

Extracting Event Metadata from Proceedings Titles

1st Wolfgang Fahl

Chair of Information Systems
RWTH Aachen University
Aachen, Germany
fahl@dbis.rwth-aachen.de

2nd Kai Eckert

Stuttgart Media University
Stuttgart, Germany
eckert@hdm-stuttgart.de

3th Christoph Lange

RWTH Aachen University & Fraunhofer FIT
Aachen, Germany
lange@cs.rwth-aachen.de

Abstract—Conferences are considered as one of the core media of scholarly communication in many research disciplines. The articles presented via such opportunities are usually published as proceedings. Generally, the titles of proceedings are composed of a limited subset of English terms that uniquely describe an event. Most proceedings title are available in digital form these days - be it as part of the publishing or when referenced e.g. in citations. Therefore, the titles of proceedings are a potential source for automatically extracting metadata about events. This paper presents a semantic parser to extract conference metadata from English titles of proceedings of scientific events. The parser we present has been designed and tested based on over 43,640 proceedings titles extracted from four complementary, high-quality sources. To improve the flexibility and simplicity of the parser, a combined parser / dictionary based approach is applied.

The implementation of the Proceedings Title Parser (PTP) is made available as an open source project at <https://github.com/WolfgangFahl/ProceedingsTitleParser>.

Index Terms—Metadata extraction, Scholarly communication Named Entity Recognition (NER), Semantic Parsing, Information Retrieval, Bibliometrics, Natural Language Processing (NLP)

I. INTRODUCTION

Conferences are considered as one of the core media of scholarly communication in many research disciplines [1]. In the careers of researchers, which heavily depend on multiple factors related to publishing and exchanging their research results, conferences play an important role. Therefore, reliable metadata about conferences is useful in many scenarios, such as finding conferences that provide relevant information to support a scholar’s research, and identifying venues for publishing research results. Our research is in the scope of the ConfIDent project¹, which aims at building a service for open information about scientific events and thus has the prerequisite of acquiring metadata about events from different sources. One possible source is the titles of conference proceedings and similar events. By extracting metadata from the proceedings titles the identification and description of events is facilitated.

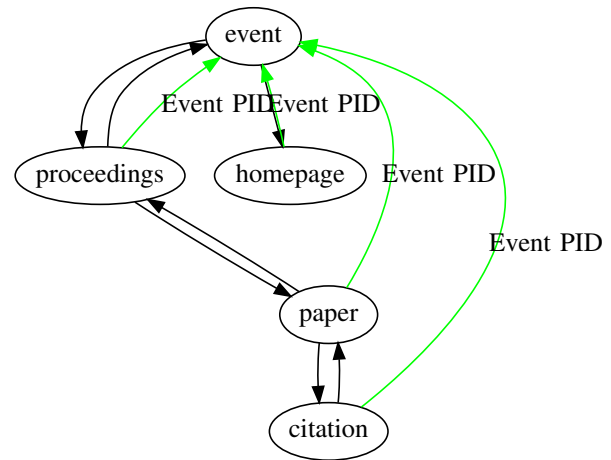


Fig. 1. Links between scholarly communication artifacts

A. Proceedings titles as a source for event metadata

A proceedings title such as “Proceedings of the 2016 ACM on Cloud Computing Security Workshop, CCSW 2016, Vienna, Austria, October 28, 2016” (see <https://dblp.org/db/conf/ccs/ccsw2016.html>) contains metadata such as the event’s acronym, location and date.

To automate the extraction of the metadata and subsequently make it available as linked open data, i.e., optimized for machine consumption and comprehensibility, the empirical findings and the ontological structure of metadata needed for the use cases need to be aligned.² The goal is to link an event’s proceedings to the event itself, and the event to its corresponding event series. In parts, such links are already available on a few platforms including dblp and Wikidata. However the distinction of an event and its proceedings might not properly done e.g. in the Scholia [3] presentation of Wikidata content. Moreover, cross-links between the state-of-the-art platforms are still missing. In particular, as pointed out in the literature [4], [5], there is not yet a standard for persistent identifiers (PIDs) for events. Overall, we aim at establishing the links depicted in figure 1.

¹<https://projects.tib.eu/en/confident/>

²We align with the Academic Event Ontology AEON (<https://github.com/tibonto/aeon/>), which is inspired by prior work such as the Scientific Event Ontology SEO [2] and aligned with multiple standard ontologies.

Erscheinungsjahr

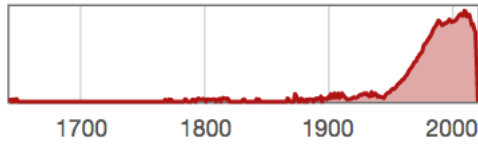


Fig. 2. TIB catalog's frequency distribution of proceedings by year

B. Limitation of current data sources

Figure 2, shows how the number of publications in proceedings has vastly increased in the digital publishing era. E.g., searching the catalog of TIB, the German National Library of Science and Technology, for entries of the category “Konferenzband”, i.e., “proceedings volume”, yields more than 600,000 results.

Digitization of library catalogs started with punched cards [6] and continued to modern online catalogs.

For the key steps of the lifecycle of a scholarly event digital traces are available for the Call for Paper (CFP), the event's homepage the accepted papers and finally the proceedings which bundle these papers. Some of these lifecycle traces are accessible via central portals like Wikicfp and others like the event homepages are scattered around the web. Libraries keep track of the papers and proceedings, projects like Wikicite try to link the different parts of the traces.

The main limitation we found in the availability of metadata from these different sources is that key references such as event and geographical context are still represented in natural language or as HTML content optimized for human readability.

An event's location, e.g., is referenced as “Vienna, Austria” instead of linking to

Vienna's Wikidata (Q1741), GND (4066009-6) or GeoNames (2761369) record.

A positive example would be <https://lobid.org> which presents GND data with proper references that allow to dereference the meta-data details.

For the notion of an event, there is not even a standard corpus one could link to and no standard Persistent Identifier (PID) to reference an event.

For papers DOIs have been established as a standard identifier and for people ORCIDs are getting more popular. Applying the same PID principle to events is one of the goals of the ConfIdent project. The Proceedings Title Parser (PTP) presented here is a tool which supports overcoming the current limitation of event data sources and corpora that do not provide a comprehensive PID based linking system yet. Extracting meta data from Proceedings titles to mine high quality event information for later disambiguation and matching is the core task of the PTP.

II. USE CASES AND TASKS

This section presents a subset of the Use cases of the ConfIdent project and the task necessary to support these

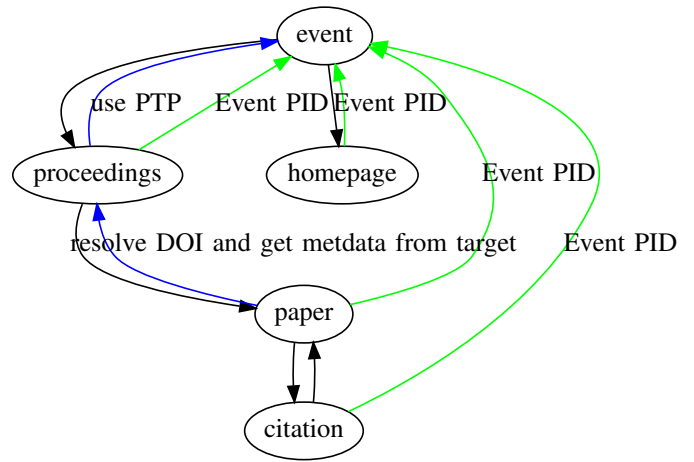


Fig. 3. Find an event via the DOI of a paper referencing it

usecases.

Our use cases value the FAIR principles [7], [8], i.e., that scientific communication artifacts should be Findable, Accessible, Interoperable and Reusable. A five star Linked Open Data (LOD) [9] level is also aimed at.

The uses cases in the following sections motivate our work and the task are needed to support these use cases. The main task is the mining of

A. Use case: Find event via DOI

Objective: given the DOI of a paper that is part of the proceedings of an event, e.g., when the paper has been cited as “in proceedings of” find the unique event.

1) Example Scenario:

- Situation: the DOI 10.1145/2362499.2362502 is given
- Action: try to find the event where the paper for this DOI has been published
- Expected Result: I-SEMANTICS 2012, Graz, Austria

The given DOI resolves to <https://dl.acm.org/doi/10.1145/2362499.2362502>. On that page the Proceedings are referenced as “Publication: I-SEMANTICS ’12: Proceedings of the 8th International Conference on Semantic Systems September 2012 Pages 9–16”.

B. Use case: Find an event via a query on the metadata attributes

Objective: given a set of metadata attributes, find the list of events (or the unique event) that match the attributes.

1) Example Scenario:

- Situation: the acronym “ICAP 2013” is given
- Action: try to find events with the given acronym
- Expected Result:
 - WikiCFP ICAP 2013: The First International Conference on Air & Space Power
 - Wikidata: Proceedings of the 3rd International Congress on Analytical Proteomics (ICAP), July 28–31, 2013, São Pedro, Brazil

The acronym “ICAP 2013” does not uniquely identify an event since the call for papers and the proceedings found do not match. The ICAP’13 Air&Space Power Conference homepage of the Turkish War Colleges is not properly linked from the wikicfp record as of 2020-10 the link <http://www.harpak.edu.tr/icap/index.html> does not work. The Wikidata entry for the Proceedings of the Analytical Proteomics event does not have a link to a corresponding event, and a page found by an internet search for the ICAP 2013 acronym which might be referencing the event such as iGroup only adds limited credibility.

C. Task: cross-check an event’s metadata against the metadata extracted from the event’s proceedings title

Objective: given an event’s metadata from a source, cross-check and verify the given set of data against the results derived from parsing the event’s proceedings title.

1) Example Scenario:

- Situation: the ICSC event series is given.
- Action: look up sources for metadata via the acronym of events and compare the retrieved metadata against each other and with metadata derived from proceedings titles.
- Expected Result: checking against dblp and OPENRESEARCH (or) should be successful, as outlined in Table II see section V-F2 for an explanation of the symbols used. For most editions of ICSC, there is no Wikidata (wd) entry. The single Wikidata ICSC 2015 proceedings title reference “Space and Situated Cognition: Proceedings of the Sixth International Conference on Spatial Cognition (ICSC 2015)” does not match. The expected ordinal in 2015 is 9 not 6.

D. Task: mine metadata for events from proceedings titles

Objective: given a source of proceedings titles, derive event metadata and match with existing event corpora to identify existing events or add new events.

1) Example Scenario:

- Situation: the Wikidata entity scholarly article (Q13442814) is given as a potential source of proceedings titles.
- Action: select articles that contain “Proceedings of” in the English label of the item. Parse selected titles with the PTP. Look up events with metadata that match the parse result in OpenResearch, dblp, ConfRef and CrossRef. Link found events or create a new event if the event was not found in any of the sources.
- Expected Result: a list of links/matches between the proceedings titles in the source document and potential existing or new events, optionally with a rating of the likelihood of the match or need for creation of a new event.

III. RESEARCH QUESTIONS

The use cases and task raise the following research questions:

- What metadata about events can be extracted from proceedings titles with what frequency?

- How can the extracted metadata be verified in the absence of a universal gold standard?
- How can deficiencies in the retrieved metadata be mitigated?

Let us elaborate each question in more detail.

A. What metadata can be extracted from proceedings titles with what frequency?

Proceedings titles seem to be created with the goal of uniquely identifying the event whose contributions shall be published. On the other hand the title shall be short. This leads to a dilemma – the shorter the title the fewer the available metadata. An empirical analysis of this question will allow to find a statistical distribution of the most commonly used syntactical and structural elements of proceedings titles.

Proceedings titles are a subset of natural language. We have limited the scope of our research to English.

Consider the following sample of a few titles of the proceedings of events located in Vienna from different sources:

- CEUR-WS: Proceedings of the 9th Transformation Tool Contest (TTC 2016), Vienna, Austria, July 8, 2016.
- crossref: Proceedings of the 20th Annual European Real Estate Society Conference – Vienna, Austria
- dblp: Proceedings of the 2016 ACM on Cloud Computing Security Workshop, CCSW 2016, Vienna, Austria, October 28, 2016
- Wikidata: [New technological capacities of the 3d millennium mammography. Proceedings of ECR-2000, Vienna, March 4-10, 2000]

These examples have in common the following structural elements typical of proceedings titles:

- Acronym: e.g., TTC2016 / CCSW2016 / ECR-2000
- Location: e.g., city: Vienna, and optionally country: Austria
- Date: e.g., July 8, 2016 / October 28, 2016 / March 4–10, 2000
- ordinal: e.g., 9th / 20th / note the special case “of the 3d millenium” where the ordinal does not refer to the event.

These and other structural elements are taken as a basis to extract the associated metadata.

Being able to uniquely identify metadata elements such as ordinal, location (city, region, country) or date (year, month, date range) allows for proper querying and lookup of events in event corpora. The lookup should lead to an attributed catalog entry that allows to verify the context. E.g. the lookup of “Vienna” leading to “Vienna, Austria” and the corresponding catalog ids and extra attributes like population may be used for verification of the lookup result.

B. How can the extracted metadata be verified in the absence of a universal gold standard?

The meta-data rich sources depicted in table I are potential event corpora gold standard candidates. By combining and disambiguating sets of event metadata to a state that is reliable and consistent enough a proper gold standard needs to be

created. Then it will be possible to derive metrics like precision and recall for queries derived from PTP results. For the time being only human inspection of samples has been performed see e.g. Table V.

IV. RELATED WORK

Our search for related work on the specific task of extracting metadata from proceedings titles did not yield an exact match. We found some references in comparable tasks of scientometrics, information extraction and named entity recognition.

A. Scientometric analysis of titles of scholarly publications

Lewison and Hartley [10] analyse how generally the length of scholarly article titles is increasing and how the use of colons increases the length of the titles based on 216513 UK and 133217 world wide oncology related articles extracted from the Science Citation Index (SCI) CD-ROM for the years 1981,1986,1991,1996 and 2001. They found an increase in the use of colons and the length of the titles. In our finding in section VI-B colons are a common but not the most common delimiter. In CEUR-WS every single title might have a colon if the "Submitted by:" information is kept. Wikicfp also systematically uses a colon to delimit the acronym of an event from the title and keeps this in the "title" metadata field e.g.

HAI 2020 : Human Agent Interaction

B. Information extraction from semi structured text

Melli [11] uses IOB/BIO tagging [12] to chunk offering titles e.g. finding identifying, product feature, -category, -brand, and -line, merchant, offering feature and functional terms for a product. The evaluation dataset has 2,437 annotated product titles. T For disambiguation the order brands, product lines, product features and products is used so that "Apple" would be considered a brand name and not a product.

C. Event metadata corpora

For the creation of the confref event corpus metadata extraction has initially been applied - there is unfortunately no publication on the details.

D. Named Entity Recognition and Linking

We evaluated the existing solutions for Named Entity Recognition (NER) and Linking as outlined in the following sections.

As an example input we used "Proceedings of the IEEE 14th International Conference on Semantic Computing, ICSC 2020, San Diego, CA, USA, February 3-5, 2020".

1) *DBpedia Spotlight*: DBpedia Spotlight [13] The DBpedia Spotlight demo annotated IEEE, USA and San Diego. The links created correctly identify the organization, country and city elements of the sample.

[Proceedings of the IEEE] [14th] [International] [Conference] on Semantic [Computing], [ICSC] 2020, [San Diego, CA], [USA], [February] 3-5, 2020

Fig. 4. OpenTapioca Result for a sample proceedings title

2) *Falcon*: Falcon [14], [15] In the ICSC 2020 example, Falcon wrongly identified "USA" as Usa (Ōita), a city in the Japanese Region Ōita. CA was linked to a song by Cole Porter. The term IEEE was linked correctly. Falcon2 linked CA to the country Canada and San Diego to the San Diego Museum of Art. The link to USA is correct. Context is not taken into account, which leads to astonishing and unexpected results.

3) *geograpy*: The geograpy ³ Python library was created in 2013 based on the Natural Language Toolkit. It is focused on the subproblem of extracting geographic information from natural language text. From the ICSC 2020 proceedings title, geograpy (original version) extracted the following city/region/country information:

```
countries=['Canada', 'Venezuela, Bolivarian
  Republic of', 'Colombia', 'United States']
places=['San Diego', 'ICSC', 'CA', 'USA',
  'International Conference', 'Semantic',
  'IEEE']
cities=['San Diego']
```

This includes the correct result but also countries. The result is still a natural language result instead of entity URIs.

4) *GROBID*: The GeneRation Of Bibliographic Data (GROBID) [16] tool allows to analyze PDF and Text documents for bibliographic content. The "Process Citation" service retrieves the following details for the ICSC 2020 sample

```
... addrLine: San Diego, CA, USA
... date type published when: 2020
... biblScope unit volume: 2020
```

in XML format. The result is only the first step in moving from Natural Language to Linked Data since only the text is extracted but not matched yet.

5) *OpenTapioca*: OpenTapioca [17] annotates text with locations, organizations and people from Wikidata. Figure 4 shows the annotation of the ICSC 2020 sample. The result for the sample contains three Wikidata items which do not correlate well with the academic event the sample is about. 14th is interpreted as a reference to a British music duo. Such an annotation is very unsatisfactory if the scientific context is clearly known or stated.

V. METHOD AND IMPLEMENTATION

A. Sources of Proceedings Titles and Event Metadata

The ConfIDest project has identified around 80 platforms for scholarly communication that might be potential sources for metadata about academic events. The focus of this paper

³<https://github.com/somnathrakshit/geograpy3>

is on sources of proceedings titles and event corpora as shown in table I.

The table shows the access method in the "via" column. The events column denotes the number of events found.

The proceedings titles were collected as plain text files from the sources as directly as possible. The getsamples shell script which. Each event is represented as a single line containing the title and source-specific ID, as shown below.

The relevant data from the sources and event corpora has been converted to JSON in case it was not readily available in that format to unify the accessibility of the data and to facilitate analysis. The JSON files are cached for sources where the preparation of these files is time consuming or the providing API might occasionally be out of service. In a second step, the data has been converted to a tabular structure and stored in an SQLite database format.⁴

The following sections describe each source and corpus with the detailed method for retrieving proceedings titles and metadata.

1) *CEUR-WS.org*: CEUR-WS.org is a free open access publishing platform for computer science workshop proceedings. Since its inception in 1994, a focus has been put on making metadata about the proceedings available. The 2014–2016 Semantic Publishing Challenges made further attempts to add value to the metadata [18]. The proceedings titles were extracted directly from the main */index.html* page with an awk script. The result for Volume 1 looks as follows:

```
Proceedings of the KI'94 Workshop
KRDB'94, Saarbrücken, Germany,
Sept. 20-22, 1994.
Submitted by: Manfred Jeusfeld|id=Vol-1
```

Besides the Proceedings Title Parsing option we have implemented a scrape mode that will get CEURVOLACRONYM, CEURFULLTITLE and CEURLOCTIME details for a Volume from the Volumes' detail page. We looked into using GROBID as a tool for getting more metadata details but decided this to be future work.

2) *dblp*: The dblp computer science bibliography regularly makes its corpus available as XML dumps. From these, we extract proceedings titles and converted them to JSON using xq, which is part of the jq command line JSON processor toolkit. Besides the title, the fields mdate, key, editor, year, publisher, isbn, series, volume, ee and url are available in this source.

3) *GND*: The Integrated Authority File (GND) has a high level "event" entity. We use the authorities-kongress dump of and loaded it into an instance of the Apache Fuseki SPARQL server. The GND.fromRDF function converts the SPARQL query result to the selected target cache format. Acronym, variant, date, areaCode, place, topic and homepage and of course title are available this way.

⁴Originally, we intended to use a graph database or triple store, but this approach did not perform well enough, as explained at Choice of Database storage system.

TABLE I
LIST OF SOURCES FOR EVENT METADATA

Source	events	titles	acronyms	via
CEUR-WS.org	2667	2477	971	HTML/RDFa
ConfRef	37945		37713	JSON API
CrossRef	46099	11490	12149	JSON API
dblp	43976	13741	42260	XML dump
GND	14281		560	RDF, SPARQL
OpenResearch	8825		8768	SMW Ask
WikiCFP	81966		72056	HTML scraping
Wikidata articles	15951	15932	396	RDF dump

4) *OpenResearch*: OpenResearch is a public wiki for conferences. It exposes event metadata via the Semantic Media-Wiki *ask* API. Our query retrieves acronym, series, homepage, city, country, start_date and end_date.

5) *Wikidata*: The Wikidata [19] entity scholarly article (Q13442814) is labeled as "article in an academic publication, usually peer reviewed". From 2018 to October 2020, the number of instances of this entity has increased from 12.4 millions to 36.5 millions. Queries over such a high number of entries with the public Wikidata query service will likely timeout even if only trying to count the instances. (try it!)

To mitigate the problem, we tried to use our own copy of Wikidata. The Blazegraph-based copy of 2018 answers the count query in 59 seconds. The Apache Fuseki based 2020 copy needs approximately one hour. Therefore we created a Wikidata dump containing the metadata of 36510243 scholarly articles. The proceedings titles were extracted with the grep command line tool. A search for "Proceedings" yields 62,320 results; "Proceedings of" yields 27,248 results. In the RDF representation of Wikidata, one such line represents a triple:

```
<http://www.wikidata.org/entity/Q62498425>
rdfs:label "Proceedings of the The 2011
International Conference on Intelligent
Computing (ICIC 2011), August 11-14, 2011,
Zhengzhou, China"@en .
```

Titles in other languages, starting with, e.g., "Tagungsband", "Konferenzband", "Anais", "Actes", etc., were not considered given our English focus. Given the Wikidata ID, the full set of metadata for the proceedings is potentially accessible – we did not make use of that information yet.

B. Core approach

DBpedia Spotlight [13] uses the following three step approach:

- 1) Tokenization
- 2) Prefix-Tree (Tries) search
- 3) Finite Automaton usage with Aho-Corasick algorithm [20])

In our approach, the first two steps are supported by the dictionary approach outlined below. The third step uses a Python parser as the finite automaton skeleton so that a standard library can be applied.

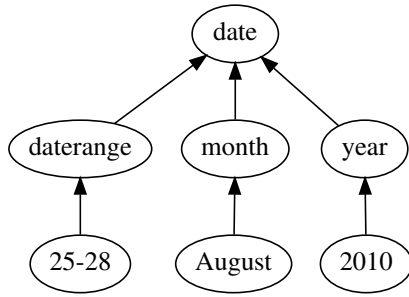


Fig. 5. Parse tree for a date reference

C. Dictionary Approach

In the first phase of our work, we applied the dictionary approach outlined in Section VI-A.

D. Shallow Semantic Parser

The next step in improving the parser is applying a proceedings titles specific grammar, which is implemented by an efficient finite automaton. The list of dictionary elements found with the dictionary approach has shown that it might be feasible to parse the proceedings titles by assuming that each dictionary element can be described as a grammar element. Also there are elements such date range, which were found but not handled by the dictionary approach yet. E.g., to analyze time metadata such as “25-28 August, 2010” the three grammar elements

- date range
- month
- year

would be used with the expected parse tree shown in figure 5.

In this step we added extra flexibility. We now have the option to derive the grammar element from a dictionary element or specify a rule based approach. E.g., for the dictionary element “year” it seems awkward to have a dictionary of all years that might be valid. It is much simpler to specify a grammar rule that specifies a year to be a 4 digit number. For the month both options seem to be equally valid – there are only 12 different options in theory. The dictionary approach has the advantage that it would be easier to, e.g., internationalize the parser because the parser itself could be fixed but use different dictionaries for different languages. For the date range, the grammar approach is definitely to be preferred because otherwise the dictionary would have to contain entries for a few hundred valid combinations of dates that do not otherwise add value to the dictionary.

We use the pyparsing library to describe the grammar in a human readable way right in the source code.

E. Combined Parser and Dictionary Approach

While standard parser generator tools such as ANTLR, Bison, Yacc, JavaCC need a tool chain to go from the parser description to the running parser the Python pyparsing library is capable of creating parsers at runtime. This makes it possible to integrate a dictionary to supply non-terminal

TABLE II
ICSC CONFERENCE SERIES METADATA CROSS-CHECK

#	year	city	region	country	wd	dblp	or
14	2020	San Diego	CA	USA	–	✓	✓
13	2019	New Port Beach	CA	USA	–	✓	✓
...							
9	2015	Anaheim	CA	USA	X	✓	✓
...							
1	2007	Irvine	CA	USA	–	✓	✓

symbols dynamically. This makes the efficiency of the created finite automaton available for our parsing process.

F. Demonstrator / Use case prototype implementations

At <http://ptp.bitplan.com>, we host a demonstrator.

The demonstrator is implemented in a test and relevance driven way, which means that the order of priority is along the use case scenarios being:

- standard cases: e.g., top 80-95% of scenarios starting from the most common ones
- special cases: long tail
- edge cases: rare long tail cases

The following sections describe the demonstrator’s state in respect to the use cases.

1) *Use case: find event via DOI:* try it!) which will find the event references:

- <https://dblp.org/db/conf/i-semantics/i-semantics2012.html>
- <http://portal.confref.org/list/i-semantics2012>

2) *Use case: crosscheck:* Table II was created by manually creating a list of ICSC acronyms for this event series and using the already implemented acronym lookup for human inspection of the results. try it! The following symbols are used in the table for three-state logic:

- ✓ - the meta data matches
- – - there is no entry available
- X - the meta data does not match

3) *lookup sources:* The main menu of the demonstrator has links to the GitHub repository as well as a wiki based documentation. There is also a chat channel available. The available modes of the Proceedings Parser are:

- Proceedings title parsing
- Named Entity Recognition (Search)
- Scrape mode

The proceedings title parsing mode extracts metadata from the proceedings title, aiming at uniquely identifying an event associated with the proceedings. The default behavior is to immediately try a lookup in the lookup database and display the results as a table.

The Named Entity Recognition mode allows to enter a set of search terms and initiate the lookup based on the search results.

As of 2020-10 the lookup is only done by the acronym found.

search	#	Source	Acronym	Url
Enter titles, :				
ICSC 2021	1 -	or	ICSC 2021	https://www.openresearch.org/wiki/ICSC 2021
ICSC 2020	1			
ICSC 2019	1 -	wikicfp	ICSC 2021	http://www.wikicfp.com/cfp/servlet/event.showcfp?eventid=104822
ICSC 2018	2			
ICSC 2017	2 -	or	ICSC 2020	https://www.openresearch.org/wiki/ICSC 2020
ICSC 2016	1			
ICSC 2015	2 -	dblp	ICSC 2020	https://dblp.org/db/conf/semco/icsc2020.html
ICSC 2014	2			
ICSC 2013	2 -	wikicfp	ICSC 2020	http://www.wikicfp.com/cfp/servlet/event.showcfp?eventid=87236
ICSC 2012	3			
ICSC 2011				
ICSC 2010				
ICSC 2009				
ICSC 2008				
ICSC 2007				

Fig. 6. Proceedings title parser lookup example

Figure 6 shows a typical lookup result.

Scrape mode retrieves proceedings titles and possible meta-data from a web page. Three such modes have been implemented:

- CEUR-WS – get the proceedings title of a CEUR-WS.org volume
- DOI – uses the CrossRef API for events
- WikiCFP – get event title and metadata for a call for papers from the given WikiCFP URL

VI. FINDINGS

A. Structural elements of Proceedings titles

The structural elements of proceedings titles are the potential source for metadata to be extracted.

In a first step of analysis we simply looked up the words most commonly used in the proceedings titles in a dictionary. For each word we decided what structural element it belongs to. We kept manually adding and classifying words until 50% of all words in our sample were covered. Table VII shows the result for the Wikidata source at this point.

About 1/3 of the titles reference one of the 63 city entries in the initial dictionary we used. At this time we did not care about multi-word city names (“York” could also have been a part of “New York”) or the completeness of the city dictionary. The assumption was that by adding more entries to the dictionary the coverage would increase later, but we would probably not find new structural elements in the “long tail” not covered by the word analysis yet.

About 2/3 of the titles have a reference to a month in full word form – the most common month being September. With 12 dictionary entries this style of month reference is fully covered. Other month references could still occur and might be ambiguous with enumerations such as “1., 2., ..., 12.”. We addressed this by taking the position of words in the title into consideration.

For the enumeration of events in a series we created a few hundred dictionary entries on the basis of enumeration sequences in different styles such as “1st, 2nd, 3rd, ...”, “1., 2., 3., ...”, “First, Second, Third, ...”, “I., II., III., ...”. About 1/2 of the titles in this sample have such references.

TABLE III
COVERAGE EFFECT OF ADDING MORE COUNTRY ENTRIES TO THE DICTIONARY

entries	coverage	CrossRef	CEUR-WS	dblp	Wikidata
30	mc 50%	1%	25%	78%	36%
62	1000	1%	88%	88%	40%
98	10000	1%	89%	89%	40%

TABLE IV
MOST FREQUENT DELIMITERS IN PROCEEDING TITLES

delimiter	usage	per title
space	556,672	12.756
,	111,351	2.552
:	7,443	0.171
)	6,396	0.147
'	1,800	0.041
/	1,579	0.036

Based on these findings we went on with the assumption that the structural elements shown in Table VII are the relevant ones to use as candidates for metadata extraction. With the three other sources we got comparable results.

1) *Probability distribution of structural elements*: The frequency distribution of structural elements is not a uniform distribution. E.g., it is much more likely that an event will happen in the country USA rather than in Uganda. This means that even a small set of dictionary entries already leads to a high coverage. Table III shows the effect of adding more country entries to the dictionary. In the first step, we added countries and other structural elements to the dictionary until we had reached a coverage of 50% (labeled “most common (mc) 50%” in the table) of the most common words/tokens in the proceedings titles. 30 country entries were sufficient to reach 50% coverage.

In the next step we analyzed the most common 1000 tokens for potential country entries (ignoring multi-word country names such as “United Kingdom”). This led to an addition of another 32 country entries. The effect is marginal on CrossRef and Wikidata, but the coverage of country metadata extraction from events in the CEUR-WS source increased from 25% to 88%. As a third step we analyzed the most common 10,000 tokens and found another 36 countries leading to a total of 98 country entries. The effect of this addition is now marginal over all sources. Note that the CrossRef source does not have the country information in the title (any more – it’s already been extracted to a separate meta data field). The conclusion here is that the general availability of the metadata in the source is more important than trying to increase the coverage by better extraction means.

B. Delimiters

We found that blanks and commas are the most often frequent delimiters in proceeding titles. Other delimiters are rare. Table IV shows an analysis of 43,640 proceedings titles from our sources.

TABLE V
CONFREF COUNTRY REFERENCES BY FREQUENCY

#	frequency	country
1	8775	USA
2	2922	Germany
...		
85	13	Jordan
86	12	München
87	11	Berlin
88	11	Hamburg
89	11	Korea (Republic of)

TABLE VI
AMBIGUOUS LOCATIONS NAMED VIENNA

name	country	region	pop	geoname id
Vienna	AT	9	1840573	2761369
Vienna	US	VA	15687	4791160
Vienna	US	WV	10861	4825976
Vienna	US	NY	5440	4833322
Vienna	US	GA	2973	4228440
Vienna	US	MO	610	4413085
Vienna	US	MD	271	4372341
Vienna	US	IL	?	4252025

C. Wordcount analysis

Figure 7 show the frequency distribution of word counts in proceedings titles derived from the crossref data source. The distribution indicates that the corner cases of titles with word counts much lower than 10 or much higher than 25 might lead to deficient entries based on the idea that proceedings titles try to be as short as possible and as long as necessary.

D. Incorrect Location references

Table V shows an excerpt on a query on the confref country references sorted by frequency. 34 events have country references for München, Berlin and Hamburg which are in fact city references.

E. Disambiguation of locations

Natural language location references may be ambiguous. Table VI shows 9 human settlements with the name "Vienna". To disambiguate such a location reference a proper persistent identifier for the location like the geoname id or some extra disambiguating information like the country or region is needed. A useful heuristic for disambiguation is to assume that the location with the highest population / number of events already registered at the location may be assumed to be the target of the reference. On the other hand the needed disambiguation information might not even be available as is the case for the missing population information for the Vienna Illinois entry in wikidata as of 2020-10. We have improved the geograpy library in the geograpy3 fork so that the ICSC example now correctly identifies:

```
San Diego (US-CA(California)
- US(United States))}
```

With links to the Wikidata and geonames ids. To achieve this result the wikidata and geonames corpora for locations were combined and the logic for selection and disambiguation

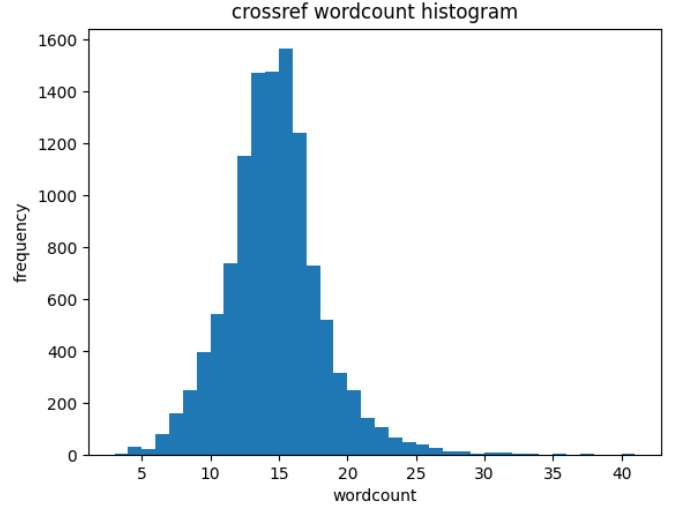


Fig. 7. Proceedings title word count histogram for source crossref

of results improved by using available context information. So Vienna on it's own will give Vienna, Austria as a result while Vienna, Illinois will select the settlement in the US. The Proceedings title parser now uses the geograpy3 library and the corresponding dictionary for the subproblem of recognizing the city-region-country part of a proceedings title.

VII. CONCLUSION

Proceedings titles generally have a pretty well defined structure that simplifies their natural language processing. The Proceedings Title Parser presented (PTP) takes advantage of the limited number of 12 relevant structural elements empirically found in a sample of over 43,600 proceedings titles from four different high quality sources. An event corpus serving as a dictionary for the lookup of the elements improves parsing quality. Even a small dictionary already reaches a coverage of over 50%. The combined parser / dictionary approach delivers higher quality results than a parser or dictionary only approach. The general availability of the metadata in the sources is more important than trying to increase the coverage by better extraction means. For the time being there is unfortunately no generally accepted Persistent Identifier (PID) system for events yet. Therefore it is not possible to identify the result by a PID yet. It is a goal of the ConfIdent project to improve this and assign PIDs to events. An example for a proper LOD link is the use of the Wikidata property P4745 "is proceedings from" which will link proceedings to the corresponding event. The Proceedings Title Parser is intended to be an essential tool for making this linking happen.

A. Limitations

The Proceedings Title Parser is currently in a minimum viable result/demonstrator state. It is a first step in the iterative development needed to provide a production ready service for the ConfIdent project.

TABLE VII
STRUCTURAL ELEMENT COVERAGE OF INITIAL DICTIONARY FOR
PROCEEDINGS TITLES FOUND IN WIKIDATA

type	# entries	found	coverage	most common examples: count
city	63	4424	(27.7%)	York: 422, London: 362, Washington: 274, Paris: 203, Tokyo: 199
country	30	5771	(36.1%)	USA: 1022, Italy: 659, Japan: 527, Germany: 501, France: 468
enum	505	7237	(45.2%)	2nd: 378, 1: 323, 3rd: 315, 2: 303, 4th: 299
eventType	20	13372	(83.6%)	Symposium: 2279, Conference: 1984, symposium: 1797, meeting: 1617, Meeting: 1578
extract	11	69	(0.4%)	In: 42, Abstract: 17, Selected: 10
field	84	10487	(65.5%)	Health: 566, Clinical: 417, Medical: 380, Cancer: 340, Medicine: 324
frequency	4	1756	(11.0%)	Annual: 969, annual: 735, Biennial: 45, Triennial: 7
month	12	10754	(67.2%)	September: 1520, June: 1308, October: 1306, May: 1100, April: 995
organization	16	6312	(39.5%)	Society: 3079, Association: 995, Societies: 489, University: 334, Group: 273
province	19	1993	(12.5%)	California: 399, Maryland: 246, Florida: 172, Massachusetts: 133, Pennsylvania: 124
publish	6	18220	(113.9%)	Proceedings: 16006, Abstracts: 777, Research: 755, Advances: 239, research: 231
scope	31	8307	(51.9%)	International: 3650, international: 606, American: 598, European: 597, British: 529
syntax	23	70187	(438.7%)	of: 25932, the: 17700, and: 7258, on: 4733, in: 3719
year	70	11745	(73.4%)	1988: 439, 1986: 412, 1997: 411, 1989: 410, 1994: 393

- The disambiguation of events is currently not done at the needed level of confidence.
- There are currently no likelihood ratings available in the presented search results.
- The search is limited to acronyms.
- The dictionary is not yet extended to the needed coverage of a reasonable part of the long-tail records with e.g. rare locations.

B. Future Work

The list of limitations outlined above calls for research to mitigate these limitations. The Proceedings Title Parser github issue list shows a list of enhancement to be implemented in the near future.

The lookup database might be improved using a Graph database and/or triplestore. This would allow for querying the dataset via SPARQL and/or GraphQL queries.

C. Mitigating deficiencies in the retrieved metadata

Campos addresses the problem of how to model and combine bodies of knowledge while maintaining an explicit representation of the unknowledge and of the conflict among the bodies [21]. Given incomplete, imprecise, contradictory, vague, non-reliable, fragmented or otherwise deficient input, how can the arising uncertainty be handled? We intend to follow the approach of assessing the plausibility and likelihood of metadata based on statistical evaluation against a corpus based on a set of highly reliable metadata. Although each source itself may still have a considerable amount of deficiencies, the combined set allows for better results than considering only a single source.

ACKNOWLEDGMENT

Part of this work has been funded by the DFG project “ConfIDent – A reliable platform for scientific events” (Grant agreements LA 3745/4-1 and SE 1827/16-1). The <http://ptp.bitplan.com> platform is kindly provided by BITPlan GmbH, Willich.

REFERENCES

- [1] M. Y. Jaradeh *et al.*, “Open research knowledge graph: Next generation infrastructure for semantic scholarly knowledge,” in *K-CAP '19, November 19–21, 2019, Marina Del Rey, CA, USA*, event: K-CAP 2019.
- [2] S. Fathalla, S. Vahdati, C. Lange, and S. Auer, “SEO: A scientific events data model,” in *The Semantic Web*, ser. Lecture Notes in Computer Science, C. Ghidini, O. Hartig, M. Maleshkova, V. Svátek, I. Cruz, A. Hogan, J. Song, M. Lefrançois, and F. Gandon, Eds., no. 11779. Cham, Switzerland: Springer Verlag, 2019, pp. 79–95.
- [3] F. Å. Nielsen, D. Mietchen, and E. L. Willighagen, “Scholia and scientometrics with wikidata,” *CoRR*, vol. abs/1703.04222, 2017. [Online]. Available: <http://arxiv.org/abs/1703.04222>
- [4] M. Ackermann *et al.*, “Global persistent identifiers for conferences and crossmark for conference proceedings,” 2018. [Online]. Available: <https://docs.google.com/document/d/1cnYoFFtasDM7H7DuRcLmw141K-1YL9C6o5SoJHKvMA/edit#heading=h.gjdgxs>
- [5] A. Birukou, “Conference identity: persistent identifiers for conferences.” [Online]. Available: <https://de.slideshare.net/birukou/conference-identity-persistent-identifiers-for-conferences>
- [6] H. Dewey, *Punched card catalogs—theory and technique*, 1959.
- [7] E. C. . G. for Research and Innovation, “Turning fair into reality: Final report and action plan from the european commission expert group on fair data.” in *European Commission – Directorate General for Research and Innovation, Brussels*, 2018.
- [8] M. Wilkinson *et al.*, “The fair guiding principles for scientific data management and stewardship,” *Sci. Data*, vol. 3, 2016.
- [9] T. Berners-Lee, *Linked Data*, <https://www.w3.org/DesignIssues/LinkedData.html> (Accessed: 2020-10).
- [10] G. Lewison and J. Hartley, “What’s in a title? numbers of words and the presence of colons,” *Scientometrics*, vol. 63, no. 2, pp. 341–356, 2005. [Online]. Available: <https://doi.org/10.1007/s11192-005-0216-0>
- [11] G. Melli, “Shallow semantic parsing of product offering titles (for better automatic hyperlink insertion),” in *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*, S. A. Macskassy, C. Perlich, J. Leskovec, W. Wang, and R. Ghani, Eds. ACM, 2014, pp. 1670–1678. [Online]. Available: <https://doi.org/10.1145/2623330.2623343>

- [12] L. A. Ramshaw and M. Marcus, "Text chunking using transformation-based learning," in *Third Workshop on Very Large Corpora, VLC@ACL 1995, Cambridge, Massachusetts, USA, June 30, 1995*, D. Yarowsky and K. Church, Eds., 1995. [Online]. Available: <https://www.aclweb.org/anthology/W95-0107/>
- [13] J. Daiber, M. Jakob, C. Hokamp, and P. N. Mendes, "Improving efficiency and accuracy in multilingual entity extraction," 2013, event: ISEM 2013.
- [14] A. Sakor, I. O. Mulang, K. Singh, S. Shekarpour, M.-E. Vidal, J. Lehmann, and S. Auer, "Old is gold: Linguistic driven approach for entity and relation linking of short text."
- [15] A. Sakor, K. Singh, A. Patel, and M.-E. Vidal, "Falcon 2.0: An entity and relation linking tool over wikidata," 2020. [Online]. Available: <https://www.researchgate.net/profile/AhmadSakor/publication/338158541Falcon20AnEntityandRelationLinkingTooloverWikidata>
- [16] P. Lopez, "Grobid: Combining automatic bibliographic data recognition and term extraction for scholarship publications," 2009, event: ECDL 2009. [Online]. Available: <https://core.ac.uk/download/pdf/38300913.pdf>
- [17] A. Delpeuch, "Opentapioca: Lightweight entity linking for wikidata," in *Proceedings of the 1st Wikidata Workshop*, 2019. [Online]. Available: <https://arxiv.org/pdf/1904.09131.pdf>
- [18] A. Dimou, S. Vahdati, A. Di Iorio, C. Lange, R. Verborgh, and E. Mannens, "Challenges as enablers for high quality linked data: Insights from the semantic publishing challenge," *PeerJ Computer Science*, 2017. [Online]. Available: <https://peerj.com/articles/cs-105/>
- [19] M. K. Denny Vrandečić, "Wikidata: a free collaborative knowledge-base," 2014. [Online]. Available: <https://doi.org/10.1145/2629489>
- [20] A. V. Aho and M. J. Corasick, "Efficient string matching: an aid to bibliographic search," 1975.
- [21] F. Campos, *Decision Making in Uncertain Situations: An Extension to the Mathematical Theory of Evidence*. Dissertation.com, 2006, ISBN: 1-58112-335-3.