dblp Advisory Board meeting 2022-09-15

- September 15-16, 2022, Begin: 16:00
- Schloss Dagstuhl, Oktavieallee, 66687 Wadern, Germany

Participants on location

- Hannah Bast (University of Freiburg, Germany)
- Silvio Peroni (University of Bologna, Italy)
- Lydia Pintscher (Wikimedia Deutschland, Germany)
- Ruzica Piskac (Yale University, CT, USA)
- Ralf Schenkel (University of Trier, Germany)
- Raimund Seidel (Saarland University and Schloss Dagstuhl, Germany)
- Marcel R. Ackermann (Schloss Dagstuhl, Germany)
- Florian Reitz (Schloss Dagstuhl, Germany)
- Benedikt Maria Beckermann (Schloss Dagstuhl, Germany)
- Martin Blum (Schloss Dagstuhl, Germany)

Participants via Zoom

- Guillaume Cabanac (University of Toulouse, France)
- Martin Fenner (Front Matter, Germany)

Could not participate

- Rüdiger Reischuk (University of Lübeck, Germany)
- Oliver Hoffmann (Schloss Dagstuhl, Germany)
- Michael Ley (Schloss Dagstuhl, Germany)

Meeting minutes

Thursday, Sep 15, 2022

item 1: opening remarks

Meeting notes of September 2021 have been approved without changes.

ToDo:

- Marcel R. Ackermann will publish approved minutes on the web pages

item 2: dblp progress report

Presentation: Marcel R. Ackermann (slides can be found in shared folder)

Discussion:

- How much of computer science is covered in dblp?
  - concrete figure is unknown, some older estimates exist
  - Petricek et al (2005) estimate a coverage of about 24% or the entire CS literature
  - Reitz & Hoffmann (2010) estimate a total coverage of 65% of the the conferences listed in the Brazilian CAPES CS evaluation process from 2005.
  - In the GI-DBLP survey among the German CS community from 2013, 71% of all mentioned conferences and 63% of all mentioned journals were already indexed in dblp.
  - since then, missing venues from those surveys have been added to dblp
  - OpenCitations reference links are used by the dblp team to uncover further missing venues.

- Hiring new staff for the dblp team
  - finding new team members for the funded projects turns out to be very difficult
  - the (remote) location of Trier might be a problem?
  - the board suggests to investigate opportunities for more remote work in the dblp team
- Integration of OpenAlex data
  - Who is behind this? Will the infrastructure last? (small organization, run by activists, supported by grants)
  - spiritual successor of the abandoned Microsoft Academic Graph
  - high visibility in the community
- Observation of "dark patterns" for numerous established journals/conferences
  - some journals have grown exponentially in number of papers per year, while declining in terms of quality
  - editorial oversight lacking after publisher change, accepting every paper
  - problem most prominent in "special issues"
  - indexing of suspicious venues has been paused, will be re-evaluated in about a year

item 3: dblp RDF data & knowledge graph

Presentation: Marcel R. Ackermann (slides can be found in shared folder)

Discussion:

- Enriching dblp RDF by further data facets?
  - no plans to allows users to add data directly to dblp
  - gender: no intention to add, information barely available from publishers, first-names are an insufficient proxy
  - for most auxiliary data: rather linking to Wikidata than collecting the data ourselves
  - adding dataset publications: work in progress, see item 4 below
- public SPARQL endpoint
  - prototype already running
  - based on QLever by Hannah Bast
- inclusion of OpenCitations and Wikidata data
  - should we aim for federated files/servcies or a joint RDF file/service?
  - HTTP federation to remote servers has very bad performance in practice
  - board suggests aggregation of a joint RDF file
- API / endpoint
  - practical advise: always set timeouts for user queries
  - idea to explore: hide SPARQL behind REST API ?

ToDo:

- dblp team and Hannah Bast will coordinate to have a dedicated dblp SPARQL endpoint running

Friday, Sep 16, 2022

item 4: data publications in dblp

Presentation: Benedikt Maria Beckermann & Martin Blum (slides can be found in shared folder)

Discussion:

- project objectives
  - making dataset publication metadata findable in dblp
  - enable computer scientist to cite data publications properly
  - motivation: insights from user survey / projects that approached us / dblp team discussions
- expectations
  - data publications should be listed on dblp bibliography

    - assignment of datasets to their true creators
    - giving incentive to cite data publications
    - enriched information, e.g. relations between publications and data sets
    - how commonly do journals ask for data along with publications?
        - artifact evaluation addresses that to a degree
        - we asked ACM about their "artifact badge" data, answer still pending
        - should evaluation result be listed in dblp?
- data sources
    - different approaches for structured and unstructured sources (cf. NFDI vs. Unknown Data project)
    - structured sources:
        - primarily looking at Zenodo and OpenAIRE
        - currently not using open secondary sources like DataCite, Wikkidata, ...
        - GoogleDatasetSearch is a structured (secondary) source - License issues?
        - Martin Fenner can give a number of good pointers
    - unstructured sources:
        - often severe metadata issues from unstructured sources
        - individual data wrappers necessary for each source
        - uncovering sources in Unknown Data project
    - relevant dataset selection strategies:
        - dataset cited by a CS paper?
        - dataset creator is in dblp?
        - dataset classified as CS?
        - dataset published in specific repository?
- data modeling
    - still only focusing on "core bibliographic data" (author, title, source, DOI/URL), also for datasets
    - currently no concept of "versions" in dblp, needs to be developed
    - version modeling also relevant for publications, e.g. arXiv preprints
    - "keep it simple": over-modeling of relations will significantly increase workload for the team
    - focus on export to BibTeX / following data citation principles of FORCE11
- open problems
    - what about duplicates in different repositories, or derived datasets?
    - how to attribute creators when the team changes from version to version?
    - viability of indexing data publication needs to be reevaluated after the initial 3 years
- known limitations
    - minding the small team size, being careful to not introduce more work than the team can handle
    - being very selective in order to not be overwhelmed
    - choosing (semantic) data quality over quantity
    - handling of software properly is non-trivial, will be treated "just as data" in first iteration
    - datasets are still rarely cited (properly), even more rarely via DOI

item 5: PhD theses in dblp

Presentation: Florian Reitz (slides can be found in shared folder)

Discussion:

- coverage of PhD theses is emerging in dblp but very much biased by availability of metadata
- since 2022, 1 FTE in the dblp team is contacting individual US universities for data
- ideas to broaden access to PhD thesis metadata:
    - contact individual seminar participants while at Dagstuhl
    - CRA in the US could have this kind of data
    - DataCite may have has dissertation information
    - Italy has agreement between universities and central library, Silvio Peroni can help
    - Martin Fenner can probably help with CalTech and Australia

item 6: dblp Advisory Board turnover

- The first formal 4-year board terms end in 2023. All board members are eligible for a second 4-year term, ending in 2027.
- Decision: the date for regular board membership turnover is fixed to be July 1st of each year
- We shall aim for a "rolling turnover", such that there will always remain a number of experienced members in the board. To allow for this, we should start adding new members in two years from now at the latest.
- Suggestions for institutions we may ask for new members:
  - Informatics Europe
  - OpenAlex
  - OpenAIRE
  - Google Scholar?

Next meeting

- Thu, Oct 12, 2023 – Fri, Oct 13, 2023
- on-site at Schloss Dagstuhl

We also already fixed the date for the 2024 meeting:

- Thu, Sep 12, 2024 – Fri, Sep 13, 2024

ToDo

dblp team

- build guided interface for user requests, esp: affiliation changes (still open from Sep 2019 meeting)
- overhaul of the dblp search interface and search API (open from Feb 2021 meeting)
- re-evaluate the coauthor graph implementation at some later point (open from Sep 2021 meeting)
- coordinate with Hannah Bast to have a dedicated dblp SPARQL endpoint running (done)

Marcel R. Ackermann

- publish approved minutes on the web pages (done)
- coordinate with Hannah Bast about a search overhaul (open from Feb 2021 meeting)
- Marcel R. Ackermann will look for ways devising data download / API statistics (open from Sep 2021 meeting)