Semantification of CEUR-WS with Wikidata as a target Knowledge Graph

Wolfgang Fahl[1], Tim Holzheim[1], Christoph Lange[1,2] and Stefan Decker[1,2]

[1]RWTH Aachen University, Computer Science i5, Aachen, Germany
[2]Fraunhofer FIT, Sankt Augustin, Germany

Abstract

For modern scholarly communication infrastructure, knowledge graphs are already ubiquitous. Some platforms of the publication lifecycle still struggle with catching up. In this paper, we use the example of the CEUR-WS publishing platform to show a viable approach to transition from a traditional HTML/PDF/text based environment to a single point of truth that separates the data and metadata storage from its presentation. This is possible using a public infrastructure such as Wikidata, which minimizes the maintenance effort and improves the publishing workflow.

The CEUR Workshop Proceedings (CEUR-WS) publishing platform (https://ceur-ws.org/) has been introduced in 1995 as a means to publish proceedings of scientific workshops (and smaller conferences) in computer science.

Technically HTML, PDF and a filesystem directory hierarchy are the core formats being used and delivery is via the HTTP and FTP protocols.

There have been multiple attempts in the past to make the metadata of the CEUR-WS platform available for computer based analysis and querying. None of these attempts has been consistent and continuous so far.

We report on the successful start of semantification of CEUR-WS with Wikidata as a target knowledge graph with the goal to achieve consistency and continuity for the future.

The challenge was in handling the textual natural language description parts of the CEUR-WS content that is inherently still part of the semi-digitized approach of using HTML and PDF. We propose to have a better separation of concerns of metadata, display and storage and started implementing it.

Keywords

Publishing, Information Extraction, Knowledge Graph, Linked Data, Metadata Extraction, Named Entity Linking, Named Entity Recognition, NLP, RDF, Semantification, Semantic Web, Wikidata

1. Introduction

Modern scientific publishing infrastructure uses knowledge graphs as a native basis. Google Scholar[1] is a popular and famous example. Microsoft Academic Graph [1] had the "graph" in its name. Although Microsoft terminated the service by end of 2021 the idea and content has been

Text2KG 2023: Second International Workshop on Knowledge Graph Generation from Text, May 28 - Jun 1, 2023, co-located with Extended Semantic Web Conference (ESWC), Hersonissos, Greece

fahl@dbis.rwth-aachen.de (Wolfgang Fahl); tim.holzheim@rwth-aachen.de (Tim Holzheim); lange@cs.rwth-aachen.de (Christoph Lange); decker@dbis.rwth-aachen.de (Stefan Decker)
0000-0002-0821-6995 (Wolfgang Fahl); 0000-0003-2533-6363 (Tim Holzheim); 0000-0001-9879-3827 (Christoph Lange); 0000-0001-6324-7164 (Stefan Decker)

[1]https://scholar.google.com/

picked up and extended by OpenAlex [2]. Publishing and indexing platforms that are based on older technology such as MARC [3], PICA [4] or XML struggle to catch up but gradually move to graph based approaches as DBLP did in March 2022[2] by offering SPARQL dumps that may be queried, e.g., using the QLever dblp SPARQL endpoint[3] [5].

1.1. CEUR Workshop Proceedings

CEUR-WS is a free online service that provides open access to proceedings of workshops (and smaller conferences) in the field of computer science and information technology, hosted by by RWTH Aachen University's Chair for Information Systems and Databases. CEUR-WS is operated by the CEUR-WS Editors – a team of unpaid volunteers – working as a non-profit organization. Over 3,300 Volumes containing over 65,000 PDF documents in total have been published until March 2023.

CEUR-WS does not use a Content Management System for publishing but relies on pure HTML and PDF for rendering its public website[4]. The metadata for these publications is only indirectly available by indexing services such as dblp [6] and K10plus[7][5]. Unfortunately both dblp and k10plus do not have a complete set of metadata records for all volumes as of March 2023 and in both cases there is a delay of some weeks/months before new volumes are picked up for indexing.

1.2. The Trend towards FAIR Data and Open Science: Semantification

Since inception the FAIR principles [8] have been a success. They have been adopted by various industries (e.g., the pharmaceutical industry) and national and international projects (e.g., the Common European Dataspaces). Persistent Identifiers (PIDs) and rich metadata are the core components of the FAIR principles, and they provide the means to create Knowledge Graphs [9]
.

Representing digital traces of scholarly communication in Knowledge Graphs (KGs) [10] is useful for supporting use cases such as literature search and recommendation of events for attendance or publishing. The metadata of the most relevant entities as outlined in Figure 1[6] need to be made available to offer such a knowledge graph.

The term "Semantic Web" [11] has been coined by Tim-Berners Lee et al. to describe the effect of resources on the Web interlinked making use of such metadata. Therefore we choose "Semantification" as the title of the project and this paper to describe the process of creating a Knowledge Graph and making the results available in Semantic Web fashion. The Semantification of CEUR-WS has been attempted multiple times in the past [12, 13] – always under the assumption that a local RDF/SPARQL endpoint would be the goal to achieve. The results have not been consistent and durable since the publishing workflow has not been adapted and the single point of truth for the metadata is still buried in the HTML/PDF/text documents. The

[2]https://blog.dblp.org/2022/03/02/dblp-in-rdf/
[3]https://qlever.cs.uni-freiburg.de/dblp
[4]https://ceur-ws.org
[5]https://dblp.org/, https://opac.k10plus.de
[6]The original SVG graphic is clickable and leads to the corresponding Wikidata properties and entity types

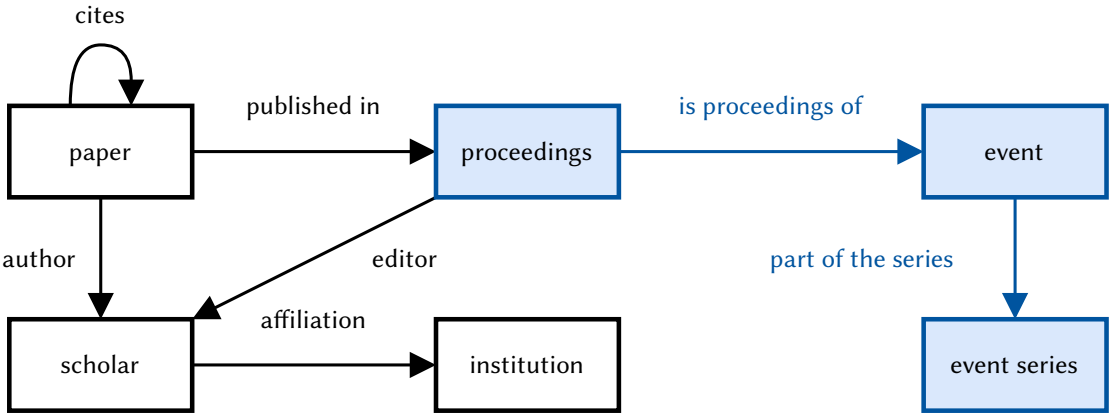

Figure 1: Most relevant entities for scientific publishing – the entities at the core of this work are blue

fear of maintenance follow-up problems, given that CEUR-WS is a non-profit service with no budget, was a major obstacle.

1.3. Challenges in making CEUR-WS more FAIR via Semantification

To semantify CEUR-WS, the following requirements where most relevant:

- the metadata should follow the FAIR [14, 8] principles
  - F1: (Meta)data are assigned globally unique and persistent identifiers (see Section 2.2).
  - F2: Data are described with rich metadata.
  - F3: Metadata clearly and explicitly include the identifier of the data they describe.
  - F4: (Meta)data are registered or indexed in a searchable resource.
- relevant queries should be supported, as derived from the original set of queries of the 2014 Semantic publishing challenge as outlined in Section 2.1
- the metadata should reuse an established ontology
- the manual and automatic curation of entries should be possible with public access for all stakeholders, e.g., editors, authors, organizers, publishers, indexers
- the infrastructure should be stable and there should be sufficient trust in its long term availability
- an open source non-profit infrastructure is preferred since this is also the mode of operation of CEUR-WS

Given the HTML/PDF/text input of CEUR-WS, we need to create the corresponding Knowledge Graph and separate the single-point-of-truth computer readable metadata from the different representations such as HTML so that the above requirements are fulfilled.

Both the HTML and PDF encoding of the original scientific content are structured for the purpose of optimizing the display / output on paper or screens; therefore, there is a structure

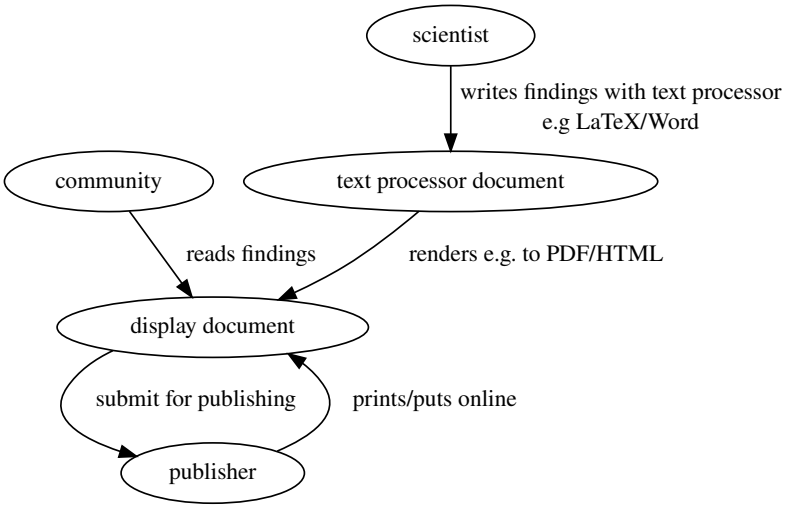

Figure 2: Textprocessing step of digital Scientific Publishing

loss compared to what was originally available in the text document processor files that the authors might have been using [15]. In Figure 2 shows the part of the publishing process where the rendering step causes this loss. Most scholars are not aware of this loss in the daily use of published content just because the documents are optimized for display and human consumption [16]. For the metadata extraction and use in knowledge graphs the difference is sometimes disastrous, e.g., when a simple text from a PDF document can not be extracted any more due to exotic styling and formatting or simply because only a scanned graphic image version of an older document is available that needs Optical Character Recognition to extract the textual content.

The metadata needed for creating the knowledge graph is only available in natural language/text form and follows rules that have been changed multiple times during the history of CEUR-WS. From 2013 to 2023, there have been 33 different versions of the index file template with no proper tracing of what was changed from version to version.

1.4. Contributions of this Paper

Our first contribution is to provide the tools, infrastructure and approaches to modify the CEUR-WS publication workflow to consistently supply FAIR high quality metadata (Semantification).

Secondly, splitting the metadata for the three main entities Event, Event Series and Proceedings is a major step towards improving the metadata quality, e.g., by allowing event series data to be checked for completeness and be completed from different sources where possible. Unfortunately, currently non of the stakeholders have the motivation to do this completion, while it is valuable, e.g., for assessing the quality of an event series. The necessary change of perspective to convince the stakeholders is a chicken-egg problem, which the CEUR-WS semantification will help to facilitate.

Provide a bootstrapping [17, 18] approach to allow for getting rid of manual editing of the CEUR-WS website content and instead using a CMS approach based on the single-point-of-truth

metadata that separates the concerns of storage and display. Making sure the results are already visible and usable during the ongoing project.

These contributions are fit to be generally applied to other publishing use cases.

2. Related Work

2.1. Semantic Publishing challenge

The Semantification of CEUR-WS has been publicly pursued by the Semantic Publishing challenge [12] from 2014. Given an excerpt of CEUR-WS, scholars were asked to prepare a RDF knowledge graph to allow for a set of 20 queries to be answered to complete Task 1.

One original task of the Semantic Publishing Challenge (SemPub2015)[7] read as follows: *Task 1: Extraction and assessment of workshop proceedings information. Participants are required to extract information from a set of HTML tables of contents, partly including microformat and RDFa annotations but not necessarily being valid HTML, of selected computer science workshop proceedings published with the CEUR-WS.org open access service. The extracted information is expected to answer queries about the quality of these workshops, ...*

Kolchin et al. [19, 20] have submitted results to the challenge twice with an approach using XPath Queries on the HTML DOM markup and converted the results directly to triples; see ceur-ws-lod repository on GitHub[8]. The reusability of this approach is limited since the parsing and generation code are intermixed.

Sateli and Witte [13] applied the GATE framework [21, 22] and a pipeline to create triples from the parsing result [23]; see also their supplementary material[9].

The 2015 work of Milicka and Burget [24] used awk text pattern matching[10] as a tool to parse the table of content files per Volume.

The objective of these attempts was to create a local SPARQL endpoint to allow to perform the required queries and fulfill the role of a target knowledge graph and point of truth.

Tasks 2 and 3 called for the detailed analysis of the PDF files.

All challenge contributions had a purely scientific focus and where not fit for making the results operational and being used in the actual CEUR-WS publishing workflow.

2.2. Persistent Identifiers in a scholarly publishing context

A study about Linked Data [25] found that each year about 10% of Linked Data URIs are no longer dereferencable. One way to mitigate the issue is to introduce persistent identifiers (PIDs), which aim to fulfill the following principles [26]: longevity, scalability, extensibility, and security. As noted in [27], the importance to ensure the longevity of PIDs is that "persistence is purely a matter of service". Thus, PIDs can only remain persistent if someone is committed to ensuring they stay accessible to users. This requires an engagement or a service level agreement for PID availability, in contrast to URIs, where no such agreement exists.

[7] https://github.com/ceurws/lod/wiki/SemPub2015

[8] https://github.com/ailabitmo/ceur-ws-lod

[9] https://www.semanticsoftware.info/sempub-challenge-2015

[10] https://github.com/FitLayout/ToolsEswc/tree/master/awk

As [9] notes, PIDs can be resolved via a URI, which follows the first principle of the Den Haag Manifesto from 2011[11]. With this alignment, Semantic Web tools, standards, and concepts to link, map, query, and integrate different data formats and data sources, and knowledge graphs become usable data based on the FAIR data principles.

Franken et al. [28] have promoted the idea of using persistent identifiers (PIDs) for scientific events in the same way as there are already persistent identifiers for papers (DOI), people (ORCID), organisation (GRID, ROR) and books (ISBN). They argue that it is also becoming more and more common practice to use PIDs to identify other important entities or objects. But as mentioned by Bryl et al. [29], it will only be beneficial if more metadata is provided and the PID is actively used to interlink with other entities.

Introducing PIDs in form of DOIs to CEUR-WS brings up the follow-up problem of who should be responsible for minting the DOIs, when the minting should be done and what the target URL of the DOI should be - not all organizers might like the landing page not to be under their own control. Wikidata Entity identifiers (Q-identifiers) seem to be a better alternative, since a rich set of other identifiers may be linked to any Wikidata Entity including DOIs, homepages and local and internationally known library and commercial and non-commercial indexing service identifiers.

2.3. Metadata Extraction from PDF, HTML and Text

As part of the Scholia Open Source Project[12], Nielsen [30] created a scraper tool capable of creating QuickStatements [31] output; see scrape/ceurws.py[13]. Using the scrape/QuickStatements chain allows for creating Wikidata entries for each paper. The Vol-3184/paper4[14] Wikidata entry has been created this way by us to show the effect. Unfortunately, the author name string (P2093)[15] property is used instead of immediately doing the disambiguate step for the author strings.

Proceedings Title Parser[16] [32] has a CEUR-WS parsing mode, which already had the RDFa extractor capability (allowing to cover more recent CEUR-WS volumes using that markup style). Part of this work has been reused and extended to a fully fledged parser in the work we are reporting here.

CERMINE (Content ExtRactor and MINEr) [33] is a software library and a web service[17] for extracting metadata and content from PDF files containing academic publications. The text content is analysed and a structured XML document containing metadata on, e.g., authors and citations created [18]. The lookup features of CERMINE are limited, e.g., to finding the ISO country code of the country of an institution an author is affiliated with.

GROBID (GeneRation Of BIbliographic Data) [34] has been gradually extended to be a machine learning library for extracting, parsing and re-structuring raw documents such as PDF into

[11] https://doi.org/10.5281/zenodo.55666
[12] https://github.com/WDscholia/scholia/issues/1438
[13] https://github.com/WDscholia/scholia/blob/master/scholia/scrape/ceurws.py
[14] https://www.wikidata.org/wiki/Q117040467
[15] https://www.wikidata.org/wiki/Property:P2093
[16] http://ptp.bitplan.com
[17] http://cermine.ceon.pl
[18] See CEUR-WS Volume 3352/Paper1.pdf example https://cr.bitplan.com/index.php/CERMINE/Example

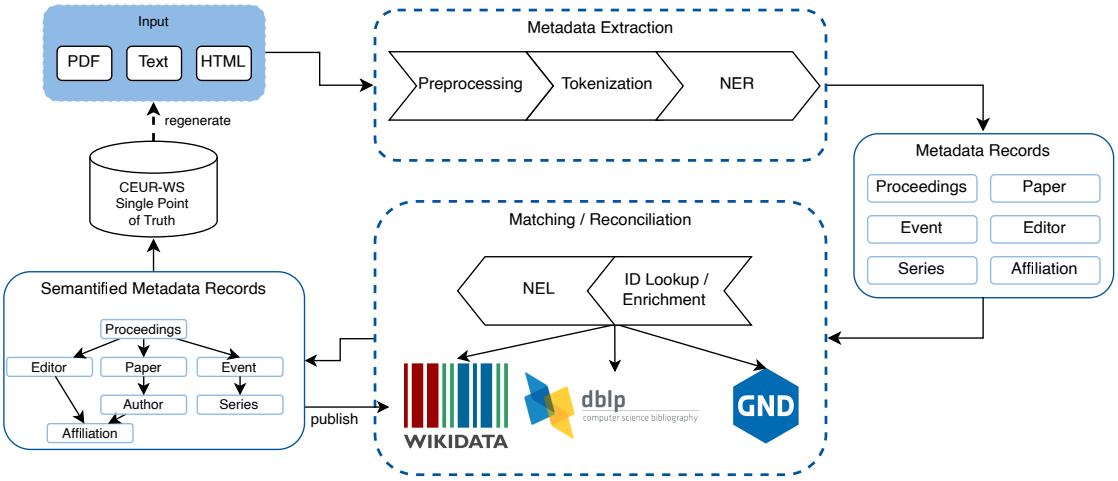

Figure 3: Workflow of the CEUR-WS Semantification

structured XML/TEI encoded documents with a particular focus on technical and scientific publications. We have tried out GROBID on over 50,000 papers so that the results are now a further potential source for disambiguation according to the original tasks 2 and 3.

3. CEUR-WS Semantification

3.1. Overview

Figure 3 shows the overview of the CEUR-WS Semantification workflow. The workflow is cyclic – it starts with what has so far been the single point of truth metadata, i.e., the ones embedded in the HTML/PDF/text of the static publication infrastructure. The Metadata Extraction step parses the input files and creates Metadata Records (which are cached as JSON records and in an SQLite relational database). These form the basis for the Matching/Reconciliation that queries Wikidata, dblp and GND with the respective SPARQL endpoints. The semantified Metadata Records are now available and may be stored in the format we see fit with JSON being the intended format for the upcoming project phase. In the next cycle the parsing of existing records is not necessary any more as long as there have been no changes. When new volumes are published, the main page, listing all proceedings volumes, is modified and HTML tables of content and PDF files are added per volume as submitted by the workshop editors.

The steps of semantification workflow for CEUR-WS as depicted in Figure 3 will be explained in the following.

3.2. Preprocessing

To extract the relevant markup elements we use the BeautifulSoup4 python library with additions to handle the RDFa-like annotations that have been applied in newer CEUR-WS volumes. The main obstacle for the extraction of the markup elements is that so far 33 different versions

have been used for the volume pages, which were often also edited manually, resulting in small differences even within the versions. The usage of the different page versions follows a long-tail Zipfian distribution with 5 versions covering 60% of all volumes.

3.3. Tokenization

3.3.1. Disambiguation using Event Signatures

As outlined in Section 2.2 PIDs would be useful for uniquely identifying scientific events. While PIDs are not available, it is necessary to use a quasi-identifier [35] consisting of a set of metadata elements that we call "Event Signature". There is neither a standardized definition of event signatures nor a recommendation for their use in references and proceedings titles. Retrieving the signature from the volume's textual description is a core step in creating the CEUR-WS knowledge graph.

As part of [28], the main author has shown that a typical scientific event "signature" consists of the following metadata (the example event being ISWC 2019[19] *The Semantic Web – ISWC 2019: 18th International Semantic Web Conference, Auckland, New Zealand, October 26–30, 2019*):

acronym a short name for the conference, often consisting of 3 to 8 upper case letters trying to be unique but actually often being ambiguous. For instance, ISWC may refer to the **International Semantic Web Conference** or to the International Symposium on Wearable Computing.

frequency annual, biennial, triennial – most events have an **annual** frequency and this is mostly **not stated explicitly**.

event reach target reach of the conference such as **international**, European, East Asian

event type such as **Conference**, Workshop, Symposium

year a two or four digit reference to the year in which the event took place – not to be confused with the year of publication of the proceedings, which might be different (**2019**)

ordinal often used to enumerate the conference series instances (**18th**)

date start date and end date or date range of the conference (**October 26–30**)

location description of the location of the conference often consisting of country, region and city – sometimes with details about the exact venue. (**Auckland, New Zealand**)

title the title often contains scope, type and subject of the conference (**International Semantic Web Conference**)

subject description what the conference is about often prefixed with "on" (**Semantic Web**)

delimiters a variety of syntactic delimiters such as blanks, comma, colon, brackets are used depending on the citation style.

[19] https://www.wikidata.org/wiki/Q48027931

Table 1
Mapping Event Signature Elements to Wikidata

property	PID	example
acronym	short name (P1813)	Text2KG 2022
title	title (P1476)	1st International Workshop on Knowledge Graph Generation From Text
event type	instance of (P13)	Workshop
date(start)	start time (P580)	May 30th, 2022 → 2022-05-30
date(end)	end time (P582)	May 30th, 2022 → 2022-05-30
location	location (P276)	Hersonissos → Chersonesos Irakliou (Q1018106)
country	country (P17)	Greece → Greece (Q41)
series	part of the series (P179)	International Workshop on Knowledge Graph Generation From Text → Workshop on Knowledge Graph Generation From Text (Q116982161)
ordinal	series ordinal (P1545)	1st
colocated with	colocated with (P11633)	ESWC 2022 → ESWC 2022 (Q110791806)
homepage	official website (P856)	https://aiisc.ai/text2kg/
URN	URN-NBN (P4109)	urn:nbn:de:0074-3184-1
publication date	publication date (P577)	2022-08-11
dblp id	DBLP publication ID (P8978)	conf/esws/2022text2kg
k10plus id	K10plus PPN ID (P6721)	1818588285

The event signature needs to be extracted from the CEUR-WS main and volume tables of content and stored as triples for the target KG.

The distinction between proceedings, event and event series needs to be made – therefore the result needs to be split and disambiguated against existing entries in the target KG.

The mapping as outlined in Table 1 has been used to map to the "event" and "proceedings" entry in Wikidata. The top rows of the table show the common properties of the proceedings and event item entries, followed by the special properties for event followed by the special properties for proceedings.

The series entry for the Text2KG example has been created and the new "colocated" property is filled for this example. The Text2KG Workshop series scholia overview[20] shows the connections.

As an example, we are using the Wikidata entry for the Text2KG@ESWC-2022[21] workshop, whose proceedings have been published as CEUR-WS Volume 3184[22] with the Wikidata proceedings item being shared by another workshop (a special but still frequent case).

[20]https://scholia.toolforge.org/event-series/Q116982161
[21]https://www.wikidata.org/wiki/Q113512465
[22]https://ceur-ws.org/Vol-3184/

3.4. Named Entity Recognition and Linking (NER & NEL)

The Named Entity Recognition (NER) and Named Entity Linking (NEL) tasks for the CEUR-WS Semantification are based on the textual input from the HTML markup that needs parsing into tokens that represent entities and then matching the textual content of the tokens against Wikidata entries that might need disambiguation. The semi-structured HTML markup helps hereby to reduce the input complexity for the parsing of the different entity types and therefore increases the accuracy of the matching process [36]. Items to disambiguate are derived from the event signature as outlined in section 3.3.1: Volumes, Papers, Editors, Authors, Locations (Country/Region/City), Dates, Ordinals, Acronym, Homepage.

For the phase of the project we are reporting, the mass creation of Proceedings and Event entries was in the focus. The Paper, Editor and Author disambiguation and Event series completion has been prepared and example results are available to show that the elements are available and may be systematically created and queried.

3.4.1. Location NER and NEL

The location of an event is described in the table of contents in *span* elements usually classified as *CEURLOCTIME*. Their value contains semi-structured information about the event's location and date range, e.g., "Hersonissos, Greece, May 30th, 2022". Besides the varying formats of the location and date definition the location information can be fairly easy separated from the date. This leaves a string that should contain information about the city and country. There exist cases were also the region or venue is named, and since more and more conferences moved to virtual meetings since 2020, the location string could also contain indications for that. Since location string can be identified on extraction Named Entity Recognition (NER) and Named Entity Linking (NEL) are done in one step to get Wikidata Qids of the mentioned locations. For the NEL we used geograpy3, a Python library that has a database with the labels of countries, regions and cities in multiple languages linking to the corresponding Wikidata Qid. The response of this label lookup is a list of possible locations of the aforementioned categories. The list is sorted by the category order city, region, country were the cities are also ordered by population. To this list we apply again a ranking since we know that in most cases the country is named within the string so we can verify that we select a city were its country was also detected. For the given example the result would then be "Hersonissos, Greece" → Chersonesos Irakliou (Q1018106)[23] (Greece (Q41)[24])

3.4.2. Editors and Authors

The author and editor name disambiguation is one of the main challenges of libraries and indexers [37]. Due to the common occurrence of duplicate names, abbreviations of first names, typos and encoding errors disambiguation is an expensive and error prone task if high accuracy is aimed.

In CEUR-WS, the editor information is given in an HTML element containing the editor signature usually given name and family name with a reference to the affiliations as shown in

[23] https://www.wikidata.org/wiki/Q1018106
[24] https://www.wikidata.org/wiki/Q41

```
<b> TEXT2KG edited by </b>
</p><h3>
   <span class="CEURVOLEDITOR">Sanju Tiwari</span> ... 1
   <span class="CEURVOLEDITOR">Nandana Mihindukulasooriya</span> ... 2
   <span class="CEURVOLEDITOR">Francesco Osborne</span> ... 3 4
   <span class="CEURVOLEDITOR">Dimitris Kontokostas</span> ... 5
   <span class="CEURVOLEDITOR">Jennifer D'Souza</span> ... 6
   <span class="CEURVOLEDITOR">Mayank Kejriwal</span> ... 7
</h3>
```

Listing 1: CEUR-WS volume page HTML markup excerpt of editor definition

Listing 1. For newer publications ORCIDs of Authors and Editors might be directly available from the PDF input. For older volumes, only the plain name, affiliation and no identifier is provided that could simplify the disambiguation. Fortunately, around 80% of the volumes are indexed at dblp. dblp provides high quality disambiguated data about the proceeding editors and paper authors also accomplished through manual curation [38]. Therefore, extracting the editors and resolving the name to an identifier by looking up the dblp id seems to be the best option. The same strategy as used for the editors also applies to the authors but here the affiliation needs to be extracted from the paper PDF.

3.5. ID Lookup/Enrichment

3.5.1. dblp and k10plus Volume matching and linking

dblp and k10plus records may be trivially matched by the unique identifier volume number of a proceeding combined with the URN of the CEUR-WS proceedings series.

We are using the QLever dblp SPARQL endpoint mentioned in Section 1 to match CEUR-WS volumes against dblp entries by volume number.

For the k10plus matching we use the catmandu library which allows to query the PICA [4] based k10plus database for URN matches see the PPN/Volume/WikidataItem matching SPARQL Query[25].

Based on the resolved IDs from the ID lookup and NEL, we can now enrich our data by querying additional or missing data. For example in Volume 3356[26], only "Tokyo" is defined as location; linking the string to Tokyo (Q1490)[27] then enables querying for the missing country information. Similar enrichments are done for editors and authors to complete the records.

3.6. Decision to use Wikidata as a target

Wikidata [39] is a knowledge graph based on an RDF triple store that has been successfully used to gather and link metadata of scholarly communication artifacts [30].

[25] https://w.wiki/6Qm5
[26] https://ceur-ws.org/Vol-3356/
[27] https://www.wikidata.org/wiki/Q1490

In 2022, we decided to directly target Wikidata instead of trying to set up our own RDF/SPARQL endpoint as outlined in Section 2.1. Wikidata is well suited to to handle the challenges listed in Section 1.3.

3.7. Workshop colocated with conference

Most CEUR-WS Volumes are proceedings of workshops, whose majority is colocated with a conference. For this "colocation" relation there was no specific property in Wikidata when we started the semantification. Kolchin [20] had already pointed out that "BIBO doesn't have an event is part of bigger event semantics" in 2015 so the need was long known. We initiated the creation of P11633 (colocated with)[28] by starting a property proposal[29] according to the Wikidata's property proposal process[30] which states "When after some time there are some supporters, but no or very few opponents, the property is created by a property creator or an administrator."

One further criterion was that at least three examples need to be supplied. Unfortunately we did present a few dozen examples but not in the format expected which held up the process by a few months.

After that there was a lively and productive discussion that lead to the clarification that the property is asymmetric. The property was considered as highly relevant and well defined. After almost a year of preparation and discussion the Property is now available and shall be used in the future.

4. Implementation and Demos

4.1. Open Source

The Python library for the CEUR-WS semantification including the source code for the CEUR-WS Volume browser[31] is available as open source[32].

A prototype for the presentation[33] of the CEUR-WS Semantification results has been created using Semantic MediaWiki [40] which is using the same open source Platform as Wikipedia but with extensions for markup that is transformed to RDF triples, which leads to "Semantification" of the Wiki.

A GitHub project for the single-point-of-truth metadata handling and conversion to different representations has been started at ceurws/ceur-spt[34].

Further background research material is supplied via the Semantic MediaWikis for Wolfgang Fahl's PhD[35] (public) and the ConfIDent requirements wiki[36] (access on request).

[28] https://www.wikidata.org/wiki/Property:colocated with
[29] https://www.wikidata.org/wiki/Wikidata:Property_proposal/colocated_with
[30] https://www.wikidata.org/wiki/Wikidata:Property_proposal
[31] http://ceur-ws-browser.bitplan.com/
[32] https://github.com/WolfgangFahl/pyCEURmake
[33] http://ceur-ws.bitplan.com
[34] https://github.com/ceurws/ceur-spt
[35] https://cr.bitplan.com/index.php/Category:Text2KG
[36] http://rq.bitplan.com

4.2. CEUR-WS Volume Browser

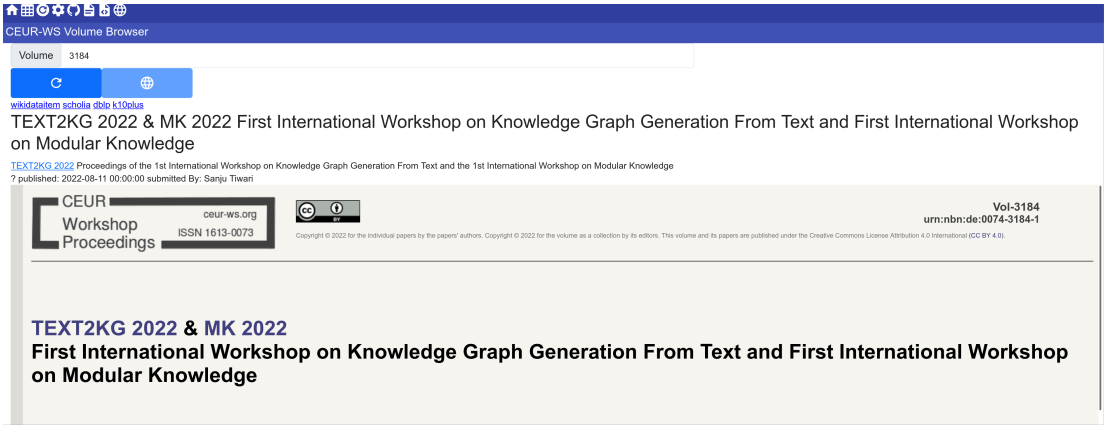

Figure 4: Screenshot of CEUR-WS Volume Browser

Figure 4 shows a screen shot of the CEUR-WS Volume Browser which we created as a means to support semantification tasks such as transferring meta-data of recently added volumes to Wikidata as well as showing the available index entries for volumes that have been published for a few weeks/months already. The example shown is for CEUR-WS Volume 3262 Wikidata Workshop 2022[37].

Figure 5: CEUR-WS Volume Browser enlarged details with links to external KGs

Figure 5 shows an enlarged section of the screenshot where the links between the proceedings volumes and the external knowledegraphs are presented. For the example these are wikidataitem[38], dblp[39], k10plus[40], and the scholia links to the proceedings, event and event series[41] (which has links to the event and proceedings).

4.3. CEUR-WS Semantic MediaWiki

The CEUR-WS Semantic MediaWiki is available as a prototype as depicted in Figure 6 that shows how a content management system approach may be applied to the metadata, which

[37] https://ceur-ws.org/Vol-3262/
[38] http://www.wikidata.org/entity/Q115053286
[39] https://dblp.org/db/conf/semweb/wikidata2022
[40] https://opac.k10plus.de/DB=2.299/PPNSET?PPN=1830580760
[41] https://scholia.toolforge.org/event-series/Q106429025

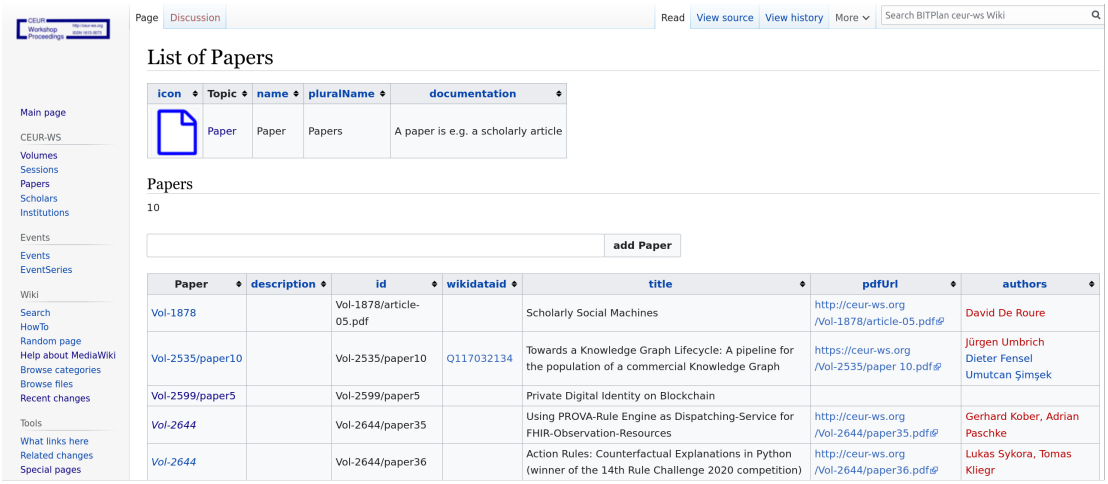

Figure 6: List of Papers in the CEUR-WS Semantic MediaWiki

allows for new features such as full text search. Semantic MediaWiki is a useful prototyping tool, since it allows to try out semantic properties and relations that are not yet fit for full public exposure via Wikidata. The example screenshot shows how a MediaWiki displays links with non-existing targets in red, allowing to judge the coverage of the disambiguation easily.

5. Results

5.1. Disambiguation

With our extraction method for editors we are able to obtain 11,764 editor signatures from 3354 volumes. Comparing these signatures against the editor records, we were able to query from dblp, it showed that 9321 signatures (4942 unique editors) can be linked to dblp and thus have at least a DBLP author id (P2456)[42]. For 2233 volumes this means that all their editors can be extracted and disambiguated to a dblp author id. But it also showed that for 387 volumes the extraction method returned fewer editors as defined in dblp with the majority of these volumes being the early 500 volumes which were manually created with a high variety in format.

With the goal to enter the editors into Wikidata, the editor signature also needs to be disambiguated to a Qid. Having the DBLP author id greatly helps in the disambiguation process as it allows to query dblp for more person identifier. This list of identifiers can than be used to query Wikidata to check if a person exists with at least on of those identifiers. The check against Wikidata showed that 1467 editors could be identified and it also showed 62 conflicting items. We found that dblp's coverage of person metadata synchronized with Wikidata is already leading to only 77 CEUR-WS editors missing. Applying the disambiguation results and linking to CEUR-WS and dblp metadata will require mass editing of Wikidata via a special CEUR-WS bot, which requires approval by Wikidata before being applicable in the next project phase.

[42] https://www.wikidata.org/wiki/Property:P2456

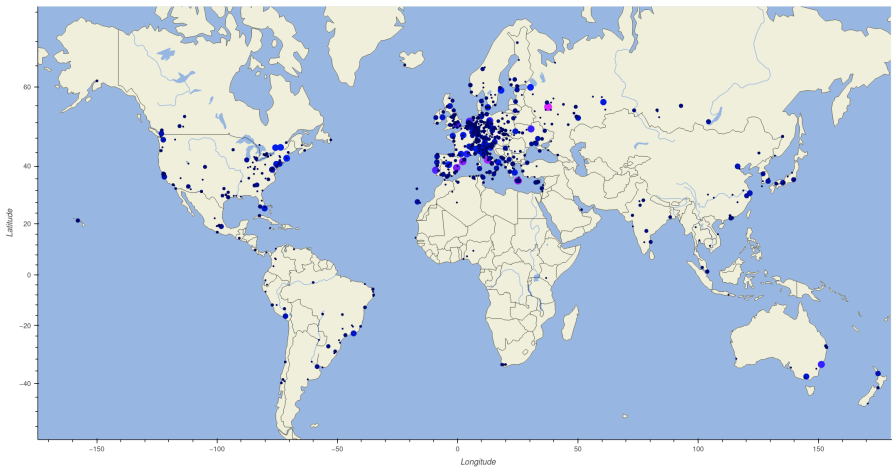

Figure 7: Locations of all CEUR-WS proceedings events (WDQS Query)

5.2. Metadata Query capability

Having the CEUR-WS metadata available in Wikidata allows for standard SPARQL queries, e.g., using the Wikidata Query Service to be applied to analyze it. Figure 4 shows a map of the distribution of event locations created with such a query. The relevance of the original set of 20 queries for Task 1 that were set as a benchmark in 2014 for different stakeholders was subjectively rated by us to sort the queries by priority[43]. The most relevant queries and the 5 queries Q1.5, Q1.12, Q1.13, Q1.16 and Q1.17 that rely on the main index have been implemented as SPARQL queries[44] that are compatible with the Wikidata Query Service endpoint to prove that our approach covers the intentions of the original challenge. Our result supplies even more capabilities given the option to do federated SPARQL queries over the connected Wikidata, dblp and k10plus knowledge graphs. The use of Wikidata ids as persistent identifiers is a core success factor here.

5.3. Further Evaluation

Making the CEUR-WS Volume metadata available on Wikidata has improved the indexing Coverage to 100% of all valid Volumes compared to 69% for k10plus and 76% for dblp.

The timeliness of the CEUR-WS metadata in Wikidata is much higher than for dblp or k10plus. For dblp it takes a few days to weeks, for k10plus it may take weeks to months before the metadata shows up. The Wikidata update may be done immediately when publishing with no delay. With the separation of event and proceeding entries it is now possible to show future events for which there are no proceedings available yet as soon as the events have been announced and later link the detailed proceedings metadata to the event record.

[43] https://cr.bitplan.com/index.php/List_of_Queries
[44] https://cr.bitplan.com/index.php/Semantic_Publishing_Challenge_Queries

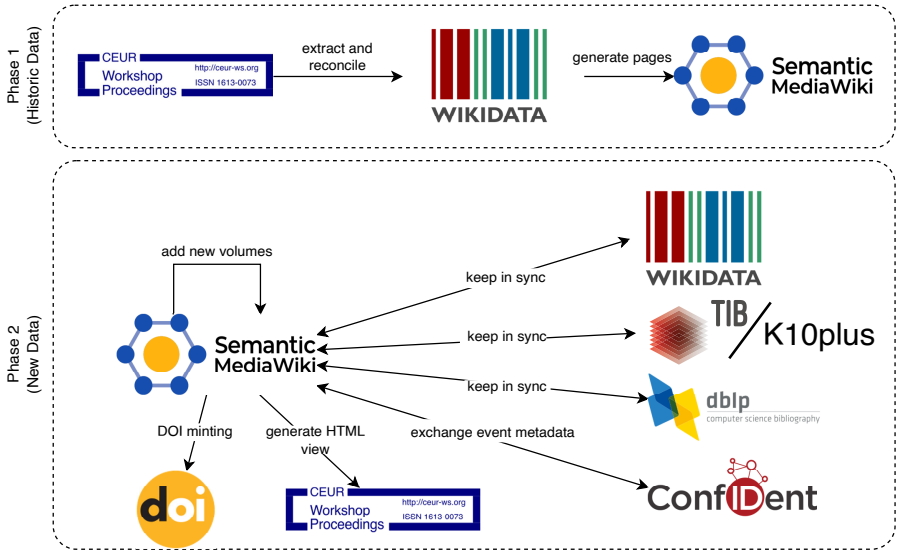

Figure 8: CEUR-WS Semantification

6. Conclusion

We have presented the first steps of CEUR-WS Semantification that result in the metadata of CEUR-WS Volumes being available in Wikidata. The linking of the relevant entities for workshops, the conferences these workshops might be colocated with, the event series the workshops and events might form as well as the linking to editor, author and paper entries and the affiliated institutions has been prototyped.

The cross-linking with dblp and k10plus has been performed and may now be continuously applied in the future.

Given that all four involved meta data sources – CEUR-WS, Wikidata, dblp and k10plus – have a lot of manual curation involved, data quality errors which derive from human errors still have to be mitigated with the goal to achieve a lower error rate than would be possible with manual efforts alone.

6.1. Future Work

Figure 8 shows an overview of a possible new approach to publishing via CEUR-WS. The core idea is to separate the concerns for displaying the content (static HTML/Semantic MediaWiki) from the storage of the metadata in a knowledge graph, e.g., Wikidata.

The new publishing workflow shall be based on single-point-of-truth metadata that is kept in computer readable format such as JSON and generate the HTML presentation from this metadata.

CEUR-WS papers do not have DOIs assigned to them during the publishing process. Assigning DOIs to papers is a feature much asked for by workshop organizers these days which CEUR-WS did not supply in the past. The new approach/architecture would simplify the DOI minting since the necessary metadata is a subset of the metadata we intend to provide anyway.

The first phase shows the current state of the workflow in the prototype phase which we report on in this paper, while the second phase is the goal of the "Semantification" project hat has been officially started in February 2023 by the CEUR-WS Editors team.

Acknowledgements. Main open source libraries being used by the work described: BeautifulSoup4[45], Catemandu[46], justpy[47], geograpy3[48] py-yprinciple-gen[49]

We would like to thank Jakob Voß for helping with the k10plus matching and creating the Wikidata property colocated with (P11633)[50] in due time.

This paper is dedicated to the memory of CEUR-WS board member Ralf Klamma † January 2023.

This research has been partly funded by a grant of the Deutsche Forschungsgemeinschaft (DFG). [51]

References

[1] K. Wang, Z. Shen, C. Huang, C.-H. Wu, Y. Dong, A. Kanakia, Microsoft Academic Graph: When experts are not enough, Quantitative Science Studies 1 (2020) 396–413. URL: https://doi.org/10.1162/qss_a_00021. doi:10.1162/qss_a_00021.

[2] J. Priem, H. Piwowar, R. Orr, OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts, 26th International Conference on Science, Technology and Innovation Indicators (STI 2022) (2022). URL: https://zenodo.org/record/6936227. doi:10.5281/ZENODO.6936227.

[3] J. Ganseman, Refactoring a library's legacy catalog: a case study, in: IAML Congress 2015, 2015. URL: http://wiki.muziekcollecties.be/images/IAML2015_JG.pdf.

[4] L. Costers, The pica catalogue system – paper 26, in: Proceedings of the IATUL Conferences 1979, Purdue University, 1979, pp. 73–77. URL: https://docs.lib.purdue.edu/iatul/1979/papers/26/.

[5] H. Bast, B. Buchhold, QLever, in: Proceedings of the 2017 ACM Conference on Information and Knowledge Management, ACM, 2017, pp. 647–656. URL: https://doi.org/10.1145/3132847.3132921. doi:10.1145/3132847.3132921.

[6] M. Ley, DBLP, Proceedings of the VLDB Endowment 2 (2009) 1493–1500. URL: https://doi.org/10.14778/1687553.1687577. doi:10.14778/1687553.1687577.

[7] B. Wiermann, K10plus – Zehn Bundesländer in einem Bibliothekssystem, 2019. URL: https://blog.slub-dresden.de/beitrag/2019/03/27/k10plus-zehn-bundeslaender-in-einem-bibliothekssystem.

[8] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes,

[45] https://pypi.org/project/beautifulsoup4/
[46] https://github.com/LibreCat/Catmandu
[47] https://github.com/justpy-org/justpy
[48] https://github.com/somnathrakshit/geograpy3
[49] https://github.com/WolfgangFahl/py-yprinciple-gen
[50] https://www.wikidata.org/wiki/Property:P11633
[51] ConfIDent project; see https://gepris.dfg.de/gepris/projekt/426477583

T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, B. Mons, The FAIR guiding principles for scientific data management and stewardship, Scientific Data 3 (2016). URL: https://doi.org/10.1038/data.2016.18. doi:10.1038/sdata.2016.18.

[9] H. Cousijn, R. Braukmann, M. Fenner, C. Ferguson, R. van Horik, R. Lammey, A. Meadows, S. Lambert, Connected Research: The Potential of the PID Graph, Patterns (New York, N.Y.) 2 (2021) 100180. doi:https://doi.org/10.1016/j.patter.2020.100180.

[10] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. de Melo, C. Gutierrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier, A.-C. Ngonga Ngomo, A. Polleres, S. M. Rashid, A. Rula, L. Schmelzeisen, J. Sequeda, S. Staab, A. Zimmermann, Knowledge graphs, ACM Computing Surveys 54 (2021) 1–37. URL: https://doi.org/10.1145/F3447772. doi:10.1145/3447772.

[11] T. Berners-Lee, J. Hendler, O. Lassila, The semantic web. a new form of web content that is meaningful to computers will unleash a revolution of new possibilities., Scientific American 284 (5) (2001) 34–43. URL: https://www.scientificamerican.com/article/the-semantic-web/.

[12] C. Lange, A. Di Iorio, Semantic publishing challenge – assessing the quality of scientific output, in: Communications in Computer and Information Science, Springer International Publishing, 2014, pp. 61–76. URL: https://doi.org/10.1007/978-3-319-12024-9_8. doi:10.1007/978-3-319-12024-9_8.

[13] B. Sateli, R. Witte, Automatic construction of a semantic knowledge base from CEUR workshop proceedings, in: Semantic Web Evaluation Challenges, Springer International Publishing, 2015, pp. 129–141. URL: https://doi.org/10.1007/978-3-319-25518-7_11. doi:10.1007/978-3-319-25518-7_11.

[14] GO FAIR International Support and Coordination Office, Fair principles, 2019. URL: https://www.go-fair.org/fair-principles/.

[15] W. Fahl, The history of scientific publishing, 2023. URL: https://cr.bitplan.com/index.php/The_History_of_Scientific_Publishing.

[16] C. Yu, C. Zhang, J. Wang, Extracting body text from academic PDF documents for text mining, CoRR abs/2010.12647 (2020). URL: https://arxiv.org/abs/2010.12647. arXiv:2010.12647.

[17] T. Bardini, Bootstrapping: Douglas Engelbart, Coevolution, and the Origins of Personal Computing, Stanford University Press, USA, 2001.

[18] Christina Engelbart, About Bootstrapping , https://www.dougengelbart.org/content/view/226/269/, 2007. Online; accessed 12 March 2023.

[19] M. Kolchin, F. Kozlov, A template-based information extraction from web sites with unstable markup, in: Semantic Web Evaluation Challenge, Communications in Computer and Information Science, Springer International Publishing, 2014, pp. 89–94. URL: https://doi.org/10.1007/978-3-319-12024-9_11. doi:10.1007/978-3-319-12024-9_11.

[20] M. Kolchin, E. Cherny, F. Kozlov, A. Shipilo, L. Kovriguina, CEUR-WS-LOD: Conversion of CEUR-WS workshops to linked data, in: Semantic Web Evaluation Challenges, Springer In-

ternational Publishing, 2015, pp. 142–152. URL: https://doi.org/10.1007/978-3-319-25518-7_12. doi:10.1007/978-3-319-25518-7_12.

[21] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan, Gate: An architecture for development of robust hlt applications, in: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02, Association for Computational Linguistics, USA, 2002, p. 168–175. URL: https://doi.org/10.3115/1073083.1073112. doi:10.3115/1073083.1073112.

[22] H. Cunningham, V. Tablan, A. Roberts, K. Bontcheva, Getting more out of biomedical documents with GATE's full lifecycle open source text analytics, PLoS Computational Biology 9 (2013) e1002854. URL: https://doi.org/10.1371/journal.pcbi.1002854. doi:10.1371/journal.pcbi.1002854.

[23] B. Sateli, R. Witte, From papers to triples: An open source workflow for semantic publishing experiments, in: Semantics, Analytics, Visualization. Enhancing Scholarly Data, Springer International Publishing, 2016, pp. 39–44. URL: https://doi.org/10.1007/978-3-319-53637-8_5. doi:10.1007/978-3-319-53637-8_5.

[24] M. Milicka, R. Burget, Information extraction from web sources based on multi-aspect content analysis, in: Semantic Web Evaluation Challenges, Springer International Publishing, 2015, pp. 81–92. URL: https://doi.org/10.1007/F978-3-319-25518-7_7. doi:10.1007/978-3-319-25518-7_7.

[25] T. Käfer, A. Abdelrahman, J. Umbrich, P. O'Byrne, A. Hogan, Observing linked data dynamics, in: The Semantic Web: Semantics and Big Data, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, pp. 213–227. doi:10.1007/978-3-642-38288-8_15.

[26] K. R. Sollins, Pervasive persistent identification for information centric networking, in: Proceedings of the second edition of the ICN workshop on Information-centric networking, ACM, New York, NY, USA, 2012, pp. 1–6. doi:10.1145/2342488.2342490.

[27] J. Kunze, E. Bermès, The ark identifier scheme, 2022. URL: https://www.ietf.org/archive/id/draft-kunze-ark-36.html.

[28] J. Franken, A. Birukou, K. Eckert, W. Fahl, C. Hauschke, C. Lange, Persistent identification for conferences, Data Science Journal 21 (2022). doi:10.5334/dsj-2022-011.

[29] V. Bryl, A. Birukou, K. Eckert, M. Kessler, What's in the proceedings? combining publisher's and researcher's perspectives, in: A. García Castro, C. Lange, P. Lord, R. Stevens (Eds.), 4th Workshop on Semantic Publishing (SePublica), number 1155 in CEUR Workshop Proceedings, Aachen, 2014. URL: http://ceur-ws.org/Vol-1155#paper-01.

[30] F. Å. Nielsen, D. Mietchen, E. Willighagen, Scholia, scientometrics and wikidata, in: E. Blomqvist, K. Hose, H. Paulheim, A. Ławrynowicz, F. Ciravegna, O. Hartig (Eds.), The Semantic Web: ESWC 2017 Satellite Events, Springer International Publishing, Cham, 2017, pp. 237–259. doi:10.1007/978-3-319-70407-4_36.

[31] M. Manske, QuickStatements, 2016. URL: https://www.wikidata.org/wiki/Help:QuickStatements.

[32] W. Fahl, K. Eckert, C. Lange, Extracting event metadata from proceedings titles, 2022. URL: https://zenodo.org/record/6568728. doi:10.5281/ZENODO.6568728.

[33] D. Tkaczyk, P. Szostek, M. Fedoryszak, P. J. Dendek, Ł. Bolikowski, CERMINE: automatic extraction of structured metadata from scientific literature, International Journal on Document Analysis and Recognition (IJDAR) 18 (2015) 317–335. URL: https://doi.org/10.

1007/s10032-015-0249-8. doi:10.1007/s10032-015-0249-8.

[34] P. Lopez, GROBID: Combining automatic bibliographic data recognition and term extraction for scholarship publications, in: Research and Advanced Technology for Digital Libraries, Springer Berlin Heidelberg, 2009, pp. 473–474. URL: https://doi.org/10.1007/978-3-642-04346-8_62. doi:10.1007/978-3-642-04346-8_62.

[35] OECD, OECD Glossary of Statistical Terms, OECD, 2008. URL: https://doi.org/10.1787/9789264055087-en. doi:10.1787/9789264055087-en.

[36] M. Cochinwala, V. Kurien, G. Lalk, D. Shasha, Efficient data reconciliation, Information Sciences 137 (2001) 1–15. URL: https://www.sciencedirect.com/science/article/pii/S0020025500000700. doi:https://doi.org/10.1016/S0020-0255(00)00070-0.

[37] S. Subramanian, D. King, D. Downey, S. Feldman, S2and: A benchmark and evaluation system for author name disambiguation, in: 2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL), IEEE, 2021, pp. 170–179. URL: https://doi.org/10.1109/jcdl52503.2021.00029. doi:10.1109/jcdl52503.2021.00029.

[38] J. Kim, Evaluating author name disambiguation for digital libraries: a case of DBLP, Scientometrics 116 (2018) 1867–1886. URL: https://doi.org/10.1007/s11192-018-2824-5. doi:10.1007/s11192-018-2824-5.

[39] D. Vrandečić, M. Krötzsch, Wikidata, Communications of the ACM 57 (2014) 78–85. URL: https://doi.org/10.1145/2629489. doi:10.1145/2629489.

[40] M. Krötzsch, D. Vrandečić, Semantic MediaWiki, in: Foundations for the Web of Information and Services, Springer Berlin Heidelberg, 2011, pp. 311–326. URL: https://doi.org/10.1007/2F978-3-642-19797-0_16. doi:10.1007/978-3-642-19797-0_16.