SemPubFlow: a novel Scientific Publishing Workflow using Knowledge Graphs, Wikidata and LLMs – the CEUR-WS use case

Wolfgang Fahl [a], Tim Holzheim [a], Christoph Lange [a,b] and Stefan Decker [a,b]

[a] *RWTH Aachen University, Computer Science i5, Aachen, Germany*
E-mails: fahl@dbis.rwth-aachen.de, tim.holzheim@rwth-aachen.de, lange@cs.rwth-aachen.de, decker@dbis.rwth-aachen.de
[b] *Fraunhofer FIT, Sankt Augustin, Germany*

Editors: First Editor, University or Company name, Country; Second Editor, University or Company name, Country
Solicited reviews: First Solicited Reviewer, University or Company name, Country; Second Solicited Reviewer, University or Company name, Country
Open reviews: First Open Reviewer, University or Company name, Country; Second Open Reviewer, University or Company name, Country

Abstract. The CEUR Workshop Proceedings (CEUR-WS) platform has been pivotal in disseminating scientific workshop and conference proceedings since 1995. This paper introduces a paradigm shift towards a semantified, consistent, and FAIR (Findable, Accessible, Interoperable, and Reusable) knowledge graph, emphasizing the critical role of Single Source of Truth (SSoT) and Single Point of Truth (SPoT) in scholarly publishing and reducing the data quality responsibility burden on CEUR-WS editors. Our SemPubFlow approach modernizes the legacy pipeline of manual HTML and PDF content curation by expecting the metadata to be supplied first. It enables the public open source collection of necessary data for event series, events, proceedings, papers, editors, authors, and affiliated institutions directly by the stakeholders of a scientific event as early as possible. The traditional Extract, Transform, Load (ETL) processes that convert existing artifacts into a comprehensive knowledge graph are only needed during the transition to this workflow. The novel approach leverages Large Language Models (LLMs) and the Wikidata knowledge graph, generating the SPoT representing CEUR-WS as the SSoT. This way our methodology not only streamlines the recreation of legacy artifacts but also addresses the "long tail" problem inherent in CEUR-WS's diverse and evolving data. This paper outlines the transition strategy, avoiding a "big bang" approach, to ensure the continuity and integrity of scholarly communication. The resulting solution is efficient in attaining the necessary level of coverage, accuracy and scalability. Data protection issues can easily be overcome in this context since even the personal data is intended to be public. The advancements presented promise to enhance publication processes across various contexts, offering a blueprint for future scholarly publishing infrastructures.

Keywords: Knowledge Graph, Linked Data, Metadata Extraction, Publishing, Semantic Web, Wikidata

1. Introduction

Traditional publishing workflows often involve redundant and manual curation of data at various stages in the lifecycle of an event, each leaving distinct digital traces. Therefore the metadata is not readily available applying state of the art FAIR, SPoT and SSoT principles. These FAIR principles are outlined in sections 1.2 mandating

Table 1
Scientific Event Lifecycle stages

Stage	Stakeholders	Artifacts Created	Format	Responsibility
Announcement	Organizers, Marketing Team	Promotional materials, websites, notifications	HTML, Email	Organizers
Call for Papers (CfP)	Organizers, Authors, Reviewers	CfP documents, emails, submission portals, responses	PDF, HTML, Email	Organizers
Performing Conference	Organizers, Chairs, Presenters, Attendees	Schedules, presentations, participant lists, recordings	PDF, PPT, HTML, Video	Organizers
Proceedings Publication	Editors, Chairs, Authors	Finalized papers, indexed and archived content	PDF, HTML	Editors, Authors, Chairs
Indexing	Libraries, Indexing Services	Catalog entries, metadata records	MARC, XML, RDF	Publishers, Libraries
Referencing/Quoting	Academic Community	Citations, references in future works	BibTeX, HTML	Scholars

the metadata being Findable, Accessible, Interoperable, and Reusable (FAIR), supplying the data based on a Single Source of Truth (SSoT) that is represented by a Knowledge Graph (KG) as a Single Point of Truth (SPoT).

Single Source of Truth (SSoT): in IT and data management SSoT refers to having one reference source for data, ensuring that everyone in an organization bases decisions on the same data. This concept is particularly important in complex distributed and networked environments where data might be stored in multiple databases or systems. The goal of a SSoT, is avoid data discrepancies and redundancies, which can lead to inefficiencies and errors. SSoT is therefore a fundamental principle in database design, data integration strategies, and enterprise information management.

Single Point of Truth (SPoT): in contexts such as software engineering, project management, and digital asset management. SPoT emphasizes that all stakeholders have access to one definitive source of information, ensuring consistency and accuracy. The SPoT principle ensures that everyone uses the most current and accurate version of digital assets.

Digital Traces of Scientific Events during their lifecyle: Table 1 shows the stages of the scientific event lifecycle that cumulatively contribute to the growing digital footprint of scholarly work, reflecting the increasing exposure to the academic community.

Moving the responsibility for curating data for relevant entities (see Figure 1) to an earlier stage and thus to the event organizers, while semi-automating the process with LLM support, has several benefits. The quality of the data will increase, and its availability will be much earlier, thus influencing FAIRness—e.g., ensuring computer-readable versions of the core entities are available for downstream systems. Involving a community such as Wikidata in open source style will further enhance the result [1].

Scientific communication increasingly relies on the use of knowledge graphs, as exemplified by Google Scholar[2] and the evolution of Microsoft Academic Graph [1] into OpenAlex [2]. Legacy systems based on technologies such as MARC [3], PICA [4] or XML are also adapting, highlighted by DBLP's adoption of RDF/SPARQL, with the QLever DBLP SPARQL endpoint[3] [5] being the most up-to date.

1.1. CEUR Workshop Proceedings

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

CEUR-WS offers a free online service that provides open access to the published proceedings hosted by RWTH Aachen University's Chair for Information Systems and Databases. CEUR-WS is operated by the CEUR-WS Edi-

[1] And will create follow-up issues for the single point of truth data to be discussed later
[2] https://scholar.google.com/
[3] https://qlever.cs.uni-freiburg.de/dblp

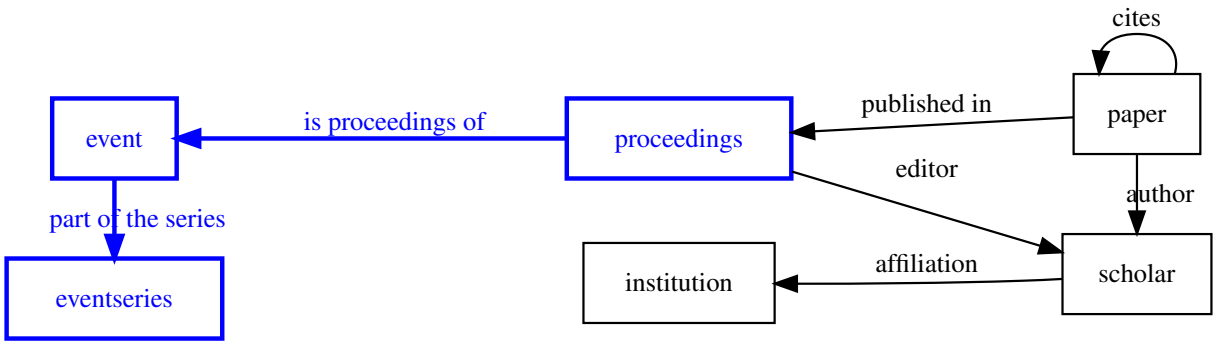

Fig. 1. Most relevant entities for scientific publishing

tors – a team of unpaid volunteers – working as a de facto non-profit organization. 3,600 Volumes containing over 66,500 PDF documents in total have been published until December 2023.

Technically, all data and metadata of the proceedings are directly represented using HTML, PDF and a filesystem directory hierarchy, and delivered via the HTTP and FTP protocols.

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 through 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 December 2023 and in both cases there is a delay of some weeks/months before new volumes are picked up for indexing.

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 – the CEUR-WS editors workflow still directly creates the HTML artifacts and still ask authors and proceeding editors to work this way as described in the CEUR-WS "how to submit" guide [8].

This work reports on the successful start of the 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.

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

Since their inception, the FAIR principles [9, 10] 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 [11].

Representing digital traces of scholarly communication in Knowledge Graphs (KGs) [12] 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 the entities at the core of this work are depicted in bold and blue.

The term "Semantic Web" [13] 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 "Semantification" was chosen as the title of the project and this paper to describe the process of creating a Knowledge Graph and making the results available in a Semantic

[4]https://ceur-ws.org
[5]https://dblp.org/, https://opac.k10plus.de
[6]The original SVG (see `http://cr.bitplan.com`) has clickable links to the Wikidata properties and entity types

Web fashion. The Semantification of CEUR-WS has been attempted multiple times in the past [14, 15] – always under the assumption that a self-maintained 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 documents often in pure text sections. The 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 were most relevant:

- the metadata should follow the FAIR [9, 10] principles
  - **Findable**
    - F1: (Meta)data are assigned globally unique and persistent identifiers (PIDs) (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.
  - **Accessible**
    - A1: (Meta)data are retrievable by their identifier using a standardised communications protocol
    - A1.1: The protocol is open, free, and universally implementable
    - A1.2: The protocol allows for an authentication and authorisation procedure, where necessary
    - A2: Metadata should be accessible even when the data is no longer available
  - **Interoperable**
    - I1: (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
    - I2: (Meta)data use vocabularies that follow FAIR principles
    - I3: (Meta)data include qualified references to other (meta)data
  - **Reusable**
    - R1: (Meta)data are richly described with a plurality of accurate and relevant attributes
    - R1.1: (Meta)data are released with a clear and accessible data usage license
    - R1.2: (Meta)data are associated with detailed provenance
    - R1.3: (Meta)data meet domain-relevant community standards
- 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, the corresponding Knowledge Graph needs to be created and the single-point-of-truth computer readable metadata be separated 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 loss compared to what was originally available in the text document processor files that the authors might have been using [16]. 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 [17]. For the

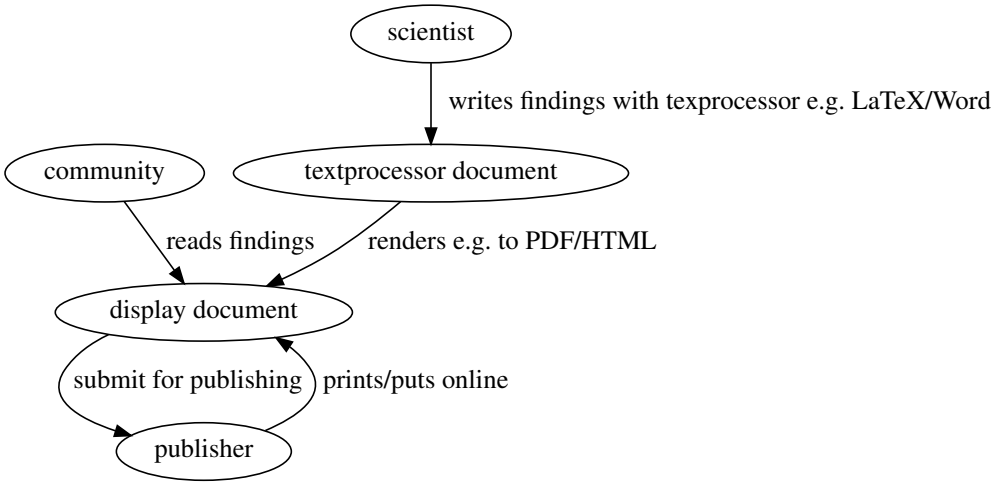

Fig. 2. Text processing step of digital Scientific Publishing

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 see table 3.

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. The index and volume HTML files were often edited manually, leading to minor, undocumented differences even within these versions. The usage of the different page versions follows a long-tail Zipfian distribution with 5 versions covering 60% of all volumes.

1.4. Contributions of this Paper

The core contributions presented in this work are:

1. To provide the tools, infrastructure and approaches to modify the CEUR-WS publication workflow to consistently supply high quality FAIR metadata (Semantification). Tools such as CEUR-WS Volume Browser, Semantic MediaWiki, and Single Point of Truth Server are detailed in section 5.
2. Splitting the metadata for the three main entities Event, Event Series and Proceedings as 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 none of the stakeholders have shown enough 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.
3. Providing a bootstrapping [18, 19] approach to get 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 CEUR-WS semantification transition project.
4. Introduction of SemPubFlow, a scientific publishing workflow that shifts from traditional manual curation to an automated tool based approach that leverages Large Language Model Systems and Wikidata. This workflow enhances the FAIRness of scholarly communications in all 15 aspects for all relevant core entities as early as possible. The benefits of this approach will predominantly manifest in the later stages of the Scientific Event Lifecycle where the high quality metadata is available for diverse usecases.

These contributions, while initially tailored for CEUR-WS have potential applicability to other publishing use cases see section 7.1.

2. Related Work

2.1. Semantic Publishing challenge

The Semantification of CEUR-WS has been publicly pursued in the Semantic Publishing challenge [14] between 2014 and 2016. From an excerpt of the CEUR-WS content, participants were asked to extract an RDF knowledge graph to allow for a set of 20 queries to be answered.

One original task of the Semantic Publishing Challenge (SemPub2015)[7] was defined 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. [20, 21] have submitted results to the challenge twice with an approach using XPath Queries on the HTML DOM markup and converted the results directly to RDF triples; see their 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 [15] applied the GATE framework [22, 23] and a pipeline to create triples from the parsing result [24]; see also their supplementary material[9].

The 2015 work of Milicka and Burget [25] used awk text pattern matching[10] as a tool to parse the table of contents files per volume.

The objective of these attempts was to create a local knowledge graph as a point of truth, which the required queries were executed against.

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

All challenge contributions had a purely scientific focus.

Without further effort, they would not have been fit for making the results operational and being used in the actual CEUR-WS publishing workflow, as doing so would have made it mandatory to operate and maintain the necessary infrastructure. For lack of resources (human as well as hardware), the CEUR-WS team did not make this effort.

2.2. Persistent Identifiers in a scholarly publishing context

A study about Linked Data [26] found that each year about 10% of Linked Data URIs become no longer dereferencable. One way to mitigate the issue is to introduce persistent identifiers (PIDs), which aim to fulfill the following principles [27]: longevity, scalability, extensibility, and security. As noted in [28], 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 that they remain accessible to users. This requires commitment or a service level agreement for PID availability, in contrast to URIs in general, where no such agreement exists.

As [11] 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 are employed to link, map, query, and integrate various data formats and sources. Consequently, knowledge graphs are rendered as usable data adhering to the FAIR data principles..

Franken et al. [29] 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

[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
[11] https://doi.org/10.5281/zenodo.55666

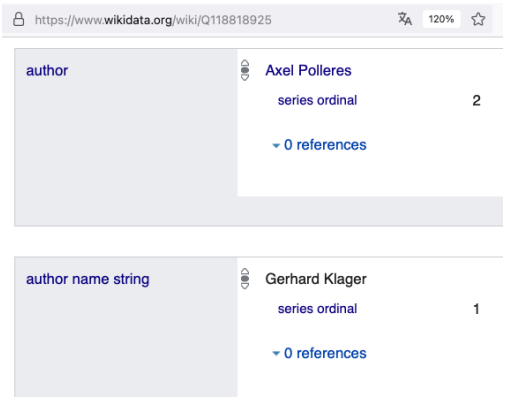

Fig. 3. author name string versus author property

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. [30], it will only be beneficial if more metadata is provided and the PID is actively used to interlink with other entities.

Introducing PIDs in the 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 workshop organizers might like that landing page not to be under their own control. Wikidata's entity identifiers (Q-identifiers) thus promise 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 [31] created a scraper tool capable of creating QuickStatements [32] output; see scrape/ceurws.py[13]. Using the scrape/QuickStatements chain enables the creation of Wikidata entries for each paper. The Vol-3184/paper4[14] Wikidata entry has been created this way by us to demonstrate the effect. In this early automation attempt, we used the author name string (P2093)[15] property, instead of immediately doing the disambiguate step for the author strings as explained in the following.

Figure 3 shows the difference in using the author (P50)[16] property. The author "Axel Polleres" already has a Wikidata entry, which is linked/clickable and all the further information of this author is available. "Gerhard Klager" is mentioned by name, which is a significant difference since the disambiguation of such entries is a major effort. In Section 4.3, we propose making the use of persistent identifiers and the immediate creation of Wikidata entries for scholars mandatory to mitigate this problem.

Proceedings Title Parser[17] [33] 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) [34] is a software library and a web service[18] for extracting metadata and content from PDF files containing academic publications. The text content is analysed and a structured XML

[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] https://www.wikidata.org/wiki/Property:P50
[17] http://ptp.bitplan.com
[18] http://cermine.ceon.pl

document containing metadata on, e.g., authors and citations is created[19]. 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) [35] has been gradually extended to be a machine learning library for extracting, parsing and re-structuring raw documents such as PDF into structured XML/TEI encoded documents with a particular focus on technical and scientific publications.

We have successfully run CERMINE on 68,524 and GROBID on 69,733 CEUR-WS papers so that the results are now a further potential source for disambiguation according to the original challenge tasks 2 and 3. The corresponding XML files are available via the CEUR-WS single point of truth server see section 5.4.

2.4. Scholarly Metadata in Wikidata

The WikiCite[20] project started in 2014 to work on citation data related to Wikipedia and Wikidata, addressing technical aspects, community building, and the future of open citations. WikiCite is aimed at creating a comprehensive bibliographic database in Wikidata to enhance all Wikimedia projects, with significant progress and community initiatives. As of 2023 the scholarly data in wikipedia is the largest subgraph see Wikidata Statistics[21] [36] with 22.6 million scholarly articles providing 31.5% of the 71.6 million instances as of 2023-12.

The Scholia[22] project [31] utilizes the scholarly data in Wikidata to provide detailed profiles for researchers, organizations, journals, publishers, individual scholarly works, and research topics. The Scholia's web portal creates on-the-fly scholarly profiles by querying the SPARQL-based Wikidata Query Service and visualizing the data in various formats. This includes lists of publications, author co-occurrence graphs, topic overlays, and more, significantly enhancing the discoverability and analysis of scholarly communication. Scholia's web portal is an open source project hosted on github[23].

3. CEUR-WS Semantification

3.1. Overview

Figure 4 gives an 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. JSON and YAML are candidate formats see table 3. 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. Currently this is done manually via the tradional workflow and semi-automatic via the single point of truth server see section 5.4.

The steps of the semantification workflow for CEUR-WS as depicted in Figure 4 are explained in the following subsections.

3.2. Preprocessing

To extract the relevant markup elements, we created parsers for the index and volume HTML Files as part of the pyCEURMake project. These parsers handle the RDFa-like annotations that have been applied in newer CEUR-WS

[19] See CEUR-WS Volume 3352/Paper1.pdf example https://cr.bitplan.com/index.php/CERMINE/Example
[20] https://www.wikidata.org/wiki/Wikidata:WikiCite
[21] https://www.wikidata.org/wiki/Wikidata:Statistics
[22] https://scholia.toolforge.org
[23] https://github.com/WDscholia/scholia

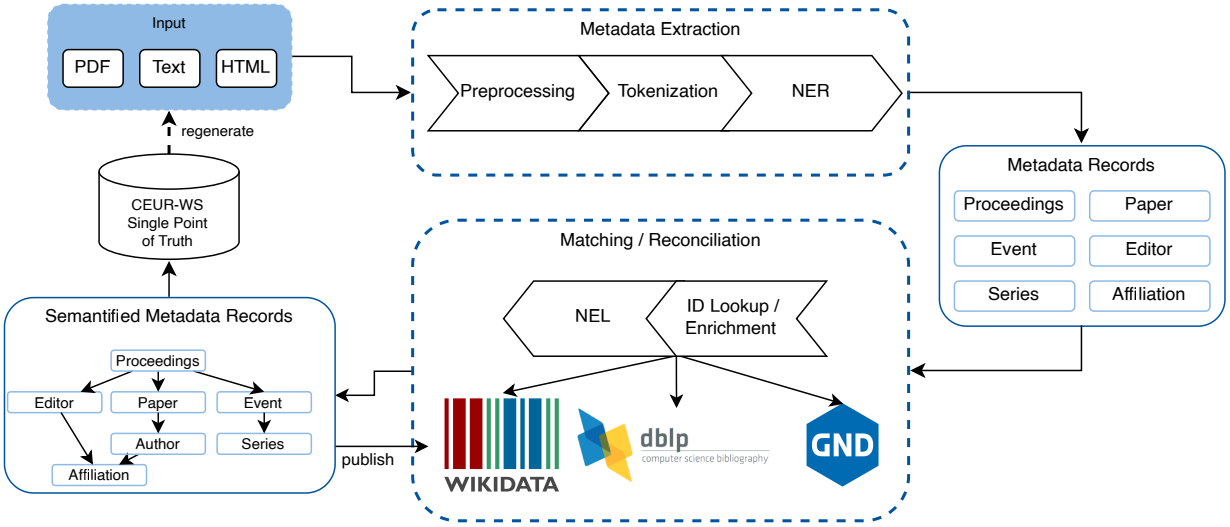

Fig. 4. Workflow of the CEUR-WS Semantification

volumes as well as the special and exotic cases arising from the long tail follow-up problems of the different over 33 styles of HTML structure being used. The BeautifulSoup4 Python library is used for lenient HTML parsing as a basis.

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. In the absence of PIDs, it is necessary to use a quasi-identifier [37] consisting of a set of metadata elements referred to as "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 [29], the first author of this article has shown that a typical scientific event "signature" consists of the following metadata (the example event being ISWC 2019[24] *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 aiming at uniqueness but actually still 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**)

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

Table 2
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

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.

The event signature needs to be extracted from the CEUR-WS main and volume tables of contents 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 2 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 events followed by the special properties for proceedings.

The series entry for the Text2KG example has been created and "colocated" property that has been added as part of this work recently is filled for this example. The Text2KG Workshop series Scholia overview[25] shows the connections.

As an example, we are using the Wikidata entry for the Text2KG@ESWC-2022[26] workshop, whose proceedings have been published as CEUR-WS Volume 3184[27] with the Wikidata proceedings item being shared by another workshop (a special but still frequent case required by the CEUR-WS submission guidelines).

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.

[25] https://scholia.toolforge.org/event-series/Q116982161
[26] https://www.wikidata.org/wiki/Q113512465
[27] https://ceur-ws.org/Vol-3184/

```
<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

The semi-structured HTML markup is simpler to handle than natural language processing of plain text since the parsing of the different entity types can be guided by structural hints such as expecting a title in an `h1` HTML tag context and therefore increases the accuracy of the matching process [38]. 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 *CEURLOC-TIME*. 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 where also the region or venue is named, and since more and more conferences have moved to virtual meetings since 2020, the location string could also contain indications for that. Since location strings 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 where the cities are also ordered by population. To this list, we once more apply a ranking, since we know that in most cases the country is named within the string, so we can select the city in the given country context. For the given example the result would then be "Hersonissos, Greece" → Chersonesos Irakliou (Q1018106)[28] (Greece (Q41)[29]).

3.4.2. Editors and Authors

The author and editor name disambiguation is one of the main challenges for libraries and indexers [39]. 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 for.

In CEUR-WS, the editor information is given in an HTML element containing the editor signature, usually given name and family name with a numeric reference to the affiliations, as shown in 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, 78% of the volumes are indexed at DBLP – less than 800 early volumes are missing. DBLP provides high quality disambiguated data about the proceeding editors and paper authors also accomplished through manual curation [40]. Therefore, extracting the editors and resolving each name to an identifier by looking up the DBLP id seems to be the best option.

[28] https://www.wikidata.org/wiki/Q1018106
[29] https://www.wikidata.org/wiki/Q41

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 in the future.

3.5. ID Lookup/Enrichment - DBLP, 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] (a library cataloging format used predominantly in German-speaking countries) based k10plus database for URN matches; for details, see the PPN/Volume/WikidataItem matching SPARQL Query[30].

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[31], only "Tokyo" is defined as location; linking the string to Tokyo (Q1490)[32] 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 [41] is a knowledge graph based on an RDF triple store that has been successfully used to gather and link metadata of scholarly communication artifacts [31].

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 as referenced below by priority:

- **Unique Identifiers (F1)**: by assigning Q-Identifiers, such as Q5 for human.
- **Rich, Identifiable Metadata (F2, F3)**: by being originally based on Wikipedia entries from multiple countries and using a very large community and bots to add current items.
- **Searchable Data (F4)**: by providing a robust Wikidata Query Service .
- **Curation and Accessibility (A1, A2, R1.2)**: by enabling community edits and offering multiple language support.
- **Complex Queries Support (A1, I1, I2, I3)**: by facilitating sophisticated querying through a SPARQL endpoint which may be federated with other Linked Open Data endpoints.
- **Ontology Reuse (I1, I2)**: by integrating and linking to external ontologies and databases and providing its own robust ontology curation process.
- **Stability and Trust (R1.3)**: by ensuring stability and trust through Wikimedia Foundation governance and having mirrored data in for profit (e.g. google) and non profit organizations[33].
- **Open Source, Non-Profit (R1.1)**: by promoting an open-source, non-profit model with Wikimedia Foundation and community contributions.

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 [21] 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)[34] by a property proposal[35] according to the

[30] https://w.wiki/6Qm5
[31] https://ceur-ws.org/Vol-3356/
[32] https://www.wikidata.org/wiki/Q1490
[33] e.g. RWTH Aachen i5 runs multiple copies of Wikidata!
[34] https://www.wikidata.org/wiki/Property:colocated with
[35] https://www.wikidata.org/wiki/Wikidata:Property_proposal/colocated_with

Format	Description
Plain text	Requires natural language processing (NLP) for metadata extraction due to its unstructured nature.
HTML	Readability varies with the structural and semantic rigor in the document. Microformats and RDFa are sometimes available.
PDF	Primarily being a collection of individual glyphs to be positioned and formatted, pose challenges in text reconstruction and structure analysis.
BibTeX	Structured nature makes it highly suitable for computer readability and metadata extraction in scholarly communications.
XML/JSON/YAML	Markup languages that balance human and machine readability with varying degrees of structure strictness.
RDF	A standard model for data interchange, facilitating data merging and schema evolution support across diverse schemas.
Multimedia (Videos, Pictures)	Visual media often require sophisticated processing such as OCR for text extraction a for content extraction.
Other (PPT, Word, SVG, etc.)	Vary in readability; structure and semantics can usually be accessed via specific APIs or tools

Table 3

Computer Readability of various formats used for scholarly artifacts

Wikidata's property proposal process[36] 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."[37].

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. SemPubFlow

4.1. Overview

SemPubFlow represents a transformative approach in the lifecycle of scientific events as outlined in Table 1, advocating for early curation and integration of publicly available machine-readable metadata separating the concerns of content and visualization.

Traditionally the created artifacts are mostly intended for human consumption and therefore optimized in this respect. Table 3 shows the computer readability challenges of some of the most common formats which often need sophisticated natural language processing capabilities to be performed. A limiting factor is the lack of standardization and separation of concerns. With the advent of LLMs the options for processing these diverse sources of metadata are vastly improving—making it feasible to get computer readable results as early as possible.

4.2. Traditional CEUR-WS Submit Workflow

CEUR-WS describe their publishing procedure in a detailed submission guide [8]. Figure 5 provides an overview of this workflow as an UML activity diagram. Even use of "PUT" in the workflow description indicates that the workflow has been in place since 1995 mostly unchanged with the idea of collecting and uploading PDF-Files and creating per Volume html pages with a single monolithic index.html to list them all.

[36] https://www.wikidata.org/wiki/Wikidata:Property_proposal

[37] 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

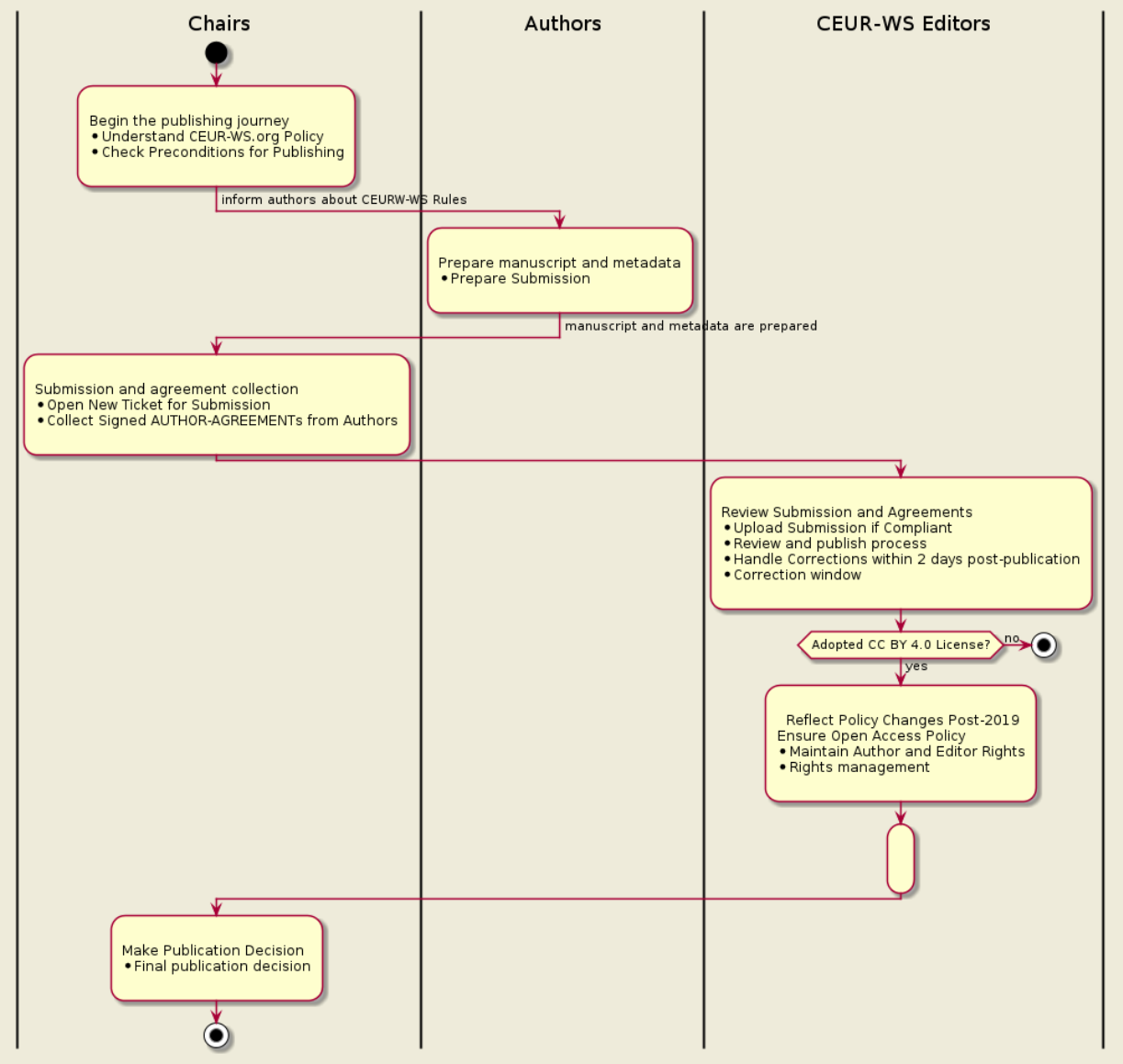

Fig. 5. CEUR-WS HOWTOSUBMIT workflow

4.3. Metadata first semantic publishing

A key goal of metadata first semantic publishing is the early use (and enforcement of this use) of persistent identifiers. Without such identifiers it is notoriously difficult to disambiguate the core entities involved in scholarly publishing.

The following PIDs are already often in use but the enforcement of their use is lacking in CEUR-WS.

- **ORCID (Open Researcher and Contributor ID):** Identifier for individual researchers.
- **ROR (Research Organization Registry):** Identifier for research organization.
- **Wikidata-ID:** A general identifier for entities in the Wikidata database, often making further identifiers available for reference.

– **DOI (Digital Object Identifier):** A general identifier for objects accessible on the internet – often used for scholarly papers.

We propose to make the use of PIDs mandatory for publishing in CEUR-WS. Given the existing index entries of CEUR-WS Volumes in DBLP and the German national library, the DBLP and GND identifiers are also candidates. For ORCIDs, RORs, DOIs there is active action required by the stakeholders to obtain these, and, in the case of DOIs, there is cost involved. Using Wikidata as the general PID Infrastructure is much more feasible since it offers a freely accessible, community-driven platform that PIDs for a wide range of entities. This approach aligns with the open and collaborative nature of scholarly communications and meets the requirements for easy integration, broad adoption, and maintenance. Using Wikidata Q-Identifiers as PIDs relies on the long term availability of Wikidata which is quite likely giving the support by a strong community and big commercial players such as google.

The Wikidata PID usage proposal calls for looking up or creating Wikidata entries for all relevant core entities as early as possible and linking these entries to the standard PIDs where applicable.

In the workflow the existing entries will later be reusable for computer assisted lookups as part of the SemPubFlow tool see section 5.5 and Scholia.

Figure 6 shows the proposed future workflow. Publication via CEUR-WS would ideally involve assigning the Volume number and transferring the metadata-rich candidate snapshot to the Single Point of Truth. This process ensures the completeness and quality of the metadata allowing for the generation of traditional HTML pages and the official stamping as a CEUR-WS proceeding.

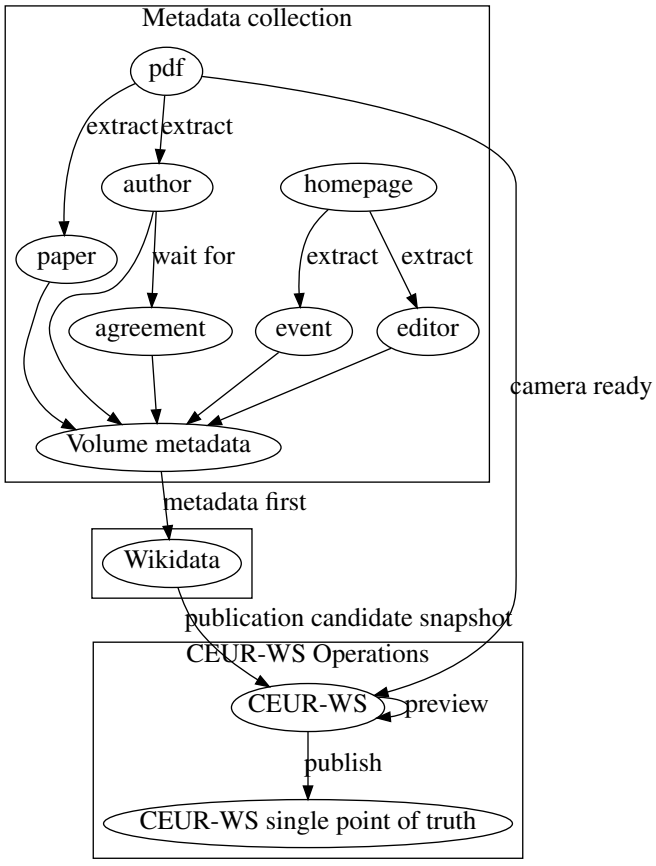

Fig. 6. Metadata First Scholarly Publishing Workflow Overview

4.4. Iterative introduction

The software engineering challenge in introducing the SemPubFlow workflow is to find a sweet spot between effort and benefit [42]. Given the long-tail distribution of issues of ever more corner and exotic cases that have shown up in the analysis of the existing data and workflow there is potentially a huge list of "small" problems that would require an inadequately high amount of effort to be solved systematically. The "sweet spot" is in solving only the most relevant problems and not fixing some of the exotic details at all or doing it manually. Therefore we will avoid a "big bang" introduction, which would replace the old workflow with the new SemPubFlow in one single big (and therefore very risky) step. Instead we propose to do smaller changes gradually. As a first step, the main index.html shall be split by years (with a link to the old legacy complete list). Then, the per-year list may be partly created using the new SemPubFlow approach. This will be continued iteratively, going from the current year back to a point where it seems not reasonable any more to spent further effort.

4.5. Legal aspects of Publishing Personal Data of Scholars

The proposed workflow calls for publishing personal data of authors and editors early in the life cycle of an event. This is perfectly legal under at least one of the following conditions:[38]

1. **Consent:** Explicit consent from individuals for their data to be published – e.g., by signing an authors agreement.
2. **Legitimate Interests:** Publication serves legitimate interests that outweighs the individual's rights and freedoms. E.g., when the scholar is a person of public interest.
3. **Public Data:** Data has already been made public by the scholar.
4. **Scientific Research:** Processing is necessary for scientific research purposes as per Article 89 of the GDPR.
5. **Contractual Necessity:** Necessary for the performance of a contract with the scholar – e.g., publishing a paper in workshop proceedings and making the metadata available for indexing by Wikidata, dblp, libraries and other interested parties.
6. **Legal Obligation:** Necessary for compliance with a legal obligation. This would, e.g., be the case if the publication was done via a traditional print outlet and thus, by German law, a copy of the proceedings has to be made available to the German National Library. On indexing, the personal data records of authors will be created as part of the German National Library's legal commitment.
7. **Vital Interests:** Necessary to protect the vital interests of the scholar or another person. This category is more relevant to medical or emergency situations and rather unlikely to be applicable for the scholarly publication use case.
8. **Public Interest:** Necessary for performing a task carried out in the public interest or in the exercise of official authority. E.g., when an investigation by public authorities is underway.

Consent, Public Data, and Contractual Necessity are crucial for scholarly publishing. Considering the analogy with Legal Obligations applying to print outlets, we expect it to be feasible to convince scholars that publishing their necessary personal data is reasonable in digital contexts.

5. Implementation and Demos

5.1. Open Source

The Python library for the CEUR-WS semantification including the source code for the CEUR-WS Volume browser[39] is available as open source on github[40].

[38] Disclaimer: this is just a summary in layman's terms of legal analysis that was done as part of the CEUR-WS semantification project
[39] http://ceur-ws-browser.bitplan.com/
[40] https://github.com/WolfgangFahl/pyCEURmake

A prototype for the presentation[41] of the CEUR-WS Semantification results has been created using Semantic MediaWiki [43] (SMW), which is using the same MediaWiki open source engine as Wikipedia. SMW has extensions for markup that is transforms subject-predicate-object statements to triples, which leads to a "Semantification" of the Wiki by making the triples available for query. Like most SMW installations we are using the standard SQL based triplestore and ask queries and not RDF/SPARQL [44].

A GitHub project for the single-point-of-truth metadata handling and conversion to different representations has been started at ceurws/ceur-spt[42].

Further background research material is supplied via the Semantic MediaWikis for Wolfgang Fahl's PhD[43] (public) and the ConfIDent requirements wiki[44] (access on request).

5.2. CEUR-WS Volume Browser

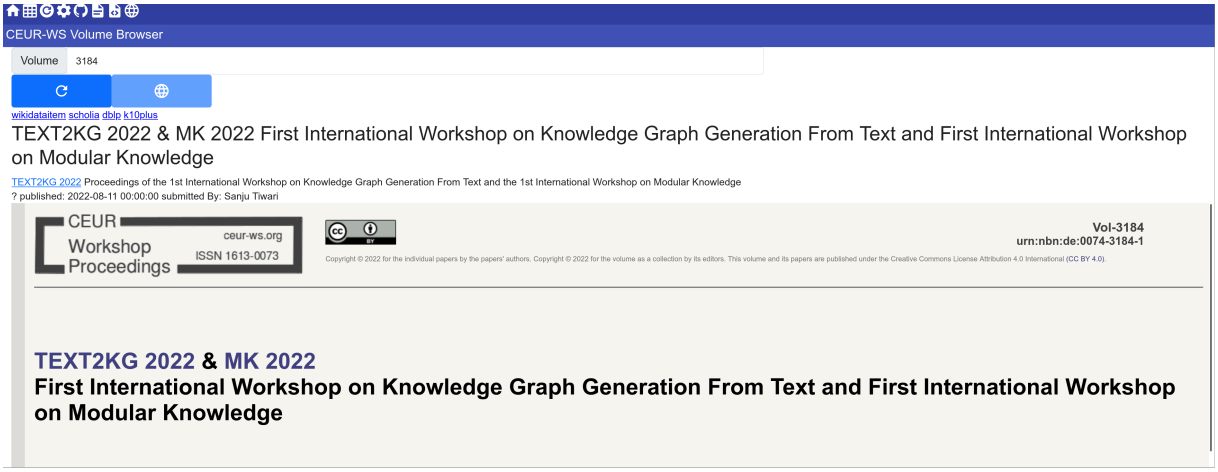

Fig. 7. Screenshot of the CEUR-WS Volume Browser

Figure 7 shows a screenshot of the CEUR-WS Volume Browser, which we created as a means to support semantification tasks such as transferring metadata 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[45].

Figure 8 shows an enlarged section of the screenshot where the links between the proceedings volumes and the external knowlede graphs are presented. In the example, these are wikidataitem[46], dblp[47], k10plus[48], and the Scholia links to the proceedings, event and event series[49] (which has links to the event and proceedings).

5.3. CEUR-WS Semantic MediaWiki

The CEUR-WS Semantic MediaWiki is available as a prototype as depicted in Figure 9, which shows how a content management system approach may be applied to the metadata, which allows for new features such as full

[41] http://ceur-ws.bitplan.com
[42] https://github.com/ceurws/ceur-spt
[43] https://cr.bitplan.com/index.php/Category:Text2KG
[44] http://rq.bitplan.com
[45] https://ceur-ws.org/Vol-3262/
[46] http://www.wikidata.org/entity/Q115053286
[47] https://dblp.org/db/conf/semweb/wikidata2022
[48] https://opac.k10plus.de/DB=2.299/PPNSET?PPN=1830580760
[49] https://scholia.toolforge.org/event-series/Q106429025

Fig. 8. CEUR-WS Volume Browser enlarged details with links to external KGs

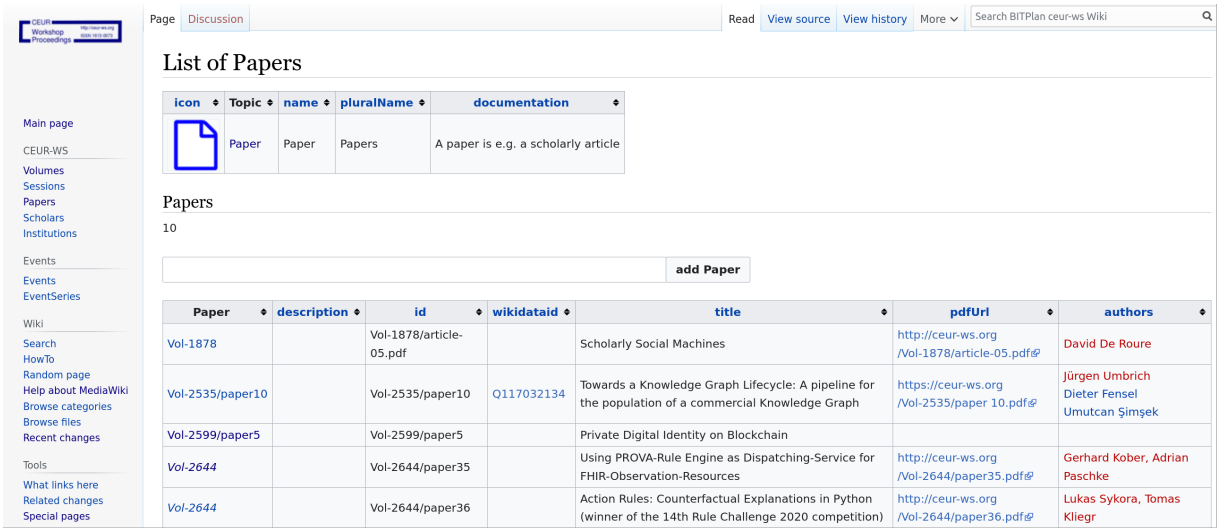

Fig. 9. List of example CEUR-WS Papers in the CEUR-WS Semantic MediaWiki

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 and allows to both visualizes the semantificiation and query the results via APIs. The example screenshot shows how a MediaWiki displays links with non-existing targets in red, allowing to judge the coverage of the disambiguation easily.

5.4. CEUR-WS Single Point of Truth Server

The Single Point of Truth Server serves as a prototype and proof of concept for the CEUR-WS semantification. Instead of static HTML, all content displayed by this server is generated from the semantified metadata. The separation of concern of metadata and presentation is implemented here.

The server software is available open source at github[50]. It runs on uvicorn ASGI server and FastAPI web framework and has only a limited list of dependencies to stable Python libraries for long term maintainability.

Key features are HTTP content negotiation for human (HTML) and computer (JSON,YAML,XML, ...) consumption and a documented RESTFul API.

The demo of the server is available at the RWTH Aachen i5 chair.

Using FastAPI version 0.1.0 with Swagger OAS3[51] compatibility, a comprehensive set of endpoints is offered. There are endpoints which are compatible with the traditional static for index and volume HTML tables of contents,

[50] https://github.com/ceurws/ceur-spt
[51] https://swagger.io/specification/

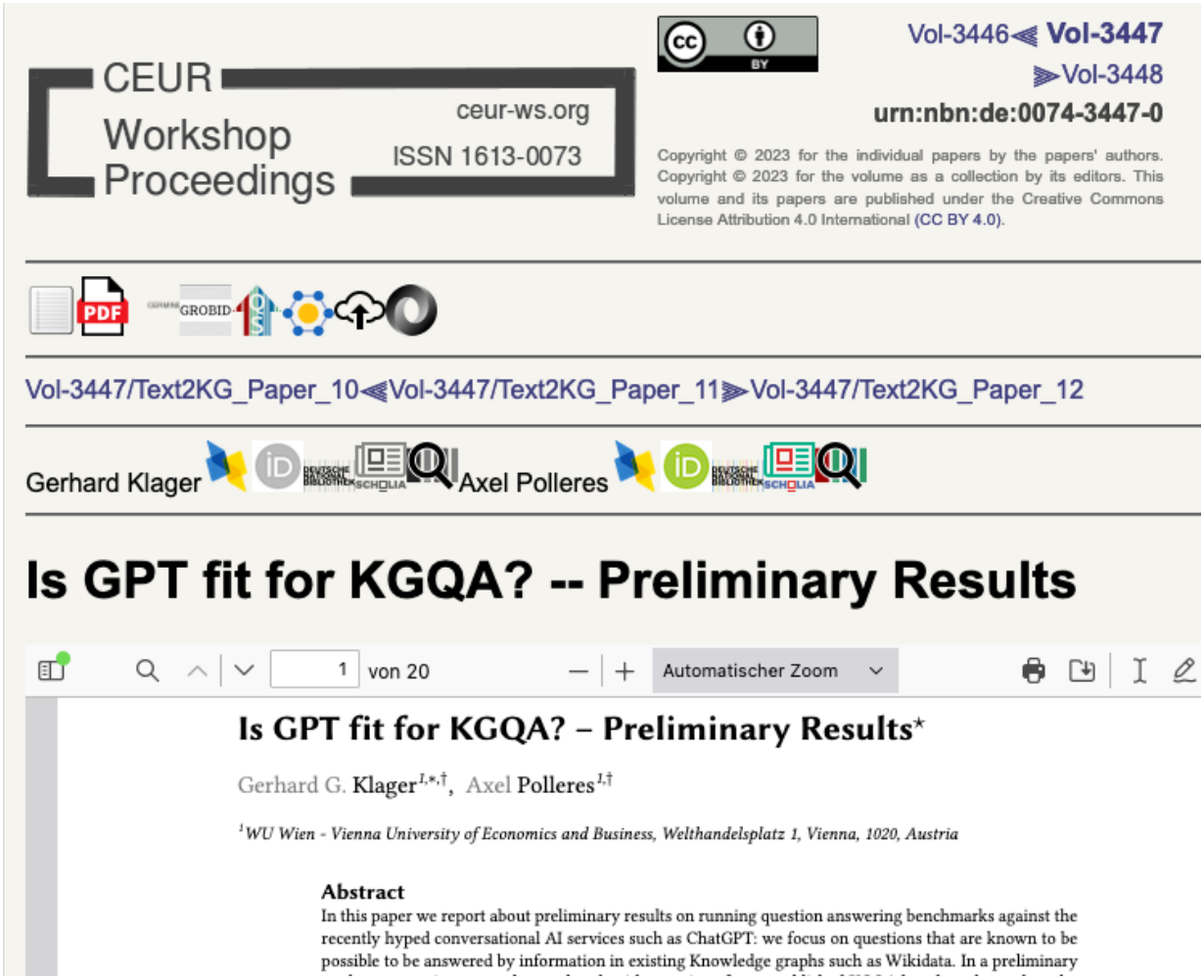

Fig. 10. Paper landing page of CEUR-WS Single Point of Truth server

as well as paper PDFs that deliver HTML content. In addition various other computer readable representations are available.

Notably, each paper is accessible via a dedicated endpoint, a feature absent in the traditional static CEUR-WS implementation. Links to corresponding volumes and author indices (dblp, GND, and Wikidata) are provided, while granting direct access to Scholia profiles based on the Wikidata items.

The paper endpoint enables retrieval of both the PDF and its plain text. It integrates the CERMINE and GROBID results for advanced XML analysis. The metadata can be exported as QuickStatements for Wikidata, wikibase-cli, Semantic MediaWiki Markup, or in standard formats such as JSON or YAML, facilitating interoperability and analysis for different use cases. Individual entity imports and amendments for different target platforms are thus possible. This is a step towards the scholarly CMS concept presented in section 7.1.

Figure 10 shows an example paper from the second TEXT2KG@ESWC2023 workshop.

When scholars change their name over time they are often inclined to ask to have their name changed in the original publications. While this is mostly not feasible the paper endpoint approach with an author bar allows to display the new name.

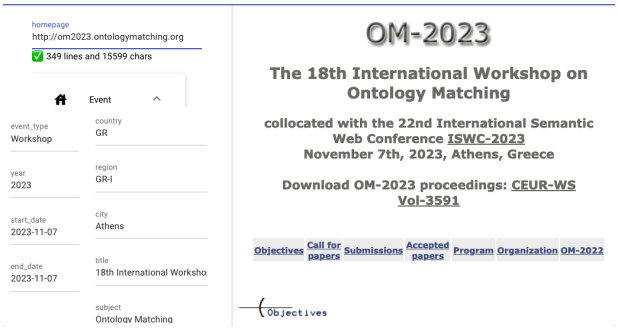

Fig. 11. Example of Event Metadata extraction from homepage

5.5. SemPubFlow tool

The SemPubFlow tool is a web based interactive system that supports Authors, Chairs and Editors in assembling the necessary metadata for the core entities needed for the publication of scientific event proceedings.

The need for such a tool was stated by proceedings editors in interviews done in the context of the ESWC@TEXT2KG workshop. Mostly the metadata for the events and proceedings is already available as part of e.g. homepages or paper and a tool for collecting this data effectively has been envisoned with CEURmake[52] earlier.

The editors explicitly asked for separation of the proceedings compilation process from the publication process itself.

SemPubflow is developed as an open source python project at github[53]. To try it out credentials are needed since there is cost involved in using some of the services—see section 6.2 [54].

Figure 11 shows an example homepage extraction for the CEUR-WS Volume 3591 OM2023 Ontology matching homepage `http://om2023.ontologymatching.org/`

The complete set of all relevant core-entity instance for proceedings is interactively collectable by SemPubFlow with a preview of the Volume index.html file available to compare with the traditional static approach[55].

6. Results

6.1. 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)[56]. 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 that 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 identifiers. This list of identifiers can then be used to query Wikidata to check if a person exists with at least one of those identifiers. The check against Wikidata showed that 1467 editors could be identified, and it also showed

[52] https://github.com/ceurws/ceur-make
[53] https://github.com/WolfgangFahl/SemPubFlow
[54] Ask the corresponding author for access of use local installation with your own OpenAI key
[55] Making this preview service available via API will be a key feature
[56] https://www.wikidata.org/wiki/Property:P2456

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 the CEUR-WS bot [45] see section 7.1.

6.2. LLM-based homepage metadata extraction

The SemPubFlow tool offers to extract event metadata from the homepage of an event to assist in obtaining the metadata for the core entities and to avoid the tedious and error prone cut & paste task involved from getting the metadata from a homepage. The OpenAI Chat-GPT API is used for this purpose. The prompts given to the LLM for the ChatGPT experiment[57] have been documented and are part ot the SemPubFlow software now. An excerpt of the prompt prefix is shown below.

```
provide the event signature elements:
- Acronym: The short name of the conference, often in uppercase.
- Frequency: How often the event occurs, like annual or biennial.
...
- Year: The year in which the event takes place.
- Ordinal: The instance number of the event, like 18th or 1st.
- Date: The start and end date or date range of the event.
...
- Title: The full title of the event, often indicating the scope and subject.
- Subject: The main topic or focus of the event.

in YAML Format.
...
Use ISO date format for dates.
Use start_date and end_date as field names.
Give the year as a 4 digit integer.
Give the location as country/region and city
Give the country using a 2 digit ISO 3166-1 alpha-2 code
Give the region using ISO_3166-2 code
...
valid answers e.g. would look like
# AVICH 2022
acronym: "AVICH 2022"
event_type: "Workshop"
...

Extract as instructed from the following homepage text:
```

An experiment was conducted with 250 random homepages selected from the 2876 CEUR-WS Volumes that have homepage references. Only 1455 of these URLs were available and only 817 made the content-length available to be able to select pages that are not too long for LLM analysis. The average content length was 20142 chars which is close to the limit of the standard 4K token boundary of the default ChatGPT models. See Histograms of content and text length[58].

An example response for the CEUR-WS Volume 1029 (ZEUS 2013) homepage `http://zeus-workshop.eu/2013/` was:

[57] https://cr.bitplan.com/index.php/Workdocumentation_2023-12-22
[58] https://cr.bitplan.com/index.php/Workdocumentation_2023-12-21

attr	event total	event #	event %	volume total	volume #	volume %	prec	recall	f1
acronym	250	240	96	250	250	100	0.43	0.43	0.43
city	250	200	80	250	234	93.6	0.78	0.74	0.75
country	250	186	74.4	250	234	93.6			
end_date	250	224	89.6	250	234	93.6	0.71	0.70	0.70
region	250	140	56	250	14	5.6	0.50	0.33	0.39
start_date	250	224	89.6	250	234	93.6	0.72	0.70	0.71
title	250	238	95.2	250	250	100			
year	250	234	93.6	250	241	96.4	0.85	0.80	0.82

Table 4

LLM based extraction of event attributes from homepages

```
- volume: 1029
  acronym: ZEUS Workshop 2013
  event_type: Workshop
  year: 2013
  start_date: '2013-02-21'
  end_date: '2013-02-22'
  country: DE
  region: DE-MV
  city: Rostock
  title: Fifth Central European Workshop on Services and their Composition
  subject: Services and their Composition
```

The responses were compared with the CEUR-WS single point of truth.

Table 4 shows the results of this experiment with ChatGPT-3.5-turbo as the model and a temperature of 0.0.

The availability and extractability of the attributes in the range of 56.0% to 96.0% with a median of 89.6% is promising and shows the feasibility of the approach. The precision, recall and f1 values are shown just for reference – in some cases fuzzy matching would be needed – e.g., for the title there has been not a single 100% match. The countries could not be matched, since we asked for ISO country codes, while the single point of truth has full country names, so "IT" and "Italy" would not match. Regions were too rarely available. Manual inspection of the data shows the consistency being over 80% with only minor changes being necessary in the case of inconsistencies.

The reproducibility of the experiment results was 82%. The median response time was 2 secs. More often than not the instruction not to add any other comments or content than requested was ignored and the YAML-loader would filter away the superfluous results.

The total cost of the API usage was US$1.07 or 0.4 cents per homepage.

Given these results and the cost and time advantage of LLMs over doing the same work with human curators we expect a strong motivation to adopt this approach.

6.3. Single Point of Truth against Wikidata validation

A major concern of the CEUR-WS editors and other stakeholders is the trustworthiness of the Wikidata entries. These entries are mostly created by the CEUR-WS Wikidata bot but also by other tools such as Scholia and individual Wikidata curators. Therefore, it is necessary to regularly check that the single point of truth data is consistent with the Wikidata entries. This is especially important for the identifiers being used. The pyLodStorage synchronization tool (see section 7.2) checks the consistency. Table 6 shows an example result for the URN-NBN persistent identifier. In 580 cases the URN identifier is set in Wikidata (since the URNs can be calculated from the volume number as explained below) but not available from the volume pages via a microformat annotation such as

```
<span class="CEURVOLNR">Vol-3601</span>
```

Volume	URN
Vol-2018	urn:nbn:de:0074-2018-6
Vol-2210	urn:nbn:de:0074-2212-3
Vol-2479	urn:nbn:de:0074-2479-C
Vol-2743	urn:nbn:de:0074-2743-1
Vol-2843	urn:nbn:de:074-Vol-2843-4
Vol-3419	urn:nbn:de:0074-3419-9
Vol-3466	urn:nbn:de:0074-XXX-1

Table 5
URN checksum Check

Table 6
URN-NBN Synchronization

left	↔	right	#	%
CEUR-WS	←	Wikidata	580	16.10%
CEUR-WS	↔	Wikidata	3006	83.43%
CEUR-WS	→	Wikidata	17	0.47%

as was the case for very early volumes.

In 17 cases, there was a mismatch originating from mistakes in the manual execution of the traditional workflow. These cases have been checked manually and could be partly fixed via the wikibase-cli command line tool with generated fixing commands such as:

```
$ wd add-claim Q116525021 P4109 "urn:nbn:de:0074-3249-2"
```

However, some of the URN entries where wrong in the input Volume HTML files see Table 5. Finding these problems was notoriously difficult since the documentation of DNB URN-NBN check digit generation is obscure and the website tool for generating the check digits does not have an API (see CEUR-Make issue 88[59] for the details) [60].

6.4. 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 7 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 of the Semantic Publishing Challenge, which were set as a benchmark in 2014 for different stakeholders, was subjectively rated by us, resulting in a list sorting the queries by priority[61]. 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[62] 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 run 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.

6.5. Evaluation by Further Quality Metrics

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.

[59] https://github.com/WolfgangFahl/pyCEURmake/issues/88
[60] different versions of the tool in Javascript, PHP and python needed to be compared due to the C-puzzle book style handling of pos++ in the code.
[61] https://cr.bitplan.com/index.php/List_of_Queries
[62] https://cr.bitplan.com/index.php/Semantic_Publishing_Challenge_Queries

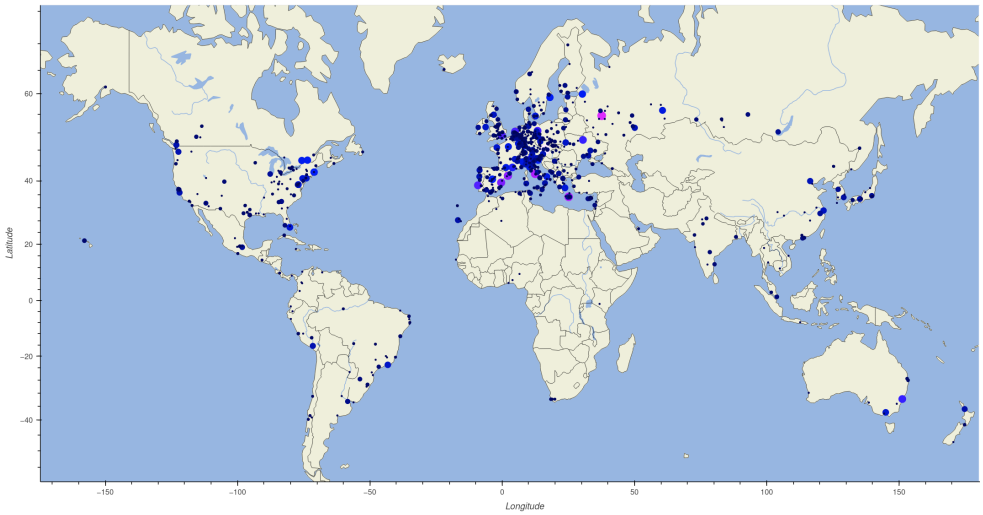

Fig. 12. Locations of all CEUR-WS proceedings events (Query 0.10)

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 SemPubFlow the data will be available as early as the event organizers see fit. With the separation of event and proceedings 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.

7. 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 that these 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 – involve a lot of manual curation, data quality errors deriving 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.

SemPubFlow introduces a metadata-first approach to scientific publishing, aiming to align the publishing process more closely with the FAIR principles. Through the integration of Language Model Systems and Wikidata, the workflow addresses several challenges in scholarly publishing, particularly around the efficiency and quality of metadata management. This paper has outlined the core components and potential benefits of SemPubFlow, including its ability to facilitate a more structured and accessible digital scholarly ecosystem.

7.1. Future Work

Figure 13 shows an overview of a possible new approach to publishing via CEUR-WS. The core idea is to separate the concerns of displaying content (such as static HTML/Semantic MediaWiki) from the storage of 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 a computer readable format such as JSON and generate the HTML presentation from this metadata.

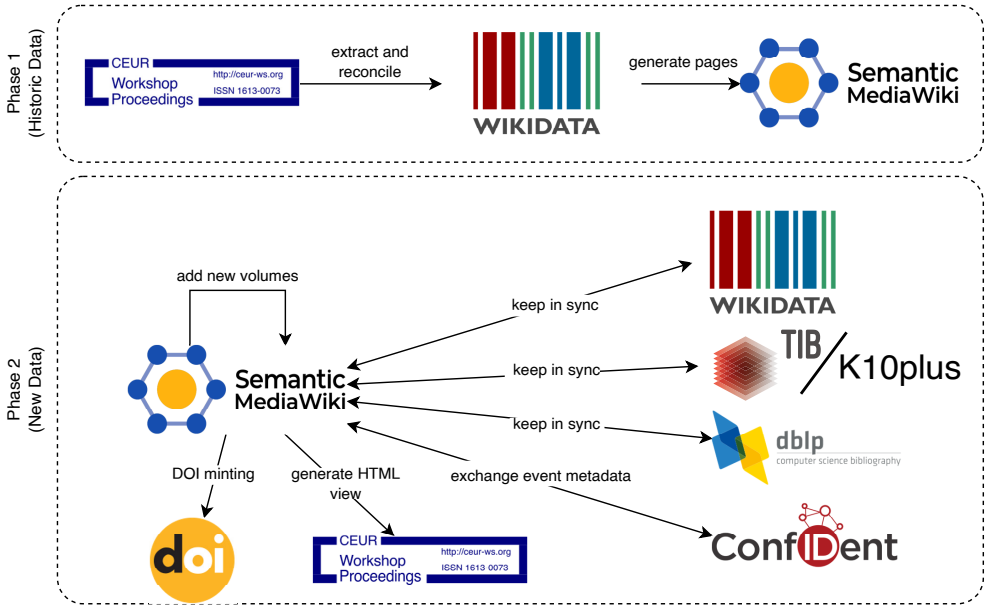

Fig. 13. CEUR-WS Semantification

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.

Mass creation, editing and disambiguation of relevant scholar entries in Wikidata has to be performed to make these entries available for lookup and reference by the SemPubFlow tool. We intend to build on the Scholia experience with this type of task.

The LLM approach already proven successful for volume homepages needs to also be applied for the much larger data sets of author homepages. An interesting research question is how well the pre processing of PDF files with CERMINE and GROBID is helpful in getting extraction and disambiguation results for paper metadata and how that might influence the efficiency, cost and accuracy.

Given the promising aspects of SemPubFlow, community feedback is essential for its iterative development to better meet the evolving needs of the academic ecosystem. Allowing the integration with other publishing outlets besides CEUR-WS will be a key success factor. Future enhancements will have to focus on increasing the automation and accuracy of metadata extraction and management, while ensuring the tools developed are accessible, user-friendly and well integrated with the publishing workflow.

In promoting metadata-first publishing, a public infrastructure like Wikidata could be used to dynamically generate artifacts, including homepages, CfP webpages, emails, and marketing materials, thus eliminating tedious manual tasks and forming a generic scholary content management system. This approach requires addressing the balance between standardization and individualization, ensuring that while a common framework is maintained for FAIRness and efficiency, the unique aspects of individual publications are not lost. Key to this is the separation of concerns: maintaining distinct layers for metadata storage, content generation, and presentation, allowing each to evolve independently while cohesively contributing to the scholarly publishing ecosystem.

7.2. Acknowledgements

Table 7 shows the main open source libraries being used by the work described.

Link	URL & Description
BeautifulSoup4	`https://pypi.org/project/beautifulsoup4/`
	HTML and XML parsing library
Catmandu	`https://github.com/LibreCat/Catmandu`
	Data processing toolkit with e.g. MARC conversion
nicegui	`https://github.com/zauberzeug/nicegui/`
	Web-based reactive UI development framework
justpy	`https://github.com/justpy-org/justpy`
	Python web framework for interactive apps
geograpy3	`https://github.com/somnathrakshit/geograpy3`
	Contextual place name extraction
pyLoDstorage	`https://github.com/WolfgangFahl/pyLoDStorage`
	List of Dict (Table) Storage library with named query support
wikibase-cli	`https://github.com/maxlath/wikibase-cli`
	Command Line Interface for Wikibase instances
py-yprinciple-gen	`https://github.com/WolfgangFahl/py-yprinciple-gen`
	Generate artifacts from y-principled metadata

Table 7
List of Tools with Links, URLs, and Descriptions

Use of generative AI: the creation of parts of this work has been assisted by ChatGPT-4.

ChatGPT Workdocumentations[63] documents some of the experiments. When creating software, ChatGPT has been extensively used, e.g., for implementing the URN check digit calculator[64].

Also, LaTeXsyntax generation has been done with ChatGPT e.g. converting tables and lists to proper format as well as grammar and spell checks.

We would like to thank Jakob Voß for helping with the k10plus matching and creating the Wikidata property colocated with (P11633)[65] in due time.

Thomas Hoeren and Jonas Kuiter of ITM, Münster kindly gave hints for legal aspects as summarized in section 4.5.

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). [66]

References

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

[2] J. Priem, H. Piwowar and 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). doi:10.5281/ZENODO.6936227. https://zenodo.org/record/6936227.

[3] J. Ganseman, Refactoring a library's legacy catalog: a case study, in: *IAML Congress 2015*, 2015. 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. https://docs.lib.purdue.edu/iatul/1979/papers/26/.

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

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

[63] https://cr.bitplan.com/index.php/Category:ChatGPT

[64] https://github.com/WolfgangFahl/pyCEURmake/blob/main/ceurws/urn.py

[65] https://www.wikidata.org/wiki/Property:P11633

[66] ConfIDent project; see https://gepris.dfg.de/gepris/projekt/426477583

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

[8] CEUR-WS Editor Team, How to Submit to CEUR Workshop Proceedings, 2019, Online; accessed 2023-12-28. https://ceur-ws.org/HOWTOSUBMIT.html.

[9] GO FAIR International Support and Coordination Office, FAIR Principles, GFISCO, 2019. https://www.go-fair.org/fair-principles/.

[10] 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, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C.T. Evelo, R. Finkers, A. Gonzalez-Beltran, A.J.G. Gray, P. Groth, C. Goble, J.S. Grethe, J. Heringa, P.A.C. '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 and B. Mons, The FAIR Guiding Principles for scientific data management and stewardship, *Scientific Data* **3**(1) (2016). doi:10.1038/sdata.2016.18. https://doi.org/10.1038/data.2016.18.

[11] H. Cousijn, R. Braukmann, M. Fenner, C. Ferguson, R. van Horik, R. Lammey, A. Meadows and S. Lambert, Connected Research: The Potential of the PID Graph, *Patterns (New York, N.Y.)* **2**(1) (2021), 100180. doi:https://doi.org/10.1016/j.patter.2020.100180.

[12] 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 and A. Zimmermann, Knowledge Graphs, *ACM Computing Surveys* **54**(4) (2021), 1–37. doi:10.1145/3447772. https://doi.org/10.1145/F3447772.

[13] T. Berners-Lee, J. Hendler and 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. https://www.scientificamerican.com/article/the-semantic-web/.

[14] C. Lange and 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. doi:10.1007/978-3-319-12024-9_8. https://doi.org/10.1007/978-3-319-12024-9_8.

[15] B. Sateli and 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. doi:10.1007/978-3-319-25518-7_11. https://doi.org/10.1007/978-3-319-25518-7_11.

[16] W. Fahl, The History of Scientific Publishing, 2023, Online; accessed 2023-03-15. https://cr.bitplan.com/index.php/The_History_of_Scientific_Publishing.

[17] C. Yu, C. Zhang and J. Wang, Extracting Body Text from Academic PDF Documents for Text Mining, *CoRR* **abs/2010.12647** (2020). https://arxiv.org/abs/2010.12647.

[18] T. Bardini, *Bootstrapping: Douglas Engelbart, Coevolution, and the Origins of Personal Computing*, Stanford University Press, USA, 2001. ISBN 0804738718.

[19] Christina Engelbart, About Bootstrapping , 2007, Online; accessed 12 March 2023.

[20] M. Kolchin and 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. doi:10.1007/978-3-319-12024-9_11. https://doi.org/10.1007/978-3-319-12024-9_11.

[21] M. Kolchin, E. Cherny, F. Kozlov, A. Shipilo and L. Kovriguina, CEUR-WS-LOD: Conversion of CEUR-WS Workshops to Linked Data, in: *Semantic Web Evaluation Challenges*, Springer International Publishing, 2015, pp. 142–152. doi:10.1007/978-3-319-25518-7_12. https://doi.org/10.1007/978-3-319-25518-7_12.

[22] H. Cunningham, D. Maynard, K. Bontcheva and 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, pp. 168–175–. doi:10.3115/1073083.1073112.

[23] H. Cunningham, V. Tablan, A. Roberts and K. Bontcheva, Getting More Out of Biomedical Documents with GATE's Full Lifecycle Open Source Text Analytics, *PLoS Computational Biology* **9**(2) (2013), e1002854. doi:10.1371/journal.pcbi.1002854.

[24] B. Sateli and 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. doi:10.1007/978-3-319-53637-8_5. https://doi.org/10.1007/978-3-319-53637-8_5.

[25] M. Milicka and 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. doi:10.1007/978-3-319-25518-7_7. https://doi.org/10.1007/F978-3-319-25518-7_7.

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

[27] 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. ISBN 9781450314794. doi:10.1145/2342488.2342490.

[28] J. Kunze and E. Bermès, The ARK Identifier Scheme, 2022. https://www.ietf.org/archive/id/draft-kunze-ark-36.html.

[29] J. Franken, A. Birukou, K. Eckert, W. Fahl, C. Hauschke and C. Lange, Persistent Identification for Conferences, *Data Science Journal* **21** (2022). doi:10.5334/dsj-2022-011.

[30] V. Bryl, A. Birukou, K. Eckert and M. Kessler, What's in the proceedings? Combining publisher's and researcher's perspectives, in: *4th Workshop on Semantic Publishing (SePublica)*, A. García Castro, C. Lange, P. Lord and R. Stevens, eds, CEUR Workshop Proceedings, Aachen, 2014. ISSN 1613-0073. http://ceur-ws.org/Vol-1155#paper-01.

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

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

[33] W. Fahl, K. Eckert and C. Lange, Extracting Event Metadata from Proceedings Titles, Zenodo, 2022. doi:10.5281/ZENODO.6568728. https://zenodo.org/record/6568728.

[34] D. Tkaczyk, P. Szostek, M. Fedoryszak, P.J. Dendek and Ł. Bolikowski, CERMINE: automatic extraction of structured metadata from scientific literature, *International Journal on Document Analysis and Recognition (IJDAR)* **18**(4) (2015), 317–335. doi:10.1007/s10032-015-0249-8.

[35] 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. doi:10.1007/978-3-642-04346-8_62. https://doi.org/10.1007/978-3-642-04346-8_62.

[36] AKhatun, Wikidata Scholarly Articles Subgraph Analysis, Online; accessed 2023-12-30. https://wikitech.wikimedia.org/wiki/User:AKhatun/Wikidata_Scholarly_Articles_Subgraph_Analysis.

[37] OECD, *OECD Glossary of Statistical Terms*, OECD, 2008. doi:10.1787/9789264055087-en.

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

[39] S. Subramanian, D. King, D. Downey and 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. doi:10.1109/jcdl52503.2021.00029.

[40] J. Kim, Evaluating author name disambiguation for digital libraries: a case of DBLP, *Scientometrics* **116**(3) (2018), 1867–1886. doi:10.1007/s11192-018-2824-5.

[41] D. Vrandečić and M. Krötzsch, Wikidata, *Communications of the ACM* **57**(10) (2014), 78–85. doi:10.1145/2629489.

[42] K. Pohl and C. Rupp, *Requirements engineering fundamentals*, 2nd edn, Rocky Nook, Santa Barbara, CA, 2015.

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

[44] Wolfgang Fahl, SMWCon Fall 2020/Backend infrastructure experience exchange e.g. SqlStore, Elasticsearch, Blazegraph, Virtuoso, Jena, 2020, Online; accessed 2023-12-30. https://www.semantic-mediawiki.org/wiki/SMWCon_Fall_2020/Backend_infrastructure_experience_exchange_e.g._SqlStore,_Elasticsearch,_Blazegraph,_Virtuoso,_Jena.

[45] Wikidata Contributors, Requests for permissions/Bot/CEUR-WS, 2023, Online; accessed 2023-12-30. https://www.wikidata.org/wiki/Wikidata:Requests_for_permissions/Bot/CEUR-WS.