Making Research Data FAIR. Seriously?

Reflections on Research Data Management in the Computational Literary Studies

Digital Humanities 2022 | Tokyo | 25-29 July 2022

Patrick Helling, Kerstin Jung and Steffen Pielström
DFG Priority Program 2207 „Computational Literary Studies“ | University Würzburg, Germany

Computational Literary Studies (CLS)
Interdisciplinary field of research combining research questions from Literary Studies with methods and technologies from Computer Sciences and Computational Linguistics

SPP 2207 | DFG Priority Program 2207

Computational LiteraryStudies

Zeta and Company – Measures of Distinctieness for Computational Literary Studies
(Christof Schöch | Universität Trier)

Was ist wichtig? Schlüsselstellen in der Literatur
(Robert Jäschke & Steffen Martus | Humboldt-Universität of Berlin)

Structuring Literature – Variants and Functions of Reflective Passages in Narrative Fiction
(Anke Holler, Caroline Sporleder & Benjamin Gittel | Georg-August-Universität Göttingen)

Relating the Unread – Network Models in Literary History
(Ulrik Brandes & Thomas Weitin | ETH Zürich/TU Darmstadt)

Quantitative Drama Analytics: Tracking Character Knowledge (Q:TRACK)
(Nils Reiter | Universität zu Köln)

Evaluating Events in Narrative Theory
(Evelyn Gius & Chris Biemann | TU Darmstadt/Universität Hamburg)

Emotions in Drama
(Christian Wolff & Katrin Dennerlein | Universität Regensburg/Universiäty Würzburg)

Computer-aided Analysis of Unreliability and Truth in Fiction – Interconnecting and Operationalizating Narratology (CAUTION)
(Jonas Kuhn & Janina Jacke | Universität Stuttgart/Universität Hamburg)

CHYLSA (Children‘s and Youth Literature Sentiment Analysis)*
(Berenike Herrmann, Arthur Jacobs, Gerhard Lauer & Jana Lüdtke | Georg-August-Universität Göttingen/Free University Berlin/Universität Basel)

The beginnings of modern poetry – Modeling literary history with text similarities
(Simone Winko & Fotis Jannidis | Georg-August-Universiäty Göttingen/Universität Würzburg)

Anomaly-based large-scale analysis of style and genre reflected in the use of stylistic devices in medieval literature
(Joachim Denzler & Sophie Marshall | Universität Jena)

Main Tasks

- Improving interdisciplinary exchange between the projects
- Supporting researchers of the priority program in questions on reseach data management (RDM)
- Developing a common, domain-specific research data management strategy

Measuring the Landscape

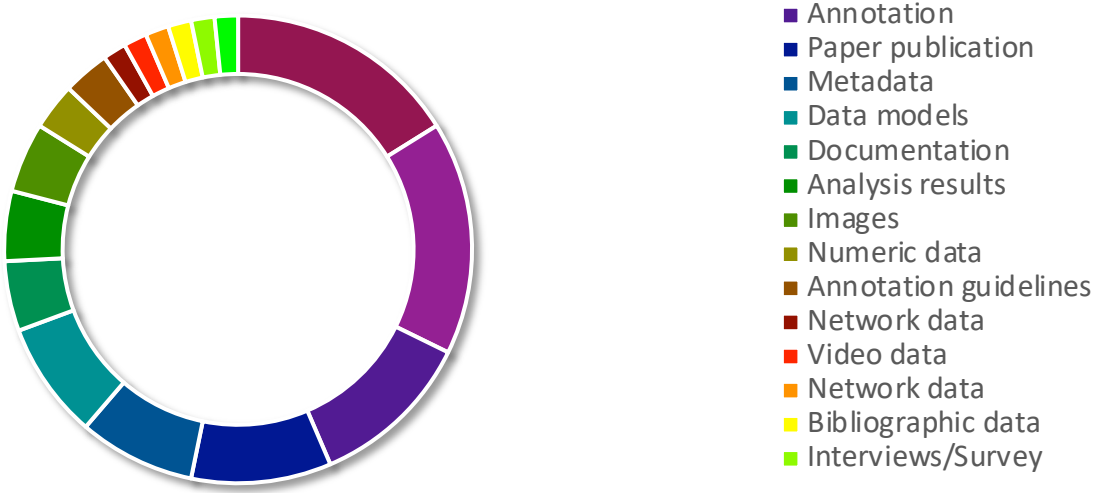
- Qualitative, guideline-based interviews with all projects (in sum: 47 questions)
- Three iterations
- Multiple reviews through the projects

Questions

- **Daily work with research data**
  - Use of tools and VREs
  - Use of methods
- **Research data management during the project**
  - Collaboration in the project
  - Backup strategies
  - Data exchange
- **Research outputs**
  - Development of software and tools (type, functionality, programming language etc.)
  - Data types and formats
- **Existing archiving strategies**
- **Existing publication strategies**
- **Handling of developed software and tools during and after the project**

Scientific outputs of projects within the priority program (and probably for the whole scientific field of the CLS) are highly heterogeneous

Data types



Scientific outputs of projects within the priority program (and probably for the whole scientific field of the CLS) are highly heterogeneous

Data formats

- python
- pdf
- plaintext
- txt
- LaTeX
- tsv
- png
- svg
- pytorch
- graph
- html
- ocr-format
- csv
- xml
- tei-xml
- json
- catma-xml
- R
- word
- java
- binary formats
- jpeg
- excel
- table formats

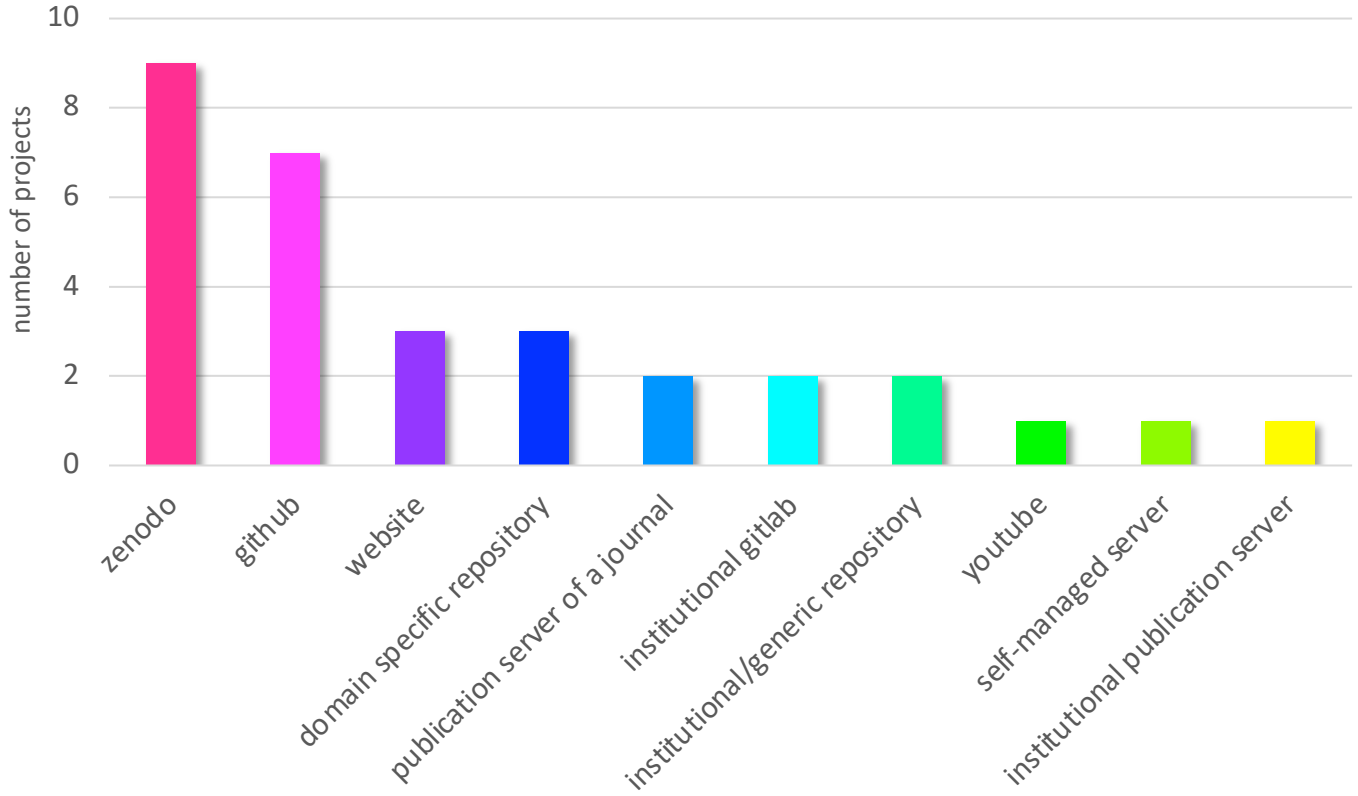
(How) Can we make all this research data and software

Findable, **A**ccessible, **I**nteroperable and **R**eusable

in the sense of the **FAIR**-Principles?

Used infrastructures for publication

Computational LiteraryStudies

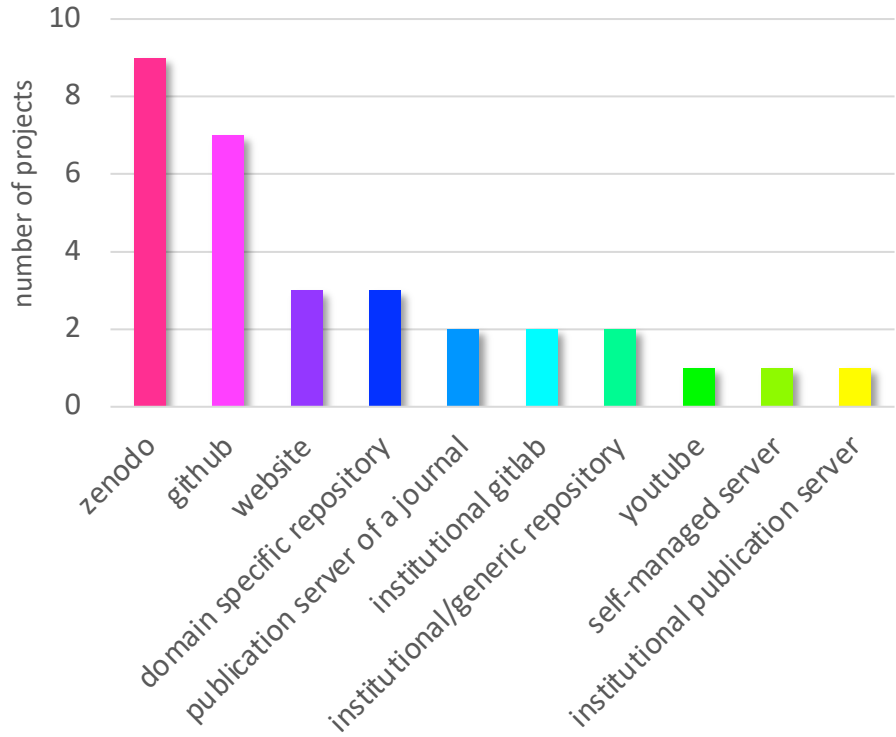


Findability and **Accessibility**

Used infrastructures for publication - findability

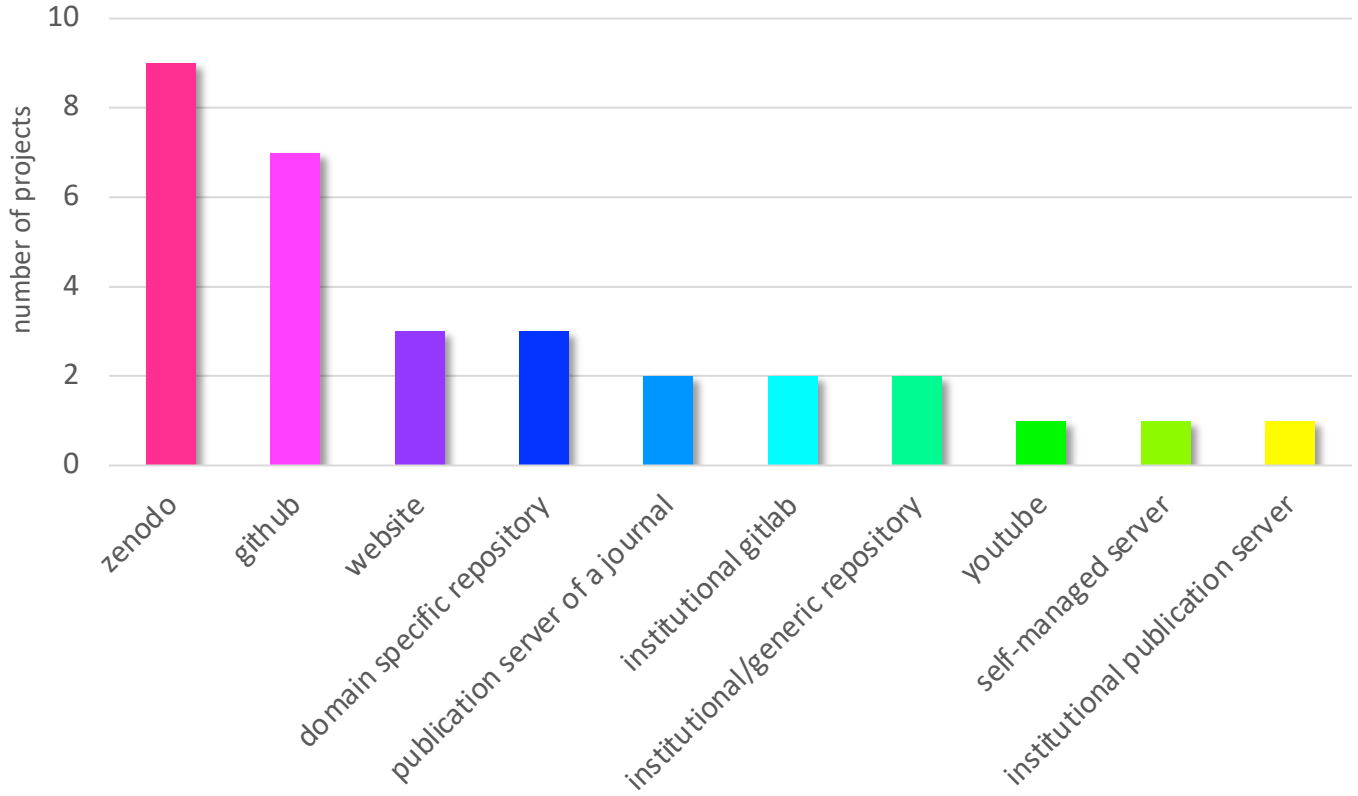
Computational LiteraryStudies

- Usually, data can be described with metadata, even if it is not a domain-specific metadata schemes
- Some of the infrastructures offer persistent identifier (although not all of the infrastructures)
- Not all the data are registered/indexed by searchable resources



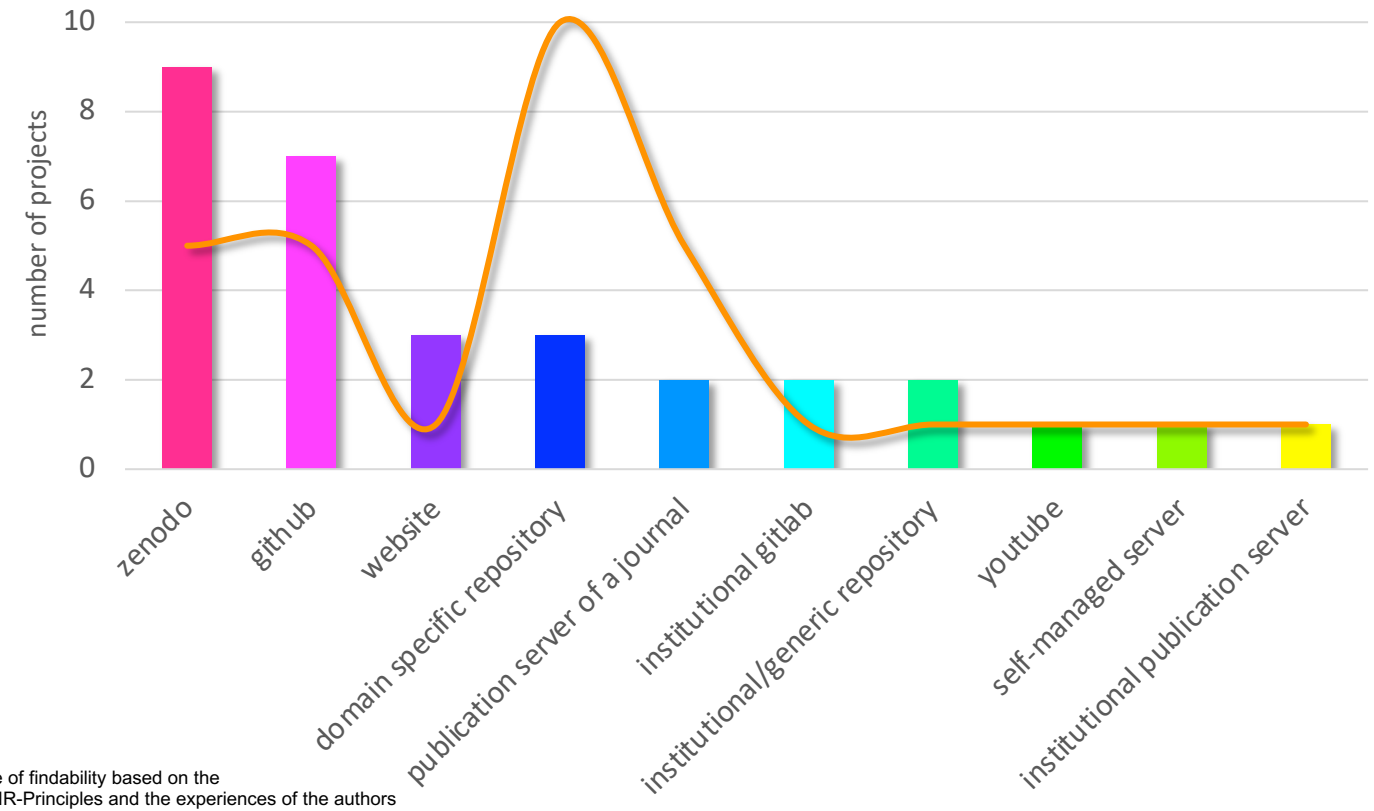
Estimation of the degree of findability

Computational LiteraryStudies



Estimation of the degree of findability

Computational LiteraryStudies

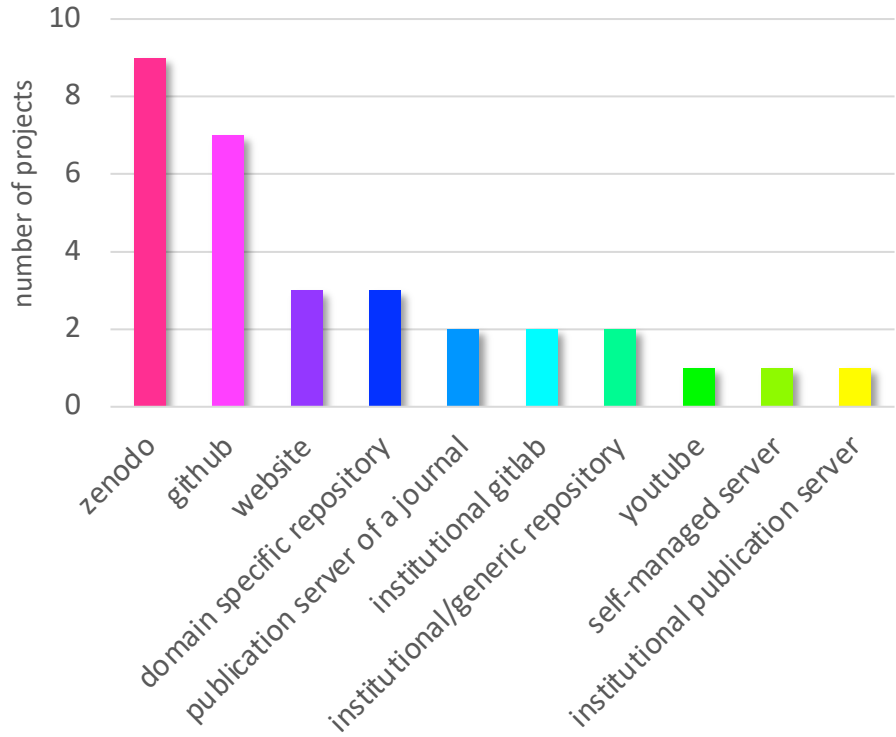


General estimation of the degree of findability based on the common interpretation of the FAIR-Principles and the experiences of the authors

Used infrastructures for publication- accessibility

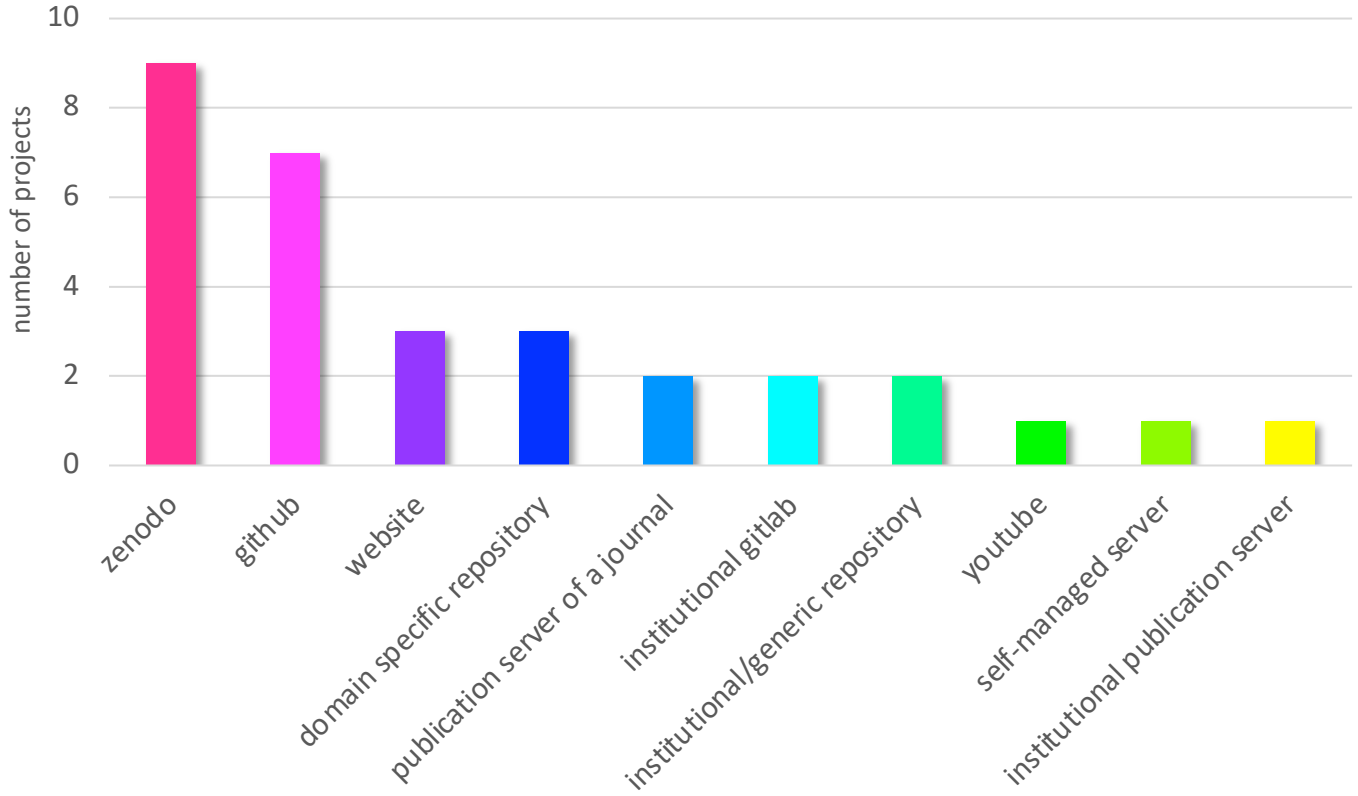
Computational Literary Studies

- Usually, data can be accessed via standardized communication protocols
- Not every infrastructure offers specific authentication and methods for authorization
- In some cases it should be possible to make only the metadata accessible



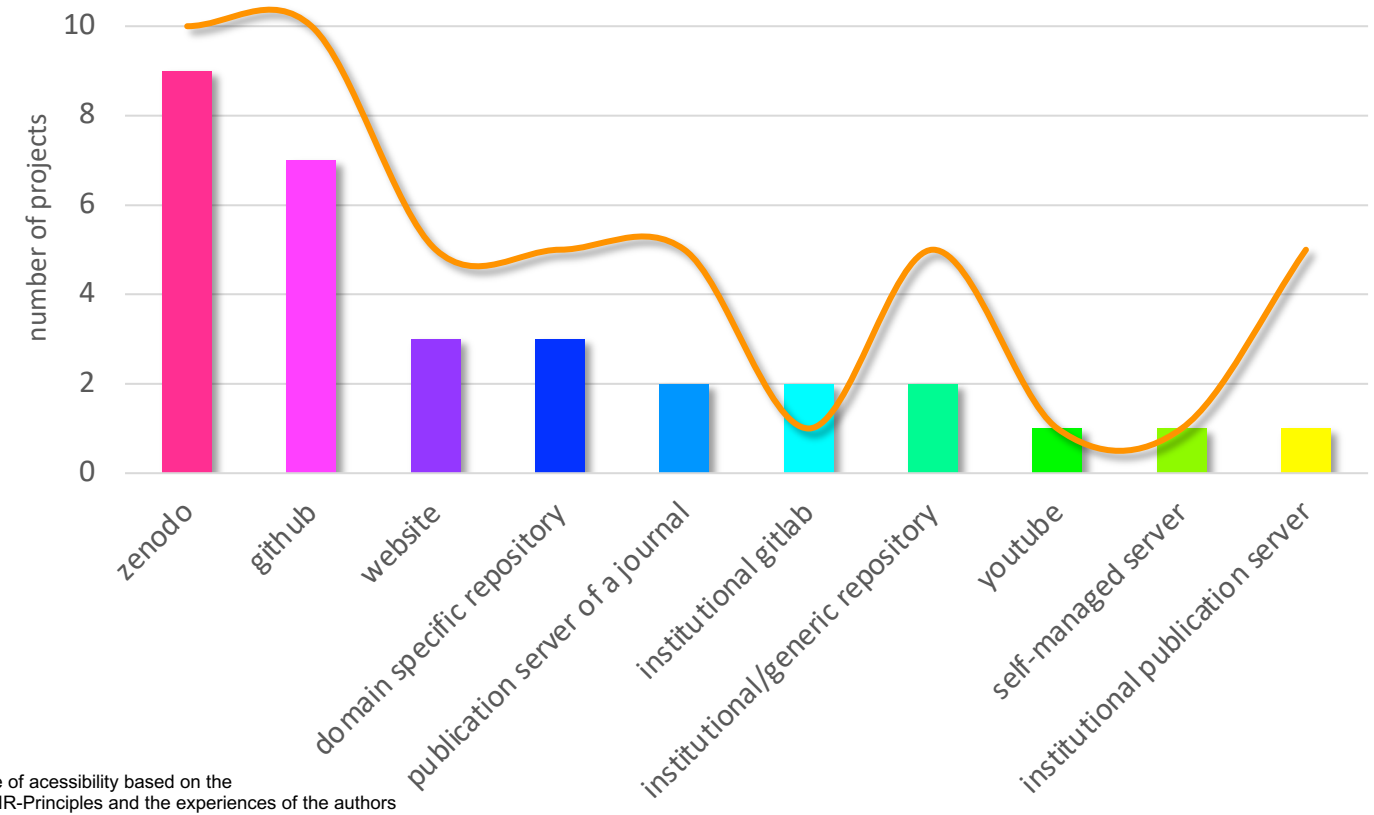
Estimation of the degree of accessibility

Computational LiteraryStudies



Estimation of the degree of accessibility

Computational LiteraryStudies



General estimation of the degree of acessibility based on the common interpretation of the FAIR-Principles and the experiences of the authors

Interoperability and **Reusability**

Aspects of reusability: legal issues

Computational LiteraryStudies

Copyright (German „Urheberrecht“)

- Recent works
- Individual negotiations with authors, publishers, libraries

Personal rights

- User studies
- Surveys

Landscape review

- Community has large interest to make data as accessible as possible and provide secure licenses

Pragmatic solutions regarding „Urheberrecht“?

- Only use primary data which is beyond copyright (e.g. by age)
  → restriction of research focus

- Derived formats (Schöch et al. 2020)
  - Scrambled words
  - Replacement by parts-of-speech
  - N-grams
  - Embeddings
  - …
    → Context size unclear

- Task-based text and data mining (Xsample, Gärtner et al. 2021)
  → gray area
- Detailed metadata for set up of collaboratation

Aspects of reusability: process metadata

Domain-specific challenge

- Equal treatment of automatic and manual steps

Typical automatic processes

- Natural language processing
- Format conversion
- Quantitative analysis
- …

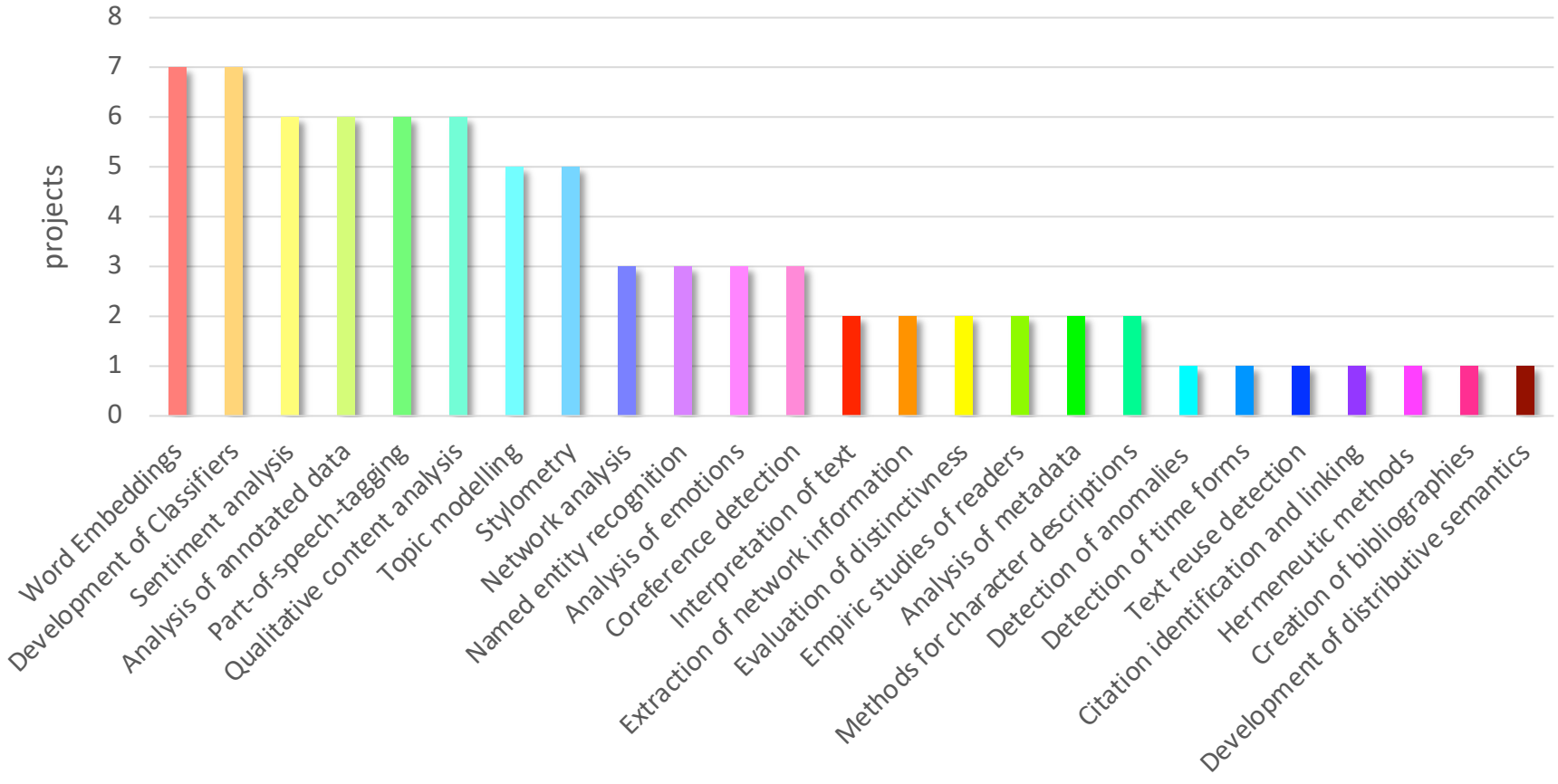
Typical manual processes

- Annotation
- Selection
- Qualitative analysis
- ...

Pragmatic solutions?

- Workflow tools often focus on automatically reproducible workflows
- No standard process metadata schema for the domain, but first objectives
  (e.g. RePlay-DH, Gärtner et al. 2018)
- Raise awareness for importance of documenting processing steps
- Handle individual documentation

Use of methods and tools



Aspects of interoperability and reusability: terms and concepts

Computational LiteraryStudies

Metadata technically describing a resource

- Format ✔
- Size ✔
- Character encoding ✔
- ...

Content-related metadata

- Author, time period → subject to research
- Genre → higly negotiated term
- Aspects of methods form different disciplines → important search feature
  - Annotation layers, guidelines, operationalisation
  - Segmentation: (sub)word / phrase / sentence / passage / chapter / document ...
  - Tool reliability
  - ...

Vocabulary

- Is important for search
- Cannot be agreed on without restricting the acutal process of research
- Comprises terms with several uses
- Originate with new data (e.g. net literature)

Pragmatic solutions?

- Domain-specific understanding of FAIR

…what does that mean for the Computational Literary Studies?

Computational LiteraryStudies

Pragmatism

(1) Define specific documentation standards
(2) Awareness of barriers regarding metadata
(3) Use of distributed infrastructures
    → preferably
    (1) domain specific solutions
    (2) generic solutions that support FAIR
(4) …do not forget the living systems
    (e.g. software, tools etc.)
(5) Individual harvest of achievements

Not FAIR in the sense of FAIR

Specific requirements

- Domain specific infrastructures for outputs of CLS-research
- Organisational and technical solutions for the sustainable storage, accessibility and reusability of living systems
- Professional handling of legal issues

Take Home Messages

Computational LiteraryStudies

Research data management (not only) in the Computational Literary Studies has to deal with **highly heterogeneous conditions and requirements**

Not every aspect of FAIR **is adaptable** to every research field

FAIR needs to be **read and implemented in a domain specific way**
FAIR needs to be **evaluated in a domain specific way**

Bibliography

- **Markus Gärtner, Uli Hahn, Sibylle Hermann**: Supporting Sustainable Process Documentation. In: Rehm G., Declerck T. (eds) Language Technologies for the Challenges of the Digital Age. GSCL 2017. Lecture Notes in Computer Science, vol 10713. Springer, Cham 2018. DOI: 10.1007/978-3-319-73706-5_24

- **Markus Gärtner, Felicitas Kleinkopf, Melanie Andresen, und Sibylle Hermann**: Corpus reusability and copyright - challenges and opportunities. In: Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-9) 2021. Limerick, 12 July 2021 (Online-Event), Mannheim, 2021, S. 10–19. DOI: 10.14618/ids-pub-10470.

- **Christof Schöch, Frédéric Döhl, Achim Rettinger, Evelyn Gius, Peer Trilcke, Peter Leinen, Fotis Jannidis, Maria Hinzmann, Jörg Röpke**: Abgeleitete Textformate: Text und Data Mining mit urheberrechtlich geschützten Textbeständen. In: Zeitschrift für digitale Geisteswissenschaften. Wolfenbüttel 2020. text/html Format. DOI: 10.17175/2020_006

SPP 2207 | Thank you very much

Computational LiteraryStudies

SPP 2207 Computational Literary Studies

E-Mail: cls-fdm-coordination@clariah.de
Twitter: @spp_cls

Patrick Helling | Kerstin Jung | Steffen Pielström

Institut für Deutsche Philologie
Lehrstuhl für Computerphilologie und Neuere Deutsche Literaturgeschichte
Universität Würzburg, Germany