BibSLEIGH: Bibliography of Software (Language) Engineering in Generated Hypertext

Vadim Zaytsev
vadim@grammarware.net
Universiteit van Amsterdam, The Netherlands
Raincode, Belgium

Abstract

The body of research contributions is vast and full of papers. Existing projects help us navigate through it and relate authors to papers and papers to venues. In this paper we list features missing from those projects and propose a solution in the form of BibSLEIGH — a work in progress on facilitated browsing of scientific knowledge objects. Through leveraging domain focus, by actively employing automated data collection and scraping tools, and with automated annotating of the corpus, we are able to gain and provide insights into scientific communities and topics, as well as surface potential interdisciplinary opportunities.

1 Motivation

BibSLEIGH has started in 2014 as a project to scratch some personal itches and solve problems that were eating away from the authors' time as well as anyone else's. These issues can be broadly categorised into four categories. In §1.1, we will discuss in some detail problems with the bib\TeX format and the unnecessary diversity of conventions for equivalent items, which has a chance of making academic publications look unprofessional and can also lead to confusion and mistakes. However, consistency enforcing is very time consuming. In §1.2, the focus will be on domain specificity, which is a specialisation, and just as any specialisation, can lead to significant optimisation. We have collected some features in §1.3 that are missing from the current widespread remedies (we refuse to call them solutions). Each of the features is missing for a good reason: each requires research, development and domain focus. This makes them both attractive to invest effort in and dangerous because most are non-trivial. Finally, in §1.4 the most obvious point will be raised about information that is interesting in bibliographical context, being distributed over various unconnected sources of not that structured data.

1.1 Bib\TeX non-uniformity across sources

If we attempt to download .bib files for the same publication from various sources, they will all look differently, sometimes drastically so. Many publishers do not curate their data, rely on automatic text recognition and only occasionally and serendipitously fix misspellings. Bib\TeX providers are often volatile when it comes to conference naming. IEEE and ACM are obviously inclined to include their affiliation ("the IEEE/ACM international conference on...") sometimes in favour of more useful information like the number of the conference in the series. DBLP has changed their policy on abbreviating venue names during the period of writing this paper, (between SATToSE in July 2015 and post-proceedings in November). When information is available, bib\TeX providers usually decide to include it — yet what was the last time someone cared about whether ESOP 1986 took place in Saarbrücken or in Passau? This information can be leveraged for other purposes, like tracking country and continent preferences and their shifting over the years, or investigating the impact of location on the number, quality and affiliation of papers. However, it is not used for any of those purposes, yet included in
the bibliographical entry. Nevertheless, many details about in which hotel near which city on which exact days the conference has taken place, find their way into bib\TeX, even though they were important only for the briefest of times, and only to immediate attendees of the event.

So, on one hand, there is too much information in the bib\TeX entries supplied by publishers and accumulators like DBLP and Google Scholar: addresses, dates, timestamps, keywords, sometimes entire abstracts. On the other hand, however, some of more useful information is routinely missed. Frequent omissions concern editor names and hyperlinks that can be used to access the actual content of the publication. Editor names play exactly the same role in events and journal special issues as author names play in individual publications: they help to identify the item but also establish community links across differently named and formally unrelated events. Hyperlinks are not always entirely missing, but oftentimes hidden behind non-standard fields like \texttt{ee} or \texttt{acmid}; not curated in a way that a doi field sometimes starts with \url{http://}; and even outdated — most if not all links like \url{http://www.computer.org/proceedings/csmr/0546/05460161abs.htm} being provided by DBLP have been dead (HTTP Status 404) for several years since the redesign of the IEEE Computer Society website made them obsolete.

Time lost in reformatting is only a part of this side of the problem. Inconsistencies lead to unprofessional look of those papers whose authors have decided against wasting time on bibliography beautification; and worse yet — to duplicate entries appearing within the same paper with slight variations in spelling and data details provided, which made searching for the right entry harder and clone detection impossible within a typical textual editor.

\subsection{1.2 Lack of domain focus}

Academic researchers tend to specialise but never limit themselves overly to one particular series of events. Yet, when we look at sources of information we have at our disposal, they come in two sizes only. On one extreme we have websites devoted to individual conferences. They usually contain a lot of information that is not immediately required for a decent bib\TeX entry, but can be quite useful in the long run for community recognition: after all, one is much more likely to submit to a conference chaired by someone whose name they recognise and whose work they can relate to that of their own. Organisation committee details and programme committee members provide refreshingly large foundation for automation of this process, as demonstrated by the recent work of Visilescu et al.\cite{VSM13,VSM*14} that harvested P C members of several top conferences and cross-checked them with authors publishing there to measure academic inbreeding. However, the focus of such a website is limited to one event, or in some lucky cases to a series of events, and such websites are very prone to disappearing forever once their organisers retire or change employers.

As the other extreme we have services that make an endeavour to collect information over a broad choice of conferences on all kinds of topics, and put them in one place for display and consumption. The most famous ones are DBLP with its 6500+ venues, Google Scholar which is based on web crawling and Microsoft Academic Search that contains ranking tables sorting conferences of one field by the number of citations their articles enjoyed over the years. Such services try to be as general and comprehensive as possible, and this is exactly where they fail short. Broad generalisations are impossible without compromises on metadata models, on information representation, on clone detection. A website of one particular conference typically shows very clearly which volume of which journal contains its post-proceedings special issue — while DBLP habitually gives you all issues of the conference and all issues of all journals and leaves the search for a match in your own hands. University libraries fall into the same category: while limiting their databases to material available physically or through subscriptions, they do not differentiate among domains, so searching for “mutation” will likely result in many items unrelated to mutation testing; and searching for “graph”, while more productive, will still yield results from graph transformation research as well as from general graph theory.

The quest for broad coverage makes the project vulnerable. For instance, DBLP covers millions of authors and thus has to be extremely careful about not confusing authors with similar names — however, many researchers, especially in the pre-google era, did not write their names always in the same fashion. This would have been known to domain experts who are familiar with key authors in their field, but domain knowledge does not scale up. Similarly, Google Scholar relies on its web crawler, and so it is not uncommon for it to point you to papers that are no longer available or are in fact no papers at all, no matter what their authors claim. Microsoft Academic Search is based on citation information — and as a result of different people citing the same venue in different ways (e.g., with “International Conference” or without it), the same venue appears several times in the ranking, both positioned much lower than they deserve.
1.3 Missing features

When we like a paper, we often begin investigating its authors to see if they have contributed to similar lines of research before or after. DBLP lookup has become a part of a routine check in many cases from research exploration to job candidate evaluation. However, a graph transformation researcher that occasionally published a model transformation paper, or a grammarware engineer masquerading as a metamodel evolution contributor, will have different styles across other of their papers, and might not be as fruitful to investigate if your interest is particular and your time budget is limited. What could have helped here is visualisation beyond textual: instead of browsing through a multi-page wall of text profile on DBLP, some of us would have wanted to take a quick look at a diagram depicting community contribution in a concise and illustrative manner.

Natural language processing techniques have a powerful arsenal: even the simplest analyses like stemming and lemmatisation can provide great aid in surfing through the oceans of papers to pick the right ones to read and cite. It is common knowledge that the names of conferences do not always completely represent their intentions: having “languages” in the name can mean one or two of a dozen of entirely different research directions; venues with “engineering” in their name can get quite science-y and theoretical, just as a name starting with “trends” does not mean all papers are surveys, overviews and vision statements. To the best of our knowledge, no currently existing bibliographic website currently provides a lot of NLP-based features, although ACM Digital Library has recently started collaborating with IBM Watson to pursue that.

Scraping older sources from document scans to websites that fell apart decades ago and have their ruins exposed though the Wayback Machine, is usually beyond the goals and capabilities of bibliographic websites. Armed with domain knowledge and the interest seriously linked to that domain, we can gather enough effort to complete such endeavours and ask senior and emeritus colleagues directly about that one long-forgotten obscure workshop that a reputable conference has grown from.

Grouping and clustering of conferences is usually either manual work, or done though event colocation, or not done at all. The first option is labour-intensive, error-prone, vulnerable to biases and prejudice. The second option delivers complications for roaming venues like BX (deliberately co-locating each year with a different community: ETAPS, STAF, VLDB, etc) and for diverging venues that stopped co-locating deliberately to emphasize pursuing a divergent path. The third option is not an option at all, since even fairly focused researchers will find themselves contemplating submission to a dozen or two reasonable venues. There is quite some space for automated clustering.

Topic-driven grouping is not the only kind of classification that would be sensible for a bibliographic portal: some venues are linked by a subcommunity of people who strongly contribute to both. For instance, there are many people who publish regularly both at MoDELS and ICSME/SCAM, even though they cannot attend both within the same year (they happen simultaneously). Having linked data about people’s contributions, we can surface such relations — and some RDF frontends to DBLP let you do that with a couple of medium-size SPARQL queries.

All that being said in § 1.1 about the state of bib\TeX entries obtainable from available sources, we still want to have some freedom in formatting: everyone in computer science research knows what LNCS is; in a paper submitted to SLIE one does not need to explain this abbreviation; editor names are nice to have but sacrificial under pressing space constraints, etc. We want flexible bib\TeX formatting: DBLP provides you with some very limited options (crossref or no crossref); IEEE Xplore and Elsevier as well (abstract or no abstract); but BibSLEIGH even in its very beginning stage provides its users with more freedom.

Desktop software for managing bibliographies like Mendeley has tagging functionality that can help its users to annotate the papers they read into different categories or add brief descriptions to them. However, there is a huge gap between doing that and providing a comprehensive annotated bibliography on the subject: in fact, such contributions are rare and properly treasured, for it takes a lot of expertise and work to craft them. Unfortunately, there are many topics and subtopics than there will even be annotated bibliographies. We need some semi-automatic way of providing us with at least bundles of related papers if we indicate the selection criteria.

1.4 Distributed information

It was already pointed out above that participating in event organisation and serving in programme committees can be seen as community binding and is therefore metadata of interest. Yet, to the best of our knowledge, there is no project currently dedicated to collecting this kind of information, and it remains scattered half over the internet and half in the Way Back Machine.

Mathematics Genealogy Project [C+] is a totally disconnected project dedicated to documenting topics of doctoral dissertations (and occasionally habilitations) and supervision information. It certainly
has a merit of its own, but we believe it can also be coupled with other kinds of metadata in a sensible way.

Affiliation information very occasionally finds its way into DBLP as well as into Google Scholar where academics can log in and update it (unfortunately, some choose to log in and prohibit Google from ever showing information about them), but there is no easy way of tracking and leveraging it. However, it is not outrageous to think of research dedicated to tracking research centres of activities on particular topics over the years.

Finally, citation information — it is available on publishers' websites in limited form (because they are not big fans of sharing it among themselves) and on Google Scholar (where it is heavily guarded against any form of automated scraping). While acknowledging some interest in it, we choose to avoid this aspect for now, because it is not static by nature: citation information available today can be totally out of date by tomorrow. However, there is a lot of potential research here that goes way beyond traditional bibliometrics: for instance, we can identify canonical sources (which often will be books, like the Dragon Book [ASU85]) that are used throughout a large fraction of papers in a specific conference, and find other venues in a different language that have the tendency to cite translations of this book.

Additionally, academic articles also contain links to web resources such as additional documentation, wikis and tool repositories, and such links have a half life of 4 years on average [Spi03]. The Software Heritage Project was recently proposed by Roberto Di Cosmo as a project to organise, preserve and share all academically produced software to provide much desired availability, traceability and uniformity. Unfortunately the project seems to be in early stages, its call to action is available on SlideShare [Cos15] but the project itself is yet unknown to public search engines. It will be interesting to see if the corpus of BibSLEIGH can be automatically mined for references to tools and clustered by technological space.

2 BibSLEIGH to the rescue!

BibSLEIGH is a work in progress. Keeping that in mind, we would like to sketch preliminary requirements and architecture decisions in §2.1, point out some related work in §2.2 and describe the state of the project as it is by the time of submission in §2.3. Next, §3 will draft some possible future directions we might decide to explore.

2.1 Proposed solution

In the centre of BibSLEIGH there is one centralised repository containing all its data in JSON format — we call it LRJ, short for Lexically Reliable JSON, because we store all key-value pairs one per line sorted by keys. This was chosen over a more classic database setup in order to allow individual traceable edits of each piece of data and at the same time to guarantee user responsiveness. Data is imported to this central place through any of the existing importers, which are usually implemented as iterative parsers (to process the DBLP dump which is around 2 GB) or web scrapers (at this moment we have those for individual DBLP pages, CEUR and EasyChair). JSON files can also obviously be added manually. There is also an ad-hoc importer that creates appropriate JSON entities from a list it reads from a textual file — this helps to properly add ancient entries.

Once the data is in the repository, it can be further curated, normalised, improved, enhanced and crosschecked with other sources. Typical maintenance activities include adding a fresh issue of an already known conference or a journal issue known to be related to one of the known conferences (automated: one just needs to run an incremental updater), improving the name of the proceedings booktitle (semi-automated: changed manually at the top and automatically propagated downwards), removing non-academic clutter such as forewords and panel summaries (manually or heuristic-based). As an example of crosschecking we can talk about adding PC members and organisers: this information is never found on DBLP, but can be harvested elsewhere and integrated into the same system.

Once normalisation reaches a point of being a valid input for analysis, we enrich the data by stemming all titles and tagging them by predefined tags — following the spirit of the rest of the project, each tag has its own definition stored in a separated JSON file which can be accessed, inspected and changed right on GitHub. Stemming provides fully automated foundation to naturally link papers to their conceptual neighbours, tags play the same role for previously known manually defined concepts (so that A-lifting falls under the same tag as \(\lambda\)-calculus, but \(\mu\)-kernel is kept away from \(\mu\)-calculus, even though the characters look similar\(^1\)).

Each tag definition can contain links to Wikipedia, Wikidata and other places that are displayed on the tag’s webpage. Stems can only rely on automatically derivable information, so their webpages display neighbours — stems that are commonly used together with them.

\(^1\) As a side remark, in Unicode there are different symbols: \(\mu\)-kernel is read as "\(\mu\)kernel" and therefore uses the micro sign character (U+00B5), while \(\mu\)-calculus is read as "\(\mu\)calculus" and is thus represented by the Greek small letter mu (U+03BC). BibSLEIGH is the only website that gets it right in all places, the readers are welcome to check.
Whenever the central dataset of BibSLEIGH is needed for inspection, it is formatted as a collection of almost-static XHTML pages: the only dynamic part of them is the pretty-printing of bib\TeX{} itself. The outlook of BibSLEIGH is less austere than that of DBLP, it makes full use of a palette of colours and a collection of icons for each covered brand of conferences.

2.2 Related work

In the field of High-Energy Physics there has been a movement concerning long time preservation of publications, datasets, repositories and relations between them [GMH\textsuperscript{+}09, GMB10, AAA\textsuperscript{+}12, Sou13], and there is a prospering project called INSPIRE-HEP at \url{http://inspirehep.net}. It covers a different domain than software (language) engineering, but otherwise partly addresses the same problems we have pointed out. It does offer additional functionality such as job listings and does not intend to cover some of our goals such as visualisations.

ACM Digital Library in recent collaboration with IBM Watson has started to provide feature called Concept Insights. For each paper, two things can be explored: “concepts in this article” that links glossary terms mined from the full text of the paper, to their definitions on Wikipedia and “recent authors with related interests” that visualises people who recently published something that share these concepts. This functionality is certainly welcome, even though it remains to be seen how such automated concept matching can compete with and complement manual research efforts in taxonomies that try to identify key publications and tie them with key concepts and relations between them: examples exist for taxonomies of domain specific aspect languages [FDNT15], reverse engineering [CC90], reverse architecting [PDP\textsuperscript{+}07], (un)parsing [ZB14], algorithm animated visualisation [KKM06], security topics [KLS09]. Information retrieval research has also demonstrated promising results in helping to select features for automated induction [YC09, LW09] and refinement [HZL06, Nov07] of taxonomies, which we have not yet explored.

One step farther from bibliographical repositories there are model repositories such as FMI (Free Model Initiative) [SHK14], ReMoDD (Repository for Model Driven Development) [FBM\textsuperscript{+}12], CDIO (Connected Data Objects) [Ed09], Atlantic Meta-Model Zoo [At105], Grammar Zoo [Zay15], GenMy-Model [Gen14], that are on a quest of collecting models for various purposes. There are quite a number of initiatives related specifically to community management and facilitation: DBLP [Ley02], Reengineering wiki [vDV02], ResearcHR [VV\textsuperscript{+}C09], Research

2.3 Terminology and current state of BibSLEIGH

By domain we mean a top group of conferences: the front page of BibSLEIGH displays logos of its domains. Right now they are defined ad-hoc with the help of some domain knowledge; in the future we will use automated clustering techniques to form such domains. A brand is a series of events with continuing numbering and, more often than not, the same name. One event can belong in several brands: a brand of MoDELS covers the UML series because they kept the numbering, but events of the brand LDTA and ATEM belong only to the domain of SLE, but not to the brand SLE. Each proceedings entity is called an issue: usually it is regular conference proceedings issue, but it can also be a journal special issue. Multi-volume proceedings have one issue per volume because \bib\TeX{} entries for such volumes are different. A tag is a predefined term such as “context-free grammar” or “visual notation” specified as a set of matching rules covering spelling variants and synonyms (so a paper with “graphical notation” in the title will be tagged with “visual notation”). There are several style-defining tags like “question” (the title ends in a question, like “Can Programming Be Liberated from the Von Neumann Style?”), “towards” (like “Towards Incremental Execution of ATL Transformations”), “considered harmful”, “past, present and future”, etc. Interestingly, one of the most popular tags (covering around 7.2% of all papers) is “named”,
which corresponds to the pattern of starting the title with a word followed by a colon or an em-dash — like “Lilith: A Personal Computer for the Software Engineer”, or “Miranda: A Non-Strict Functional language with Polymorphic Types”, or “GHC: Operational Semantics, Problems, and Relationships with CP(↓, |)”. Currently tags are created based on titles only, because that information is indisputably in the public domain and can be used fairly; there is an ongoing discussion about fair use of abstracts and keywords, but technically they can be harvested as well, so we plan to do so (perhaps not committing the results of such harvest to public repositories to avoid copyright claims). A word is what we call a stem obtained from a classic Snowball stemmer for English. We use our own lexer that tries to split camelcased words properly: not just Camel-Case to Camel and Case, but also APIExplorer to “API” and “Explorer” and “XSDtoMOF” to “XSD”, “to” and “MOF” (it also leaves “JavaScript” intact!). Figure 1 shows a typical use of a word link. A role is some facilitating role a person has played in an issue: being an editor, a keynote speaker, a PC member, etc., are roles.

By the time of submission of this paper, BibSLEIGH covered 166 brands in 26 domains, summarised on Table 1. There are 2726 issues of these brands with 144589 papers in total. There are currently 684 tags with 354720 markings. The total vocabulary is 24359 stems derived from 1183492 words.

Table 1: Snapshot of the brands and domains currently in BibSLEIGH.

<table>
<thead>
<tr>
<th>Domain</th>
<th>Brands</th>
</tr>
</thead>
<tbody>
<tr>
<td>Applied computing</td>
<td>WICSA, ECSA, CBSE, QoSA</td>
</tr>
<tr>
<td>Components / architecture</td>
<td>ASE, CASE, DAC, DATE</td>
</tr>
<tr>
<td>Design / automation</td>
<td>DocEng, DRR, HT, ICAR, PODS, SIGMOD, TPD, JCDL, VLD</td>
</tr>
<tr>
<td>Documentation / databases</td>
<td>CSEET, ITICSE, TFP, LAK, SIGITE</td>
</tr>
<tr>
<td>Education</td>
<td>WICSA, ECSA, CBSE, QoSA</td>
</tr>
<tr>
<td>Federated computing</td>
<td>PEFM, PLDI, SAS, TOC</td>
</tr>
<tr>
<td>Formal language theory</td>
<td>ALP, CIAA, DIT, ICA, LATA</td>
</tr>
<tr>
<td>Formal methods</td>
<td>FM, IFM, SEFM, SFM, VDM</td>
</tr>
<tr>
<td>Functional</td>
<td>ACP, ACP, FCAP, ICFP, IFL, ILC, LFP</td>
</tr>
<tr>
<td>Graphs</td>
<td>ICGT, AGTIVE, GA, GCM, GG, GRAPHITE, GT-VMT</td>
</tr>
<tr>
<td>High level / logic</td>
<td>ALP, FLOPS, GPCE, LOPSTR, PLLP, PDDL, QAP</td>
</tr>
<tr>
<td>Human factors</td>
<td>CI, CSCW, DHM, DUX, HCD, HCl, HIMI, IDGD, LCT, OSC, SCMS, SOFTVIS, VISIO</td>
</tr>
<tr>
<td>Information systems</td>
<td>CAISE, EDOC, ICSIS</td>
</tr>
<tr>
<td>Knowledge engineering</td>
<td>CIKM, ECR, ICML, DCR, KDD, KDIR, KEOD, KMIS, KR, LSO, MLDM, RecSys,</td>
</tr>
<tr>
<td>Language engineering</td>
<td>SEKE, SIGIR, SKY</td>
</tr>
<tr>
<td>M ode lware</td>
<td>SLE, ATEM, LDTP, ASF+SDF, WAGA</td>
</tr>
<tr>
<td>Object orientation</td>
<td>Modelers, ULM, ECMFA, ICMT, AMT, BX</td>
</tr>
<tr>
<td>Product lines</td>
<td>ECOOP, Oswald, OOPSLA, PLATEAU, SPLASH, TOOLS</td>
</tr>
<tr>
<td>Programming languages</td>
<td>SPLC, PLEASE</td>
</tr>
<tr>
<td>Reliability</td>
<td>ADA Europe, HILT, SIGAda, TRIAda</td>
</tr>
<tr>
<td>Requirements</td>
<td>ICRE, RE, REFSQ</td>
</tr>
<tr>
<td>Software engineering</td>
<td>ESEC, FSE, ICSE, GITTSE</td>
</tr>
<tr>
<td>Software evolution</td>
<td>SANER, SCAM, CSAM, WCRE, ICPC, ICSME, PASTE, MIR</td>
</tr>
<tr>
<td>System software</td>
<td>ASPLOS, CC, COVC, CCG, RPCA, HPDC, ISM, LM, LCST, OSDI, PLOS, PPO, SOSP</td>
</tr>
<tr>
<td>Testing</td>
<td>CADE, CAV, CSL, FATES, FLoC, ICLP, ICST, ICTSS, IJCAR, ISTA, LICS, MBT,</td>
</tr>
<tr>
<td>Theory of software</td>
<td>SAT, SMT, TAP, TLLA, VMCAI</td>
</tr>
<tr>
<td></td>
<td>ESOP, FASE, FosSaCS, TACAS, WRLA</td>
</tr>
</tbody>
</table>

The oldest entry so far is the First International LISP Conference held in 1963 in Mexico, with attendees like John McCarthy and Marvin Minsky. It has mostly historical value, but a nice part was that it was possible to surface most of the papers and reconstruct metadata by googling and scraping. This issue is not present on DBLP.

Many mistakes in DBLP data (and sometimes in publishers’ data) were corrected because they were becoming quite apparent once automated processing began: the longest stems were words erroneously glued together; matching heuristics work reasonably well to equate different spellings of diacritical names, etc. An example of DBLP mismatch could be seen by comparing http://dblp.uni-trier.de/db/conf/edoc/edoc2007.html to http://bibtex.github.io/EDOC-2007.html: except for 10.1109/EDOC.2007.42 and 10.1109/EDOC.2007.44, all DOIs at DBLP are incorrect but fixed at BibSLEIGH. This was spotted automatically by reporting that some entries in this issue had no page information; an attempt to fix it revealed a mismatch between DBLP and IEEE Xplore. DOI information is usually reliable; we know of only one counterexample: http://doi.ieeecomputersociety.org/10.1109/TCSS.1997.624246 resolves successfully, but http://dx.doi.org/10.1109/TCSS.1997.624246 does not.

BibSLEIGH contains profiles on 150454 people,
Figure 1: A screenshot demonstrating the usefulness of stemming: an “abstract domain” is a proper tag, but “functor” is not, but we can still jump from this paper to all 17 papers that use that word and then to any of them with just another click.

Figure 2: The front page of BibSLEIGH with 26 domains
some of them might erroneously view several name-sakes as one person — no noticeable attention was devoted to this issue so far. Some scraping for roles has begun, so far we have 4154 roles, which is almost 10 times the size of the dataset of Vasilescu et al. [VSM13], but still around 5% of total work if we optimistically estimate 10 organisers and 20 PC members on average per issue. Figure 3 and Figure 4 show two examples of person profiles, with corresponding narrations in the captions. Notice how the profile is interpreted without the usual bibliometric remarks about the number of papers!

Exploring the rest is left as an exercise to the reader:

- http://bibtex.github.io — web front end
- http://github.com/slebok/bibsleigh — partially curated JSON data
- http://github.com/bibtex/bibsleigh — JSON refactorings and visualisations

3 Future directions

What makes BibSLEIGH become more than a glorified wrapper for DBLP is harvesting its domain specificity and community specificity. While keeping the automated, semi-automated and heuristic-based transformations as maintenance activities, we can continue ingraining the bibliographic entities and their groups with information relating them to one another, as well as to concepts, methods, frameworks, approaches, toolkits, datasets. Implementing various distance metrics, as well as annotating them manually or automatically with topic information can aid clustering and linking beyond traditional methods depending on the citation information. We see this as another step towards the construction of a body of knowledge for the domain of software language engineering (SLEBoK).

Expansion of the BibSLEIGH data set will continue, but not far: most interesting next steps involve strategically adding special issues and role annotations to al-
ready imported conferences. We are afraid that overly eager expansion will deprive us of the main advantage of being domain-specific. However, if we could find a way to eventually hide irrelevant parts from sight so that a user can productively focus on a reasonable subset, that could solve the problem and open the door wider for interdisciplinary growth of this project.

Navigational support at the current stage of development is already quite strong: domains, brands, tags and words let you browse through thousands of papers quite easily to find that doze that you are interested in. However, we believe this can be improved further — through adding annotations, leveraging metadata, proper visualisations, ground-based ranking and clustering, etc.

At BibSLEIGH’s webpage the project is called “facilitated browsing of scientific knowledge”. Indeed, providing interactive access to the curated annotated corpus of academic papers on programming language theory, compiler construction, metaprogramming, software evolution and analytics, refactoring and other related topics can serve as an entrance point into the research domain as well as the foundation for some metaresearch activities. Software engineering Master students at the University of Amsterdam have already started using BibSLEIGH actively in their studies.

It remains to be seen which open problems of software language engineering can this project contribute to solving [BZ15]. SLE, besides being a subdomain of software engineering, is known to be a bridging area of research, where a fair share of activities is devoted to seeking similarities between technologies and technical spaces, and to developing techniques with wide and cross-space applicability. However, even within one space reaching a point of soundly relating concepts can take substantial time and effort — consider laying relations between attribute grammars [Kos91] or between object algebras to attribute grammars [RBO14]. We will try to push BibSLEIGH towards facilitating this, and any help is welcome.

References


[CC90] Elliot J. Chikofsky and James H. Cross II. Reverse Engineering and Design Recovery:


[Nov07] Vit Nováček. Imprecise Empirical Ontology Refinement — Application to Taxonomy Acquisition. In Jorge Cardoso, José Cordeiro, and Joaquim Filipe, editors, Proceedings of the Ninth International Conference on Enterprise Information...


