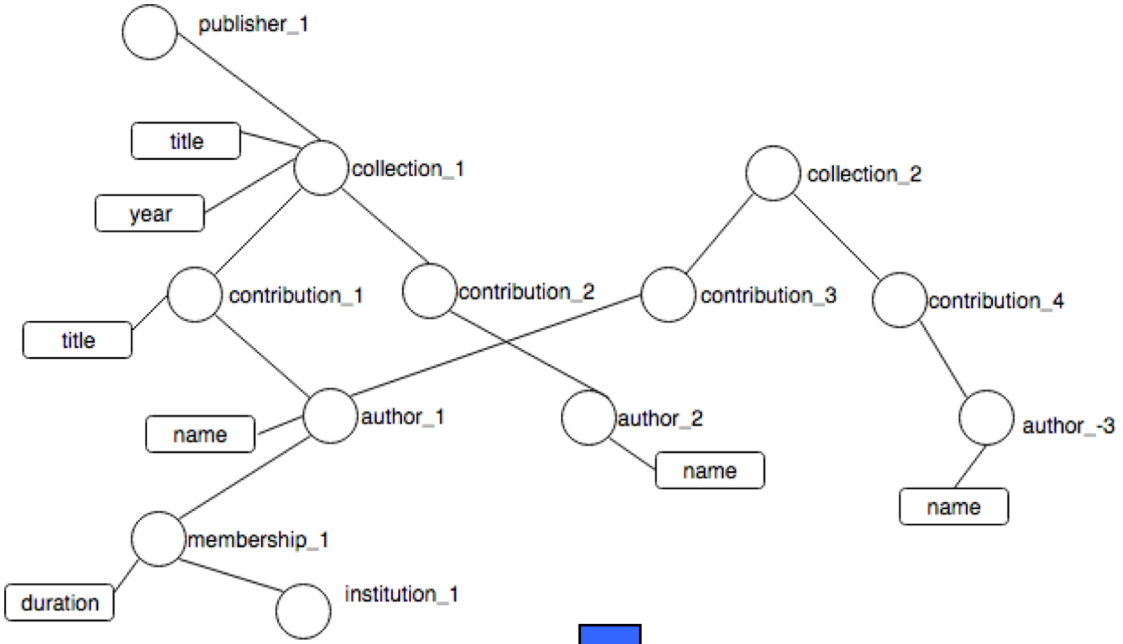
- **Method**: Social Network Analysis (SNA) + “conventional” Statistics

- What **tools** do we need for investigation?

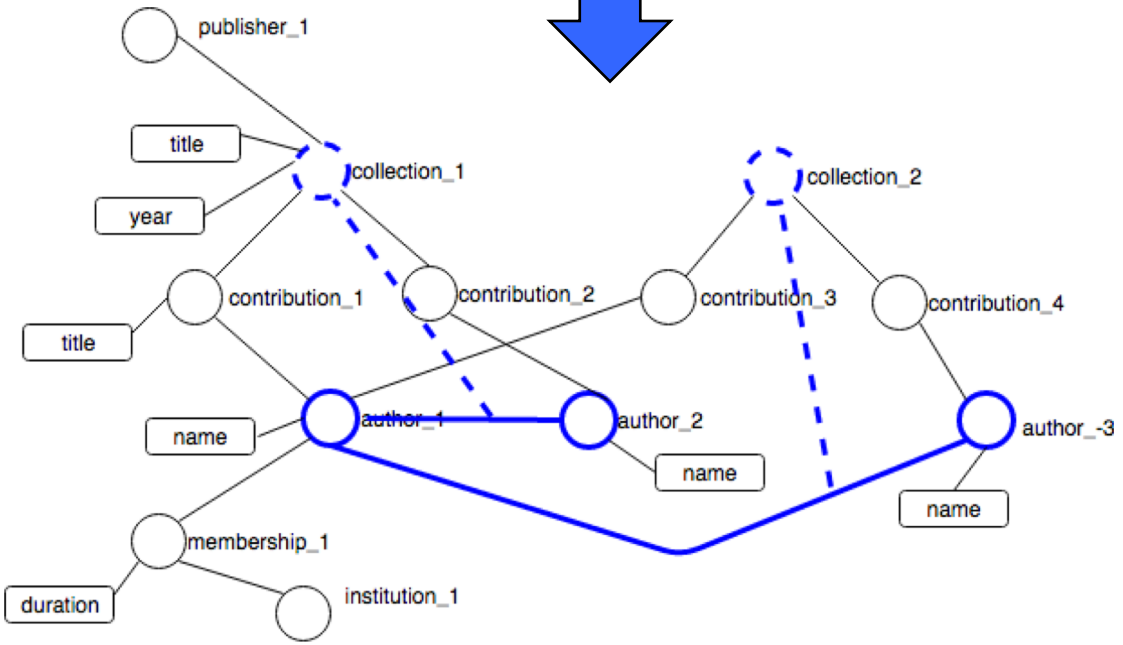
Starting Point: NetworKit

- NetworKit had been found as a **potential core engine for SNA** in 2015
- Installation of **NetworKit on MacOS** in Summer 2016
- *"Now the SNA core engine is running, **what else do we need** in order to do SNA on a particular domain using some domain specific data set?"*
  - How can the **graphs** to be analysed by NetworKit be **created**?
  - How can the graphs and the analysis results be **visualised**?
  - How can the visualisation be **explored** and **enriched** by **domain specific data** to qualitatively augment the quantitative SNA results?
- Motivation of **enrichment:**
  - **Graphs for SNA** need to be **extracted** from raw data using appropriate queries
  - But: **loss of detail information** by merging of **multiple aspects** of raw data into the **single dimension** of a relation between agents

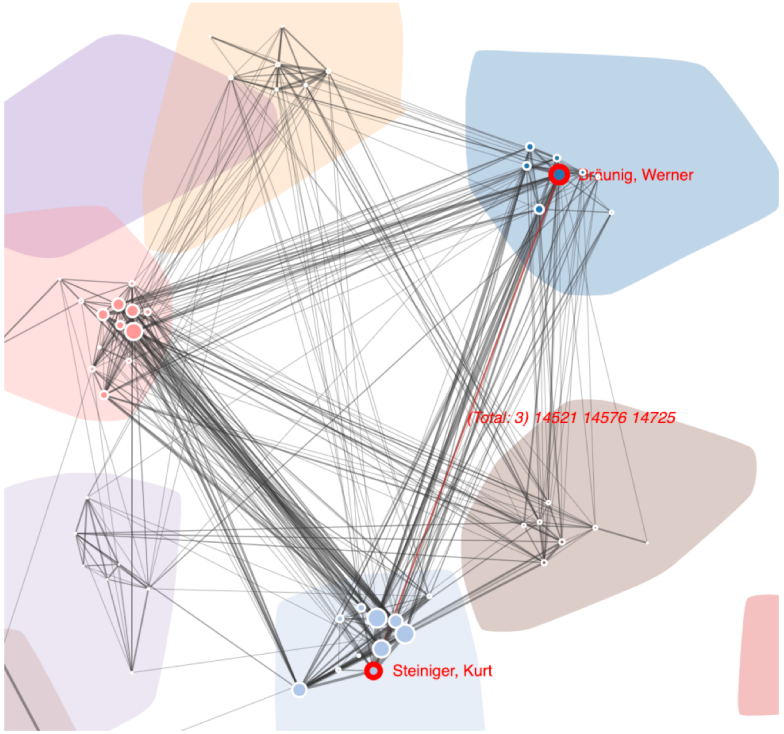
Raw Data *vs.* Graph Data



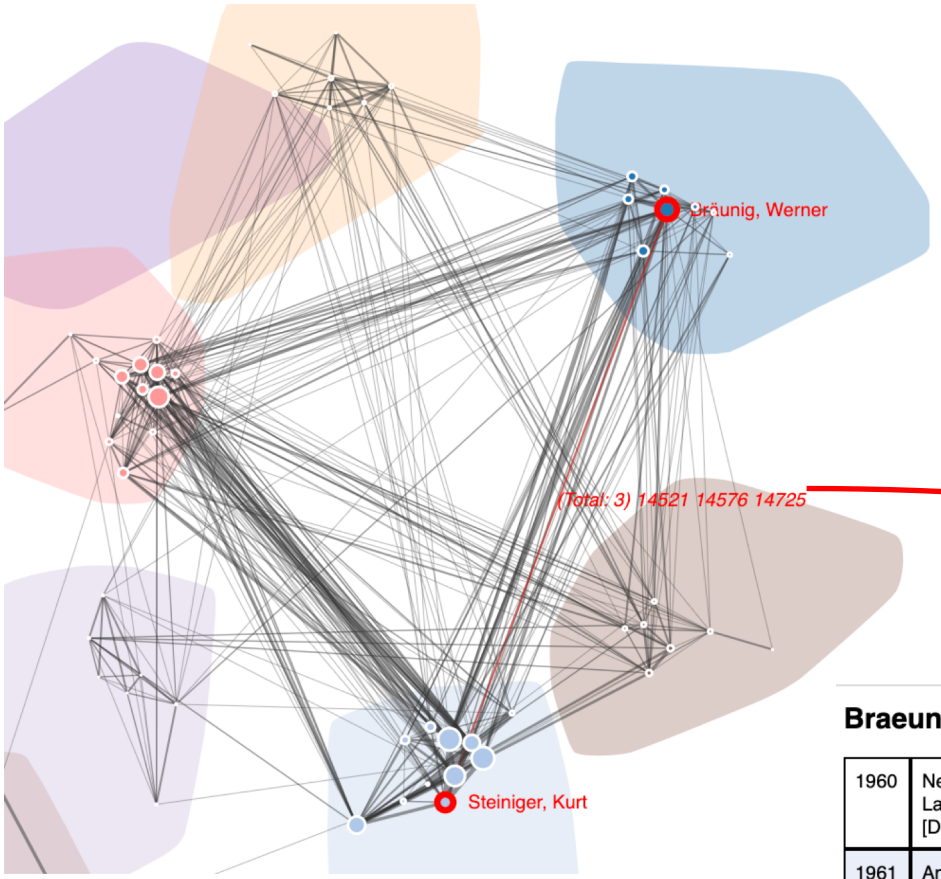
„graph extraction“



Graph itself only represents information on **related entities** and those **entities that constitute** relations („edge values“)



Graph Data Enrichment



*How can we seamlessly **mediate** between a **"distant reading"** (Moretti 2005) perspective of analysis results (e.g. based on entities' KPIs calculated by SNA tools, like centrality) to a **"close(r) reading"** of an entity's details (e.g. their attributes beyond the given network)?*

?

Braeunig__Werner ⟶ Steiniger__Kurt

1960	Neue Landpostille : [Dorfgeschichten]	9	Bräunig, Werner, Bär, Heinz, David, Kurt, Endler, Adolf, Habel, Hildegard, Jakobs, Karl-Heinz, Neumann, Margarete, Seeger, Bernhard, Steiniger, Kurt
1961	An den Tag gebracht : Prosa junger Menschen	11	Branstner, Gerhard, Bruyn, Günter de Heiduczek, Werner, Bräunig, Werner, Jakobs, Karl-Heinz, Neutsch, Erik, Paschke, Erika, Radetz, Walter, Sachs, Heinz, Steinhaußen, Klaus, Steiniger, Kurt, Wolf, Christa
1963	Ihr seid unsere Zukunft : e. Anth. über d. Hilfe d. SED für d. Jugend	23	Arlt, Wolfgang, Axen, Hermann, Becher, Johannes Robert, Bernhard, Rolf-Peter, Brecht, Bertolt, Bräunig, Werner, Cwielong, Hilde, Engelmann, Günter, Grotewohl, Otto, Jugel, Günther, Kuba, Mammach, Klaus, Matern, Hermann, Müller, Armin, Neutsch, Erik, Pieck, Wilhelm, Rebetzky, Ursula, Schädlich, Werner, Stefan, Helga, Steiniger, Kurt, Thoms, Lieselotte, Ulbricht, Walter, Weinert, Erich

Issues to be resolved

1. How can creation, analysis and visualisation of graphs be **integrated** into a **seamless** workflow?

2. How can information represented by a network graph be **enriched** by linkable data?

3. How can **domain dependent enrichment** functionality be supported by the **generic** implementation of the graph **visualisation** component?

Approach

- Necessity of a **generic software solution** as an alternative to “scripting-on-demand”
  - Fast **portability** to new domains and datasets
  - **reusability** of analyses for different datasets with identical structure
- Overall **objective**: provision of functionality via a **web-based service and UI** without necessity of on-site installation
- Related technical work:
  - Gephi and R support creation and analysis of networks based on SPARQL queries (Totet 2015; van Hage)
  - Gephi allows additional **custom attributes** for nodes and edges and foresees UI extension points for **custom java plugins**
  - BUT: **No** “rich” interactive **web based UI** provided

Functionality

Development History

- Starting in late summer 2016: “**Proof of Integration**” for SPARQL+SNA+Web-based visualisation
- From the beginning: usage of **NetworKit as SNA tool** (Staudt et al. 2015)
  - ? *“Building a web-based Environment for Network Analysis and Exploration on top of NetworKit”*
- Initial Scenario: **Co-Occurrence networks** of drama characters based on “Projekt Gutenberg” data (see, e.g., Trilcke et al.)
- Application to the above mentioned bio bibliographical data sets
- Since then **continuous enhancement** due to proven practicability and expressivity of analyses in the area of bio-bibliographical data (see Kreutel et al. 2019)

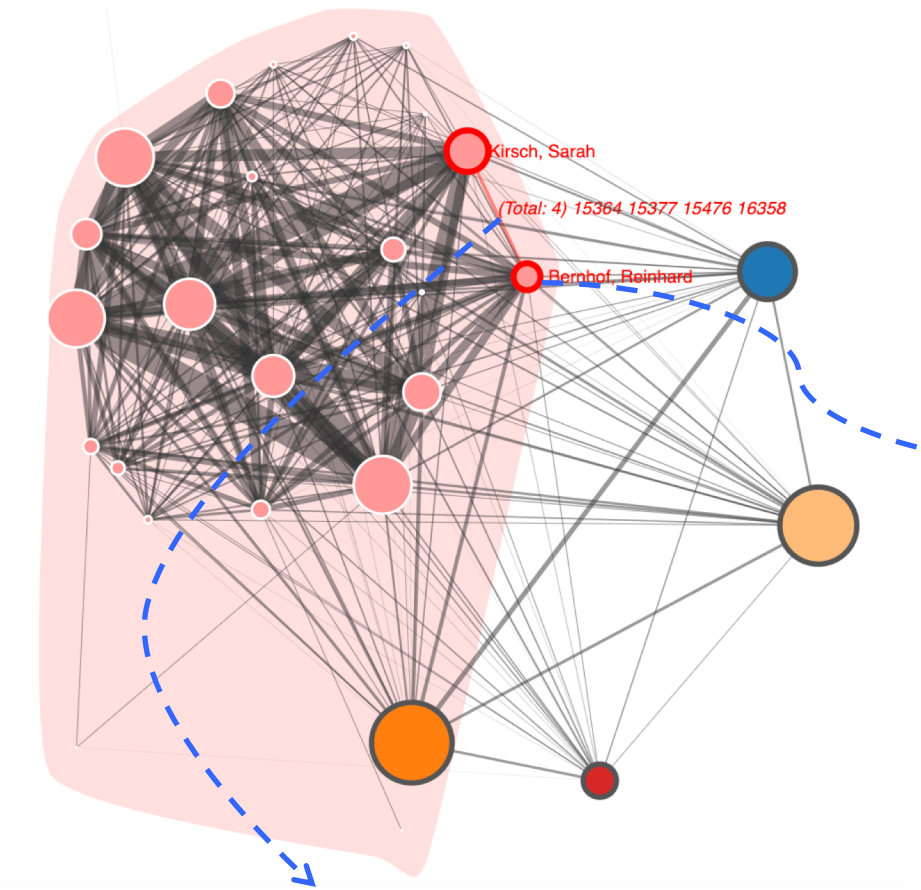
Key Features

1. Integrated **Workflow and UI**
   - **UI-based configuration** of query and analysis details (query, data sources, analysis parameters)
   - integrated **processing pipeline** for SPARQL etc. querying, graph creation, SNA and visualisation, **triggered by a single action** at UI level

2. Domain specific **Enrichment** of graph data and visualisation
   - **Network Manipulation**: manipulation of graph visualisation based on domain specific KPIs for the entities represented by the graph's nodes (e.g. node size based on number of overall publications)
   - **Network Browsing**: exploration of nodes and edges by providing domain specific views of entities based on linkable data
   - **Network Statistics**: applying analysis templates to selected nodes and/or edges of the graph (e.g. for obtaining insight into qualitative features of detected communities as a potential source of community formation); correlating domain specific KPIs with SNA KPIs

3. Enrichment as **Scripting**
   - Enrichment functionality is implemented in **JavaScript** and managed via the UI

Enrichment Example



Bernhof__Reinhard --- Kirsch__Sarah

75a0016	Don Juan überm Sund : Liebesgedichte	1975
77a0098	Vor meinen Augen, hinter sieben Bergen : Gedichte vom Reisen ; e. Anth.	1977
76a0015	Bekanntschaften : e. Anth.	1976
89a0100	Selbstbildnis zwei Uhr nachts : Gedichte ; e. Anth.	1989

edge value enrichment

Centrality

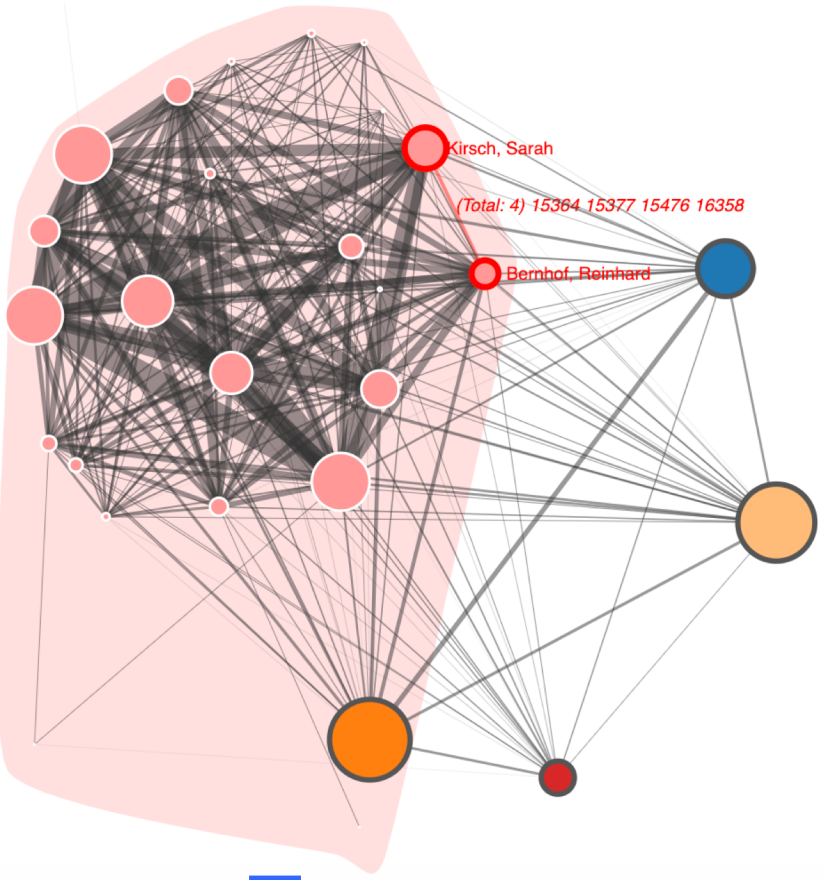
overview table enrichment

export

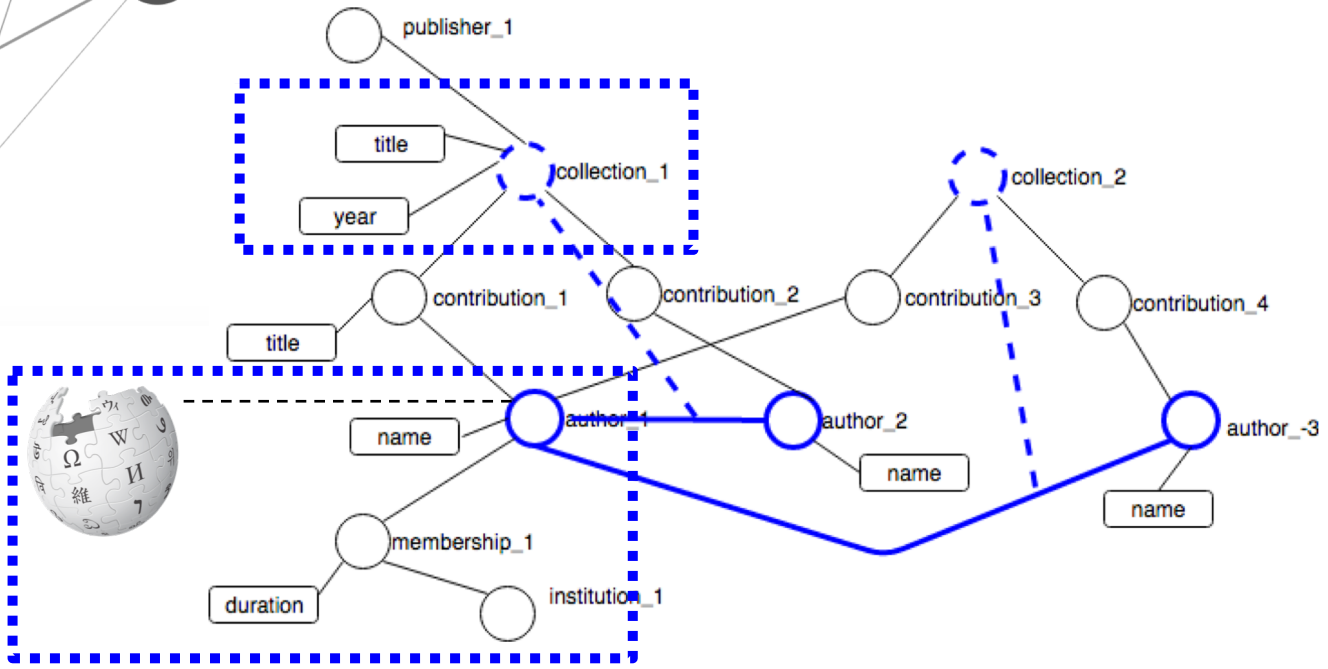
Name	Degree		Eigenvector	Betweenness	DS	FS	SK	Publications	
Werner__Walter	0.29268	24	0.45503	0.13087	1956-1959 1966-1967		1966-1967 1979-1980	42	info
Preissler__Helmut	0.10976	9	0.41144	0.0014614	1955-1956			46	info
Endler__Adolf	0.10976	9	0.39560	0.0032718	1955-1957			41	info
Jakobs__Karl-Heinz	0.18293	15	0.30579	0.0000	1956-1958			31	info
Lindemann__Werner	0.13415	11	0.28593	0.0019591	1955-1957		1968-1968	32	info
Steiniger__Kurt	0.14634	12	0.28280	0.011509	1956-1959			24	info
Czechowski__Heinz	0.17073	14	0.18028	0.024701	1958-1961		1969-1970 1984-1985	33	info
Schulze__Axel	0.19512	16	0.17653	0.0000	1964-1967		1971-1972 1984-1985 1989-1990	30	info
Braeunig__Werner	0.24390	20	0.16259	0.11764	1958-1961			32	info
Reimann__Andreas	0.085366	7	0.12406	0.0000	1965-1966		1974-1975	13	info
Baierl__Helmut	0.097561	8	0.12390	0.00080023	1955-1957			15	info
Schulz__Max_Walter	0.18293	15	0.12369	0.022270	1957-1959			18	info
Bernhof__Reinhard	0.17073	14	0.10407	0.066433	1965-1967		1972-1973 1985-1986	18	info
Kirsch__Rainer	0.17073	14	0.10250	0.069604	1963-1965			36	info
Viertel__Martin	0.18293	15	0.091075	0.015753	1956-1959			16	info
Kirsch__Sarah	0.097561	8	0.087125	0.0063951	1963-1965			24	info

node enrichment

Effect of Enrichment

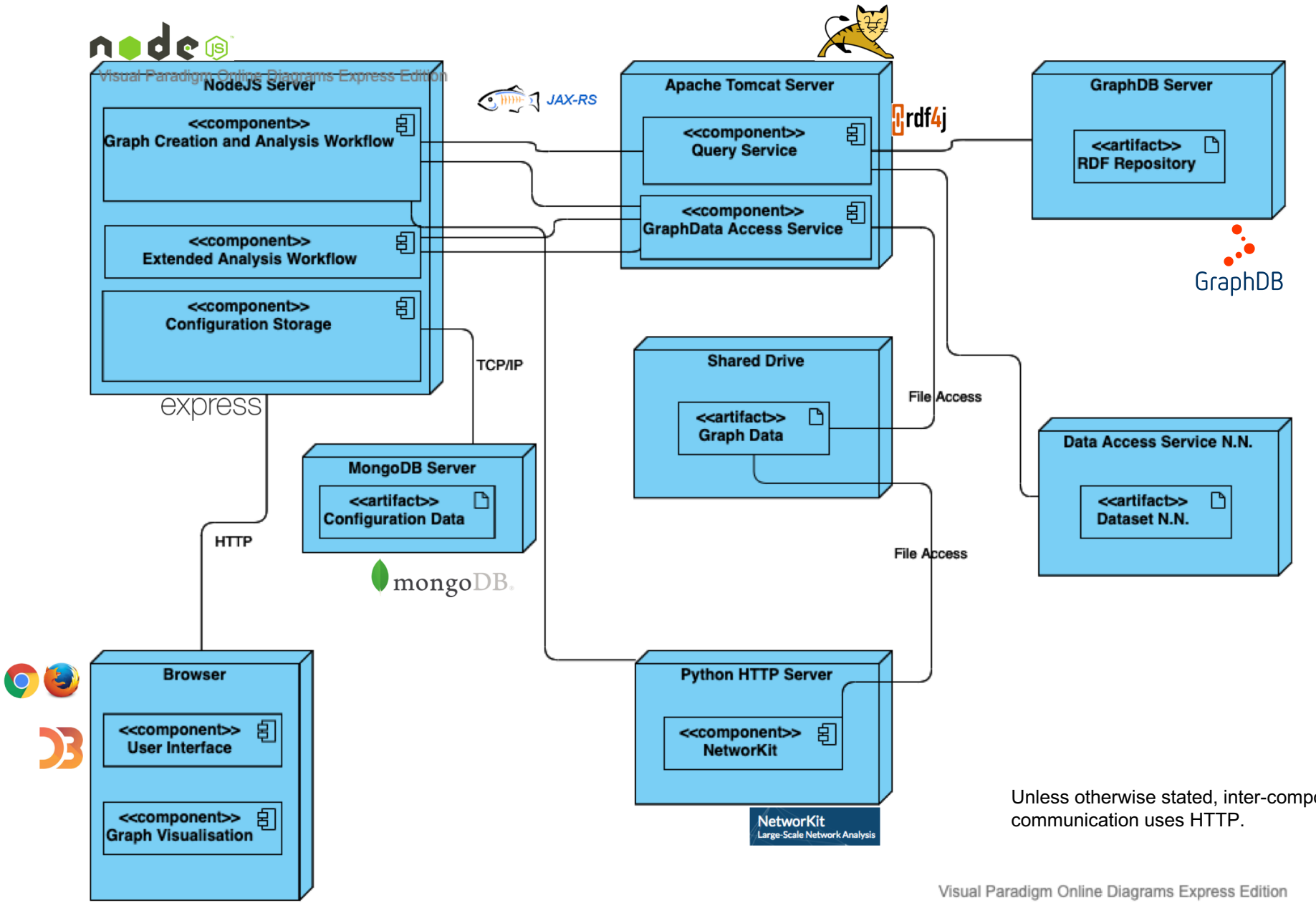


- **re-integration of data** that has been abstracted over by graph creation
- enrichment may cover **arbitrary domain specific data** that can be linked with the entities contained within the graph
- **separation of concerns** between graph data and analysis and enrichment functionality (vs. custom attributes on the graph itself)

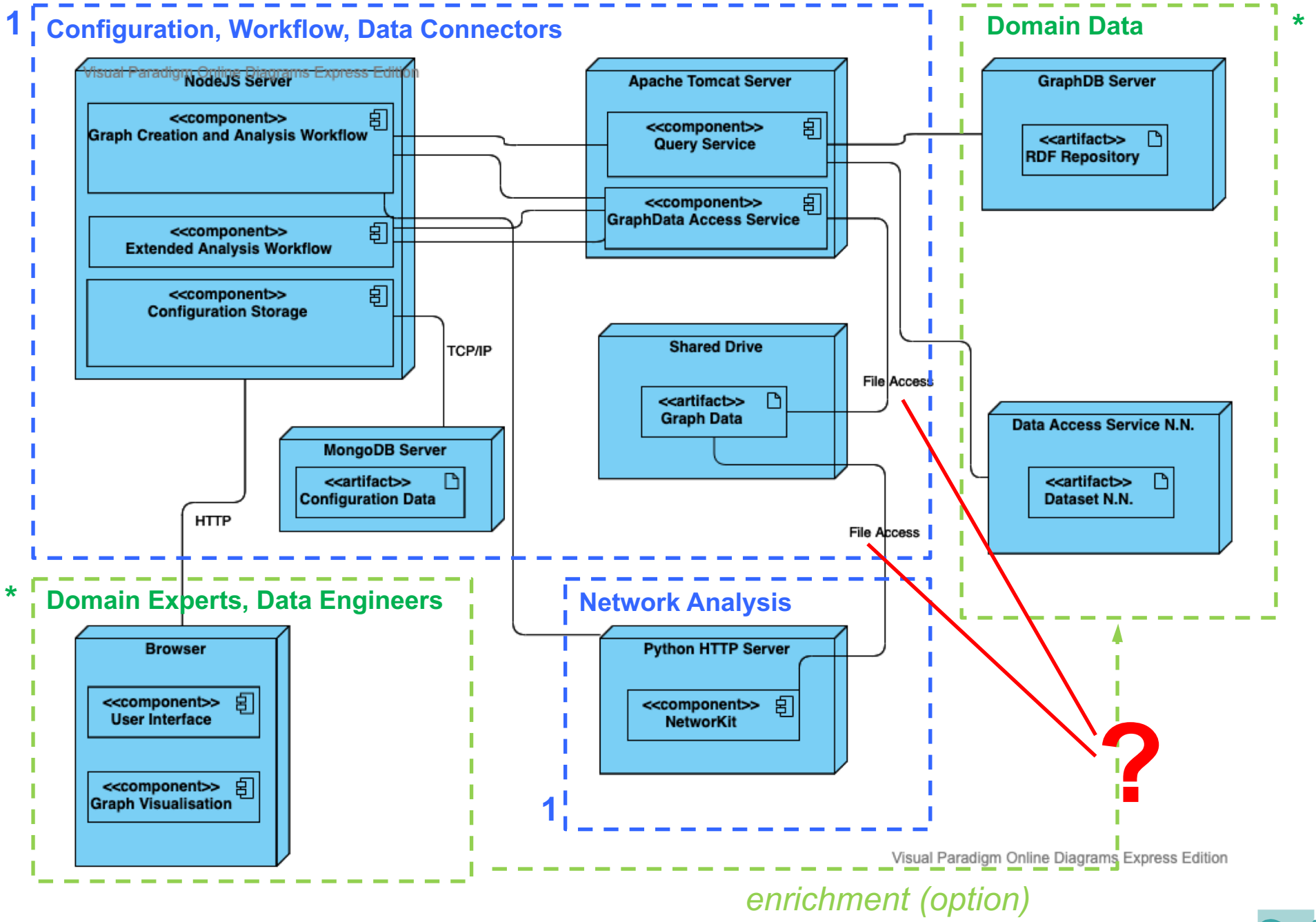


Architecture

System Architecture and Implementation



A Web-based *Platform* for SNA?



Current State and Open Issues

Current State and Open Issues

- **Domain independent solution** applied to and further developed on the basis of two example domains and various datasets
  - Straight ad-hoc usage for **analysing new domains** and datasets (e.g. Wikidata movies and actors)
- BUT: So far, **no self service by domain experts** for graph exploration and analysis (additionally backed by DH engineers engineers doing setup and developing extension scripts).
- Both UI and underlying implementation would benefit from some cleanup.
- Large Graphs seriously challenge **performance of browser-based visualisation** both on the basis of SVG (D3) and WebGL (sigma.js).
- Current environment is rather a **prototype for a web based SNA platform** providing integration and visualisation services for SNA on the basis of NetworKit than such a platform itself.

DEMO

Thank you!

joern.kreutel@beuth-hochschule.de

References

- Bastian M., Heymann S., Jacomy M.: Gephi: an Open Source Software for Exploring and Manipulating Networks. In: International AAAI Conference on Weblogs and Social Media. (2009)
- Bottero, W., Crossley, N.: Worlds, Fields and Networks: Becker, Bourdieu and the Structures of Social Relations. Cultural Sociology 5(1) (2011)
- Bourdieu, P.: Die Regeln der Kunst. Genese und Struktur des literarischen Feldes. Suhrkamp (1999)
- De Nooy, W.: Fields and Networks: Correspondence Analysis and Social Network Analysis in the Framework of Field Theory. Poetics 31(5-6) (2003), pp. 305–327
- Harris, S., Seaborne, A.: SPARQL 1.1 Query Language. https://www.w3.org/TR/sparql11-query/ (last accessed on October 09, 2020) (2013)
- Kreutel, J., Martus, S., Thomalla E., Zimmer, D.: Die Germanistik der Germanistik – Qualitative und quantitative Studien zur Wissenschaftsgeschichte eines 'Referatenorgans'. IASL 44(2) (2019)
- Kreutel, J.: Augmenting Network Analysis with Linked Data for Humanities Research. In: Kremers, H. (ed.), Digital Cultural Heritage. Springer 2019.
- Meeks, E.: D3js in Action. Manning (2015)
- Moretti, F.: Graphs, Maps, Trees: Abstract Models for a Literary History. Verso (2005)
- Newman, M.E.: Coauthorship Networks and Patterns of Scientific Collaboration. In: Proceedings of the National Academy of Sciences 101(suppl. 1) (2004), pp. 5200–5205
- Perer, A., Shneiderman, B.: Balancing systematic and flexible exploration of social networks. IEEE transactions on visualization and computer graphics 12(5), pp. 693-700 (2006)
- RDF Working Group: Resource Description Framework (RDF). http://www.w3.org/RDF/ (last accessed on October 09, 2020) (2014)
- San Martín, M., Gutierrez, C.: Representing, Querying and Transforming Social Networks with RDF/SPARQL. In: European Semantic Web Conference. Springer (2009), pp. 293–307
- Scott, J.: Social Network Analysis. Sage, 3rd edition (2012)
- Staudt, C., Sazonovs, A., Meyerhenke, H.: Networkit: An interactive Tool Suite for high-performance Network Analysis. Computing Research Repository, abs/1403.3005. https://arxiv.org/pdf/1403.3005.pdf (last accessed on October 09, 2020) (2014)
- Totet, M.: Let's Play Gephi: Dbpedia, RDF, Sparql and your favorite Actors. http://matthieu-totet.fr/Koumin/2015/09/06/lets-play-gephi-dbpedia-rdf-sparql-and-your-favorite-actors/ (last accessed on October 09, 2020) (2015)
- Trilcke, P., Fischer, F., Göbel, M., Kampkaspar, D.: Digital Literary Network Analysis. http://dlina.github.io/ (last accessed on October 09, 2020) (2015)
- van Hage, W.R.: SPARQL for R Tutorial - Hollywood Social Network Analysis. http://semanticweb.cs.vu.nl/R/sparql_hollywood/sparql_hollywood.html (last accessed on October 09, 2020)