Employing RAG to Create a Conference Knowledge Graph from Text

Daniil Dobriy[1,*,†]

[1]Vienna University of Economics and Business, Welthandelsplatz 1, Vienna, 1020, Austria

Abstract

In this paper, we present Semantic Observer, a platform that 1) defines a FAIR Conference Ontology for describing academic conferences, 2) presents an RAG architecture that constructs a Conference Knowledge Graph based on this ontology, 3) evaluates the architecture on a corpus of latest available CORE conference websites. The Conference Ontology models key entities such as conferences, workshops and challenges, organizer and programme committees, calls for papers and proposals as well as major deadlines and relevant topics. In the evaluation, we compare the performance of three leading Large Language Models: GPT-4 Turbo and Claude 3 Opus - in supporting the Knowledge Graph construction from text. The best-performing RAG architecture is then implemented in Semantic Observer and available in a SPARQL endpoint to make up-to-date conference information FAIR: findable, accessible, interoperable and reusable.

Keywords

Research Ecosystem, Retrieval-Augmented Generation, Web Crawling, Semantic Web, Knowledge Graph, Ontology Engineering

1. Introduction

1.1. Venue discovery and selection

The search for applicable academic venues takes time and, most importantly, in-depth knowledge of venues in the respective field of research, essentially creating a barrier to publishing, especially for early-stage researchers. Therefore, often, senior and experienced academics are often relied upon to play a crucial role advising less experienced researchers on venue selection. Sometimes, however, applicable venues are still overlooked, for example, in such cases, when a conference in a neighbouring area of research, which would not be normally applicable to a given academic's line of work, would offer a special interdisciplinary track for publication which in-turn would become highly relevant. Also, as such, venue selection is a process that, by itself, should logically not affect the intrinsic quality of the works themselves. Thus, freeing the enterprising researchers' time (Figure 1) from this process by supplying them with a broad overview from the start would have an awesome impact on the variety of work submitted to peer review of respective conferences and, therefore, enrich the scientific project as a whole.

1.2. Vision behind the conference intelligence platform

The vision behind the project is to provide academics with an up-to-date and reliable conference intelligence platform, which, besides supplying them with general overview on applicable

TEXT2KG'24: International Workshop on Knowledge Graph Generation from Text, May 28–June 1, 2024, Hersonissos,

Figure 1: Picture from semantic.observer depicting a scene from an enterprising researcher's life.

venues, and thus assist with developing a strategy for publishing their work, also informs academics of new potential venues for publication in their line of work and notifies them of relevant recently published, upcoming and updated deadlines for submission.

1.3. Scope and paper structure

This paper focuses specifically on the central aspect of a conference intelligence platform, namely the ability to reliably extract relevant conference information from conference websites, which includes both retrieving the full contents of conference websites as well as employing an Retrieval-Augmented Generation (RAG) architecture to extract relevant data according to the pre-defined ontology.

Thus, the remainder of the paper is structured in the following way:

- Section 2 discusses related work on website metadata, website discovery and crawling, as well as RAG for extracting structured knowledge.
- Section 3 describes the methodology used in the study, including the definition of the conference ontology, the description of the RAG architecture used, the prompt definition and the selection of academic conference websites for evaluation.
- Section 4 describes the data collection process, including the pre-processing steps.

Greece
*Corresponding author.
daniil.dobriy@wu.ac.at (D. Dobriy)
dobriy.org (D. Dobriy)
0000-0001-5242-302X (D. Dobriy)

- Section 5 presents the results of the evaluation, comparing the performance of selected Large Language Models (LLMs) in Knowledge Graph (KG) construction.
- Section 6 discusses the implication of the finding and the implementation in the Semantic Observer.
- Section 7 concludes the paper by summarizing the main contributions.
- Section 9 outlines concrete directions for future work.

2. Related work

2.1. Conference metadata and ranking

There has been a variety of platforms and services ranking academic conferences. In the computing disciplines, one of the most prominent such initiatives is the *CORE Ranking portal*, maintained by the CORE Advisory Committee.[1] The CORE Ranking categorises conferences in grades depending on their geographic extent and a set of visibility and academic quality considerations, ranking from *Regional* and *National* to more recognized classes like *B*, *A* and *A** conferences.[2] Other notable conference rankings in the field include the *QUALIS Conference Ranking* sponsored by the Brazilian Federal Agency for the Improvement of Higher Education[3] and the *ERA's ranking for conferences* from the Excellence in Research for Australia initiative.[4]

However, these portals only collect limited metadata about the conferences or the data collected is not publicly disclosed. Some platforms explicitly collect conference metadata: DBLP is an open-access repository collecting conference proceedings and metadata,[5], OpenResearch.org is a Semantic MediaWiki-powered resource for conference metadata[6] and other, larger digital libraries like *IEEE Xplore*,[7] *Scopus*,[8] *Web of Science*[9] also collect (besides other things) some metadata about conferences. Notably, Wikidata also has an active community maintaining metadata on a variety of conferences. Currently, one can find metadata related to 9945 academic conferences on Wikidata.[10] The most common (> 100 occurrences) properties (the properties linking the object to external identifiers are grey).

There are also platforms explicitly collecting Call for Paper (CfP) information. For example, WikiCFP is a manually curated platform collecting CfPs for conferences, workshops and further events in the field of Web[11]. The platform breaks down the deadlines (see Figure 2) but doesn't

[1] See, https://www.core.edu.au/conference-portal
[2] Cf., https://drive.google.com/file/d/1DQixeK53tlq_jh6IspIHroiwu1pmM6-y/view?usp=sharing
[3] See, https://www.capes.gov.br/images/documentos/Qualis_periodicos_2016/Qualis_conferencia_ccomp.pdf
[4] See, http://direction.bordeaux.inria.fr/~roussel/rankings/era
[5] See, https://dblp.org
[6] See, https://openresearch.org
[7] See, https://ieeexplore.ieee.org/Xplore/home.jsp
[8] See, https://www.scopus.com/home.uri
[9] See, https://clarivate.com/products/scientific-and-academic-research/research-discovery-and-workflow-solutions/webofscience-platform/
[10] You could use the following SPARQL query to retrieve the number on Wikidata (https://query.wikidata.org/):
```
SELECT (COUNT(?conference) AS ?NumberOfConferences) WHERE { ?conference
wdt:P31/wdt:P279* wd:Q2020153. SERVICE wikibase:label { bd:serviceParam wikibase:language
"[AUTO_LANGUAGE],en". }
```
[11] See, http://www.wikicfp.com

go into detail with regard to the types of CfPs available, mostly presenting only the information for one (main) CfP related to a given conference whereby. However, for illustration, ESWC 2024 already had 10 distinct calls for contributions,[12] not including CfPs of all the related events (workshops, challenges etc.).

Property ID	Property Label	Count
wd:P31	instance of	11243
wd:P276	location	9455
wd:P17	country	9268
wd:P580	start time	9123
wd:P582	end time	9111
wd:P1813	short name	8801
wd:P179	part of the series	7959
wd:P1476	title	7926
wd:P973	described at URL	6962
wd:P664	organizer	3307
wd:P227	GND ID	3122
wd:P823	speaker	2855
wd:P921	main subject	2369
wd:P214	VIAF ID	2148
wd:P710	participant	1928
wd:P856	official website	1016
wd:P212	ISBN-13	984
wd:P10692	DBLP event ID	557
wd:P244	Library of Congress authority ID	516
wd:P585	point in time	492
wd:P123	publisher	435
wd:P6721	K10plus PPN ID	310
wd:P5804	has program committee member	295
wd:P859	sponsor	289
wd:P5124	WikiCFP event ID	185
wd:P131	located in the administrative territorial entity	144
wd:P2936	language used	126
wd:P793	significant event	123

Table 1
Most common properties used to describe academic conferences on Wikidata.[13]

2.2. Embedding techniques for structured data

Different technologies exist to embed structured data in web pages, including RDFa, JSON-LD, Microdata. Metadata embedded in this way is indexed by search engines and can be integrated

[12] See, https://2024.eswc-conferences.org/call-for-contributions-eswc-2024/

[13] You could use the following SPARQL query to retrieve these properties on Wikidata:

```
SELECT ?property ?propertyLabel (COUNT(?conference) AS ?count) WHERE { ?conference wdt:P31
wd:Q2020153. ?conference ?p ?statement. ?property wikibase:directClaim ?p. SERVICE
wikibase:label { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en". } } GROUP BY
?property ?propertyLabel ORDER BY DESC(?count)
```

Figure 2: Breakdown of deadlines on WikiCFP.

into centralized knowledge bases [1]. Tools like the python library *extruct*[14] can be used to automatically extract a wide variety of such embedded metadata, supporting the technologies mentioned above as well as the Open Graph protocol.[15]

2.3. Ontologies for conference data

A number of ontologies has been developed that can be used to describe conference metadata:

- Semantic Web Conference Ontology [2, 3]
- Comprehensive Call Ontology for Research 2.0 [4]
- EVENTSKG Scientific Events ontology [5]
- AceKG ontology [6]
- OR-SEO: Scientific Events Data Model [7]
- SEDE: An ontology for scholarly event description [8]
- ESWC and ISWC metadata projects [9]
- schema.org[16]

Table 2 describes common aspects of a conference description covered by these ontologies. Notably, most ontologies represent basic metadata (e.g., conference title and dates) as well as topics of interest. Already less ontologies describe related events (workshops), proceedings and related events. Most importantly, the most relevant aspects of a conference for the researchers submitting their work, i.e., Calls for Papers and Deadlines, are among the least modelled aspects. Most interestingly, the submission guidelines (formats) are not represented by any ontology.

2.4. Web platform discovery

Web platform discovery is a research priority for the Semantic Web as it combines techniques and approaches from web crawling, automatic extraction of (structured) web contents and

[14]Available on pip: https://pypi.org/project/extruct/
[15]See, https://ogp.me
[16]See, https://schema.org

Table 2
Ontologies and their intended coverage of conference aspects

Ontology	Conference Metadata	Conference Series	Related Events	Conference Proceedings	Members of Committees	Calls for Papers	Other Types of Calls	Deadlines	Topics of Interest	Submission Format
SW Conference Ont.	✓	✓	✓	✓	✓				✓	
AceKG	✓								✓	
Comp. Call Ont.				✓	✓	✓				
EVENTSKG	✓		✓						✓	
OR-SEO	✓	✓	✓	✓	✓	✓	✓	✓	✓	
SEDE	✓		✓	✓		✓			✓	
ESWC & ISWC	✓	✓	✓	✓	✓	✓		✓	✓	
Schema.org	✓	✓	✓					✓		

Search Engine automation to estimate the extent of Linked Open Data (LOD) [10]. Many of the popular Search Engines provide dedicated APIs to retrieve search results using their indices automatically, includign: Google,[17] Bing,[18] and Naver[19] as well as external APIs which can query many Search Engines simultaneously, such as SerpAPI.[20]

There are a number of libraries that are used for web scraping. The basic approach to web scraping is static, for which simple requests or Python libraries like BeautifulSoup can be used. However, as the approach towards publishing websites as dynamic web applications becomes more popular (e.g., the website for ISWC 2024[21] is built with Cvent, a dynamic website application[22]), dynamic web scraping tools are increasingly needed. Such tools are also used for web automation and include Selenium,[23] Playwrite[24] and Scrapy.[25]

Best practices in the field of website-friendly crawling that we adhere to include:

1. Following guidelines set forth in *robots.txt* allowing/disallowing the automatic crawling of certain parts of the website.
2. Load moderation and appropriate timeout between requests to not overload the server with a flurry of tasks.
3. Using a descriptive user-agent in the header of requests to inform the server of the crawling procedure.

[17]See, https://developers.google.com/custom-search/v1/overview
[18]See, https://www.microsoft.com/en-us/bing/apis
[19]See, https://developers.naver.com/docs/common/openapiguide/
[20]See, https://serpapi.com/
[21]See, https://iswc2024.semanticweb.org
[22]See, https://www.cvent.com
[23]See, https://selenium-python.readthedocs.io
[24]See, https://playwright.dev/python
[25]See, https://github.com/scrapy/scrapy

2.5. Retrieval-Augmented Generation

The advent of LLMs has led to major new developments in the field of Semantic Web. On the one side, relevant to the study, LLMs have spurned abundant research in the direction of Knowledge Graph construction and completion [11, 12, 13, 14, 15, 16, 17, 18, 19, 20]. On the other hand, the broad implementation of LLMs have led to the development of a variety of Retrieval-Augmented Generation approaches, which infuse LLMs with Knowledge Graphs and structured data in various ways [21, 22, 18, 23, 24]. Thus, an array of advanced Large Language Models allows us to efficiently extract relevant information from available website markup code and text, including augmenting the process by incorporating structured knowledge in the pipeline. Table 3 gives an overview of the leading LLMs (GPT-4 Turbo,[26] Claude 3 Opus,[27] Gemini 1.5[28] and Mistral Large[29]) as of the writing of this paper. Notably, these models are closed-source, and the most advanced open-source models currently have a limited context window (e.g., 4,096 tokens[30] for Llama-2).

Table 3
Comparison of Leading Large Language Models

Model	Maximal Context Length	Training Data Cut-off Date
GPT-4 Turbo	128,000 tokens	Dec 2023
Claude 3 Opus	200,000 tokens	Aug 2023
Gemini 1.5	128,000 tokens	Feb 2024
Mistral Large	32,000 tokens	Before Feb 2024

3. Methodology

3.1. Ontology

In order to capture a wide variety of conference details (see Table 2 for aspects), we capture all of them in an OWL ontology, conforming to OWL 2 DL. The ontology is composed according to FAIR principles, emphasising re-usability and linking to existing ontologies. Figure 3 illustrates the main classes and relationships of the ontology.

3.2. Architecture

Our architecture (illustrated in Figure 4), is designed to automate the process of extracting structured data from information available on the conference website on the web. The process starts with a given website URL for an academic conference and ends with the integration of extracted structured data into the Conference Knowledge Graph, also accessible via a SPARQL endpoint. Below, we detail the components and their functions as part of the system.

[26] See, https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo
[27] See, https://www.anthropic.com/news/claude-3-family
[28] See, https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/#sundar-note
[29] See, https://mistral.ai/news/mistral-large/
[30] Cf., https://llama.meta.com/llama2

Figure 3: Illustration of the main classes and relationships of the Semantic Observer ontology

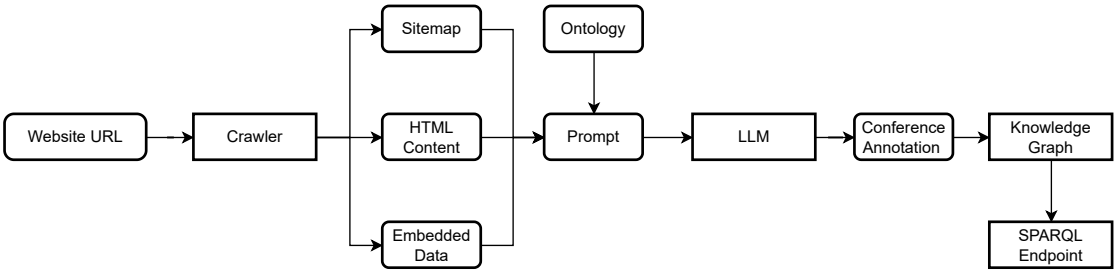

Figure 4: Architecture

1. The website URL serves as the entry point for the crawler. This URL is either retrieved from the user, or it is automatically retrieved by the discovery service [10].
2. The crawler navigates to the website and traverses the contents. The main functions of the crawler are:
   - retrieving a sitemap if it exists and extending it through traversing the links,
   - extracting embedded structured data (Microdata, JSON-LD, RDFa, OpenGraph),
   - retrieving HTML contents of the website pages.
3. Afterwards, the system uses extracted data (sitemap, embedded metadata and HTML contents) together with the pre-defined ontology to formulate a prompt for the LLM.
4. The LLM processes the prompt and returns the structured representation of the conference.
5. The representation is validated and integrated back into the Knowledge Graph.

3.3. Data selection

For the evaluation, we have opted to focus on the academic conferences present in the CORE Ranking under the field of Data Management and Data Science. To retrieve this list of conferences, one has to use the Australian and New Zealand Standard Research Classification (ANZSRC) code for the field, which in this case is 4605.[31] We furthermore limit the selection to the top 20 conferences from the CORE 2021 Ranking - the resulting subset of conferences is shown in Table 4. For further analysis, we retrieve the 2023 iterations of the conference websites as these are the first reliably available for all conferences after the Covid period, when some conferences didn't take place.

4. Data collection

In this section, we will present the initial data from the collection stage. Table **??** shows the response status of the websites as well as the presence of robots.txt and its restrictiveness, presence and extent of the sitemap as well as presence and type of embedded metadata.

[31] Cf., https://www.abs.gov.au/statistics/classifications/australian-and-new-zealand-standard-research-classification-anzsrc/latest-release

Full Name	Abbr.	Rank	Website
IEEE International Conference on Data Mining	ICDM	A*	https://www.cloud-conf.net/icdm2023/
ACM International Conference on Knowledge Discovery and Data Mining	KDD	A*	https://kdd.org/kdd2023/
ACM SIGMOD-SIGACT-SIGART Conference on Principles of Database Systems	PODS	A*	https://2023.sigmod.org
ACM International Conference on Research and Development in Information Retrieval	SIGIR	A*	https://sigir.org/sigir2023/
ACM Special Interest Group on Management of Data Conference	SIGMOD	A*	https://2023.sigmod.org
International Conference on Very Large Databases	VLDB	A*	https://vldb.org/2023/
ACM International Conference on Web Search and Data Mining	WSDM	A*	https://www.wsdm-conference.org/2023/
International World Wide Web Conference	WWW	A*	https://archives.iw3c2.org/www2023/
Conference on Innovative Data Systems Research	CIDR	A	https://www.cidrdb.org/cidr2023/
IEEE International Working Conference on Mining Software Repositories	MSR	A	https://conf.researchr.org/home/msr-2023
ACM International Conference on Recommender Systems	RecSys	A	https://recsys.acm.org/recsys23/
IEEE International Conference on Data Science and Advanced Analytics	DSAA	A	https://conferences.sigappfr.org/dsaa2023/
SIAM International Conference on Data Mining	SDM	A	https://www.siam.org/conferences/cm/conference/sdm23
Pacific-Asia Conference on Knowledge Discovery and Data Mining	PAKDD	A	https://pakdd2023.org
Extended Semantic Web Conference	ESWC	A	https://2023.eswc-conferences.org
European Conference on Information Retrieval	ECIR	A	https://ecir2023.org
ACM International Conference on Information and Knowledge Management	CIKM	A	https://uobevents.eventsair.com/cikm2023/
ACM International Conference on Advances in Geographic Information Systems	SIGSPATIAL	A	https://sigspatial2023.sigspatial.org
European Conference on Machine Learning and PKDD	ECML PKDD	A	https://2023.ecmlpkdd.org
International Conference on Database Theory	ICDT	A	http://edbticdt2023.cs.uoi.gr

Table 4
Selected Conferences from the CORE Ranking

In terms of availability, all websites responded with the *200* status code signifying their availability, except for CIKM 2023, which responded with the denial of the request. With regard to the Robots.txt policy, most websites either do not possess any (either because robots.txt is missing or empty) or the policy stipulates a delay of 20 seconds between requests. Only PAKDD 2023 disallows visiting internal/administrator pages, and surprisingly, WWW 2023 forbids any automatic crawling of the site - potentially, to induce Search Engine crawlers to index the newer conference website (2024) instead only.

Most academic conference websites do not provide a sitemap. However, it is extensive whenever it is available (RecSys 2023, PAKDD 2023, and ECMLPKDD 2023) and includes pages not normally available through link traversal starting from the homepage. Therefore, we uniformly re-create the sitemap when crawling the website to compensate for this shortcoming

and ensure comparability.

Conference	Availability	Robots.txt	Sitemap	Gen. Sitemap	2023 Only
ICDM 2023	✓	-	-	2	2
KDD 2023	✓	Delay (20 sec.)	-	-	-
PODS 2023	✓	Delay (20 sec.)	-	88	88
SIGIR 2023	✓	Delay (20 sec.)	-	-	-
SIGMOD 2023	✓	Delay (20 sec.)	-	93	93
WSDM 2023	✓	Delay (20 sec.)	-	570	64
WWW 2023	✓	Disallow (all)	-	96	95
CIDR 2023	✓	-	-	111	9
MSR 2023	✓	-	-	-	-
RecSys 2023	✓	Delay (20 sec.)	573	584	40
DSAA 2023	✓	-	-	-	-
SDM 2023	✓	-	-	-	-
PAKDD 2023	✓	Disallow (admin)	35	31	31
ESWC 2023	✓	-	-	75	75
ECIR 2023	✓	-	-	27	27
CIKM 2023	✗(999)	-	-	-	-
SIGSPATIAL 2023	✓	Delay (20 sec.)	-	38	38
ECML PKDD 2023	✓	-	67	-	-
ICDT 2023	✓	-	-	1	1

Table 5
Comprehensive Conference Data Including Initial Data Collection and LLM Filtering Results

We found that embedded metadata was relatively common across websites, with RDFa being the most common type of embedding metadata. However, all of this embedded metadata was created automatically (therefore, we put the checkmarks in parentheses), and none included any semantics beyond the language tag or page title, except for VLDB 2023. VLDB is the only conference, where schema.org has been used to describe the events in any perceivable detail. Interestingly, leading (Semantic) Web conferences like WWW and ESWC did not, notably, include any semantically rich descriptions of the event.

Therefore, in accordance with the architecture detailed in Section 3, to remedy the pervasive lack of both sitemaps and semantically rich metadata, we propose a tool for a full crawl of the websites' sitemaps, in accordance with the limitations set forth by the Robots.txt as presented in Table 5, i.e. setting the timeout between subsequent requests to 20 seconds in most cases and taking the liberty to ignore WWW 2023 no-crawl restriction.

In order to generate the sitemap for the website, we are starting with the main pages defined in Table 4 and then collecting all further links pointing to the same base URL. We recursively continue the process with all newly discovered links until reaching the point where no further links are being discovered for the base URL. Following the sitemap generation, we also request all pages from the sitemap.

Table 5 shows the actual crawled extent of sitemaps (number of website pages) and filtered-out number that excludes all non-2023 pages.[32]

[32]For this, we have employed an LLM to filter out non-2023 pages from the list of sitemap pages, utilizing

Conference	RDFa	Microdata	JSON-LD	OpenGraph
ICDM 2023	(✓)			
KDD 2023	(✓)	(✓)	(✓)	(✓)
PODS 2023				
SIGIR 2023	(✓)	(✓)		
SIGMOD 2023				
VLDB 2023		(✓)	✓schema.org	
WSDM 2023	(✓)			
WWW 2023	(✓)			(✓)
CIDR 2023				
MSR 2023	(✓)			(✓)
RecSys 2023	(✓)		(✓)	(✓)
DSAA 2023	(✓)			
SDM 2023	(✓)			(✓)
PAKDD 2023	(✓)			
ESWC 2023	(✓)			
ECIR 2023	(✓)			
SIGSPATIAL 2023				
ECML PKDD 2023	(✓)			
ICDT 2023				

Table 6
Metadata Embedded in Conference Websites

In the final step, after compiling both the sitemap and the page HTML contents into a string representation of a JSON-formatted object, we augment it with the Turtle representation of the ontology defined in 3, and a pre-defined prompt to form the final prompt for a long-context LLM, as illustrated in Figure 5.

5. Results

From the 10 websites with an extensive sitemap, 8 could be fully processed with GPT-4 Turbo and Claude Opus 3. On average, the length of the annotated data for a particular conference has been calculated at 235,13 RDF statements per conference. Notably, while GPT-4 Turbo generated only 47,13 RDF statements per conference on average, Claude Opus 3 could generate 4 times more, on average: 188 statements per conference. This is in line with commonly reported feedback that the model has better performance on needle-in-haystack tasks where it needs to find a particular data point in a large corpus of other information.

Figure 6 gives an example of the annotations generated for the ESWC conference:

Manually going through the annotations and the conference website, we confirm the validity of the created statements, therefore producing the initial assessment that the system can be successfully used to extract structured data from conference websites.

the following prompt: `You will get a list of URLs, and you should exclude every URL which is for any other year than 2023. Return a JSON and only this JSON with filtered URLs. List: [[sitemap]]. Return nothing else but a well-formatted JSON without any extra spaces!`

```
You receive: [A] an academic conference sitemap with
corresponding HTML contents and [B] a RDF ontology describing
conferences (Conference Ontology). Your task is to extract
structured data from [A] using the ontology [B] and return a
well-formed Turtle representation of the contents of the
conference website [C].

[A] Sitemap with HTML contents:

Page: [URL to the page]
Title: [Page title]
HTML: [Page HTML content]
[…]

[B] Conference Ontology:

[Turtle representation of the ontology]

[C] Task:

Return the well-formed representation of the contents of the
conference website using the Conference Ontology [B]. Be
prudent and complete, type Literals using XML Schema, include
every single detail in the result.  Only return the Turtle
formatted representation and nothing else:
```

Figure 5: Composite Prompt

6. Discussion

While many ontologies to structure conference information have been proposed, and there are numerous established embedding techniques for structured data, we have shown that few websites in the area of Computer Science research in general and Semantic Web in particular include usable structured data on their websites. Though some sites are powered by CMS platforms, while they nominally include embedded snippets of structured data, this data is unrelated to the page content and does not reflect semantic information. Furthermore, often these websites do not provide any sitemaps that would help the automatic extraction of contents. The ontology we provide is a step towards capturing this information and can be easily included in any conference website. This work has also shown that the ontology is suitable for the automatic generation of annotations with the help of LLMs.

In terms of the generation of structured content, this paper has shown that, supported with a pre-defined ontology, the leading LLMs can indeed successfully extract structured information from conference websites. However, even between the leading models, the performance varies drastically. Most notably, Claude 3 Opus has been shown to perform much better, extracting

```
@prefix so: <http://semantic.observer/ontology#> .
[... Prefixes]

<https://2023.eswc-conferences.org> a so:Conference ;
  so:title "2023 ESWC-Conferences - 20th ESWC 2023"^^xsd:string ;
  so:shortName "ESWC 2023"^^xsd:string ;
  so:website <https://2023.eswc-conferences.org> ;
  so:startDate "2023-05-28"^^xsd:date ;
  so:endDate "2023-06-01"^^xsd:date ;
  so:location "Hersonissos, Greece"^^xsd:string ;
  so:country "Greece"^^xsd:string ;
  so:topicOfInterest "Semantic Web"^^xsd:string ;
  so:hashtag "eswc_conf"^^xsd:string ;
  so:inceptionYear "2004"^^xsd:gYear ;
  so:editionNumber "20"^^xsd:int ;
  so:fieldOfWork "Semantic Web"^^xsd:string, "Knowledge Graphs"^^xsd:string ;
  so:hasKeynote <https://2023.eswc-conferences.org/keynote-speakers/#MariekEvanErp> ;
  [... Keynotes]
  so:hasDocument <https://2023.eswc-conferences.org/accepted-papers/> ;
  so:sponsoredBy <https://2023.eswc-conferences.org/sponsors-and-supporters/#[Sponsoring company]>;
  [... Sponsors]
  so:organizedBy <https://2023.eswc-conferences.org/organising-committee/#GeneralChair> ;
  so:hasSession <https://2023.eswc-conferences.org/program-overview/#Tutorial1> ;
  [... Sessions] .

<https://2023.eswc-conferences.org/call-for-papers-research-track/> a so:CallForPapers ;
  so:hasTopic "Semantic Web"^^xsd:string ;
  so:hasTopic "Knowledge Graphs"^^xsd:string ;
  so:hasTopic "Ontologies"^^xsd:string ;
  so:pageLimit "15"^^xsd:integer ;
  so:referencesIncluded "true"^^xsd:boolean ;
  so:hasSuppMaterialStatement "true"^^xsd:boolean ;
  so:hasTimezone "UTC-12"^^so:AnywhereOnEarth ;
  so:hasAbstractDeadline [
    a so:AbstractDeadline ;
    so:abstractDeadline "2022-12-08"^^xsd:date
  ] ;
  so:hasPaperSubmission [
    a so:PaperSubmission ;
    so:paperSubmission "2022-12-18"^^xsd:date
  ] ;
  so:hasRebuttalPeriodOpen [
    a so:RebuttalPeriodOpen ;
    so:rebuttalPeriodOpen "2023-01-30"^^xsd:date
  ] ;
  so:hasRebuttalPeriodClose [
    a so:RebuttalPeriodClose ;
    so:rebuttalPeriodClose "2023-02-06"^^xsd:date
  ] ;
  so:hasNotificationDeadline [
    a so:NotificationDeadline ;
    so:notificationDeadline "2023-02-23"^^xsd:date
  ] ;
  so:hasCameraReadyDeadline [
    a so:CameraReadyDeadline ;
    so:cameraReadyDeadline "2023-03-23"^^xsd:date
  ] .

  [... Other CfPs]

  [... Subsequent definitions, subevent titles]
```

Figure 6: Snippet of a Generated Conference Knowledge Graph (ESWC 2023)

a more complete representation of the conference, while GPT-4 Turbo could identify general information and a limited number of additional aspects such as CfPs, related workshops and events.

One of the limitations of the system is the reliance of a large context window of the leading LLMs to include the whole textual representation of a given conference website in the prompt. While currently all sitemaps represented as text could have been included, it must be considered that broader sitemaps might include a larger number of pages and more content, going beyond the threshold. This particular limitation could, however, be solved by the initial segmentation of the sitemap in thematic blocks (main page, CfP-related pages, related event pages, agenda and programme etc.), which correspond to distinct modules of the conference ontology. Another approach would be to chunk the sitemap in content-agnostic parts and iteratively prompt the model with single chunks and the full ontology. This approach would then necessitate combining the generated annotations from a number of consecutive responses, as well as consistency assessment and, potentially, steps of data integration.

7. Conclusion

This work has demonstrated that a RAG architecture supported by a pre-defined ontology and a pre-trained LLMs with a large context window can effectively and reliably extract structured conference information from conference websites. As most conference website do not include useful structured data representing their contents, our approach aims to make the conference information findable, accessible, interoperable and reusable (FAIR) - enabling a variety of smart applications for the research ecosystem, including venue discovery, advanced scientometrics and various types of research assistants. The developed ontology and architecture can be extended to extract information from other academic events (e.g., workshops) and formats (e.g., journal websites).

8. Sustainability plan

We plan to expand the scope of the Conference Knowledge Graph as part of the Semantic Observer platform and produce continuous updates capturing updated information. The sustainability plan includes further extension of the automatic structured data extraction in the frame of future work and other continuing research as well as tight collaboration with other researchers and platforms (e.g., Wikidata).

We are committed to sustainably host and maintain the Conference Knowledge Graph and the Conference Ontology on a standalone basis and through our institute that already hosts various widely adopted Semantic Web resources for several years now and promote the sustainability strategy within ongoing community activities such as the “Distributed Knowledge Graphs” COST Action[33], which as one of its activities aims at aligning and sustaining community services and tools. The resources are made accessible in following ways:

[33] https://cost-dkg.eu/

- The Conference Knowledge Graph is made available via the standalone and institutional repository.[34]
- The SPARQL endpoint provides an accessible way to query the Knowledge Graph.[35]
- The Conference Ontology for describing aspects of academic conference is available via a dedicated information page.[36]

9. Future work

Following the set-out vision to provide the academic community with a reliable conference intelligence platform, the future work includes:

- Evaluating the quality and timeliness of automatic search engine-supported discovery of new conference websites targeting the academic communities of different fields of research.
- Extending the evaluation of the described RAG architecture to further LLMs and broadening the scope to academic conference websites targeting other academic fields.
- Continuous feedback-based improvement of the underlying ontology with the aim of capturing further aspects of conferences relevant to the academic community.
- Extending the scope of the ontology to various formats of further academic venues (workshops, symposia etc.) as well as academic journals (incl. special tracks in academic journals).
- Create clear and accessible guidelines for conference website publishers detailing ways to including the ontology-based annotations in their website.

Acknowledgments

This work has been supported by the *WU Anniversary Fund of the City of Vienna*.

References

[1] S. Lynden, Analysis of semantic URLs to support automated linking of structured data on the web, in: Proceedings of the 7th International Conference on Web Intelligence, Mining and Semantics, WIMS '17, Association for Computing Machinery, New York, NY, USA, 2017, pp. 1–6. URL: https://doi.org/10.1145/3102254.3102265. doi:10.1145/3102254.3102265, 0 citations (Crossref) [2024-03-22].

[2] A. Nuzzolese, A. Gentile, V. Presutti, A. Gangemi, Semantic Web Conference Ontology - A Refactoring Solution, volume 9989, 2016, pp. 84–87. doi:10.1007/978-3-319-47602-5_18, 13 citations (Crossref) [2024-03-22].

[34] https://purl.org/semanticobserver/conference-kg
[35] https://purl.org/semanticobserver/conference-kg-sparql
[36] https://purl.org/semanticobserver/conference-ontology

[3] A. G. Nuzzolese, A. L. Gentile, V. Presutti, A. Gangemi, Conference Linked Data: The ScholarlyData Project, in: P. Groth, E. Simperl, A. Gray, M. Sabou, M. Krötzsch, F. Lecue, F. Flöck, Y. Gil (Eds.), The Semantic Web – ISWC 2016, volume 9982, Springer International Publishing, Cham, 2016, pp. 150–158. URL: https://link.springer.com/10.1007/978-3-319-46547-0_16. doi:10.1007/978-3-319-46547-0_16, series Title: Lecture Notes in Computer Science.

[4] V. Tomberg, D. Lamas, M. Laanpere, W. Reinhardt, J. Jovanovic, Towards a comprehensive call ontology for Research 2.0, in: Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies, ACM, Graz Austria, 2011, pp. 1–8. URL: https://dl.acm.org/doi/10.1145/2024288.2024338. doi:10.1145/2024288.2024338, 3 citations (Crossref) [2024-03-22].

[5] S. Fathalla, C. Lange, EVENTSKG: A Knowledge Graph Representation for Top-Prestigious Computer Science Events Metadata, in: N. T. Nguyen, E. Pimenidis, Z. Khan, B. Trawiński (Eds.), Computational Collective Intelligence, volume 11055, Springer International Publishing, Cham, 2018, pp. 53–63. URL: https://link.springer.com/10.1007/978-3-319-98443-8_6. doi:10.1007/978-3-319-98443-8_6, series Title: Lecture Notes in Computer Science.

[6] R. Wang, Y. Yan, J. Wang, Y. Jia, Y. Zhang, W. Zhang, X. Wang, AceKG: A Large-scale Knowledge Graph for Academic Data Mining, in: Proceedings of the 27th ACM International Conference on Information and Knowledge Management, ACM, Torino Italy, 2018, pp. 1487–1490. URL: https://dl.acm.org/doi/10.1145/3269206.3269252. doi:10.1145/3269206.3269252, 52 citations (Crossref) [2024-03-22].

[7] S. Fathalla, S. Vahdati, C. Lange, S. Auer, SEO: A Scientific Events Data Model, 2019, pp. 79–95. doi:10.1007/978-3-030-30796-7_6.

[8] S. Jeong, H.-G. Kim, SEDE: An ontology for scholarly event description, Journal of Information Science 36 (2010) 0165551509358487. doi:10.1177/0165551509358487, 9 citations (Crossref) [2024-03-22].

[9] K. Möller, T. Heath, S. Handschuh, J. Domingue, Recipes for Semantic Web dog food - The ESWC and ISWC metadata projects (2007). doi:10.1007/978-3-540-76298-0_58, 41 citations (Crossref) [2024-03-22].

[10] D. Dobriy, A. Polleres, Crawley: A Tool for Web Platform Discovery (2023). URL: https://ceur-ws.org/Vol-3632/ISWC2023_paper_496.pdf.

[11] S. Yu, T. Huang, M. Liu, Z. Wang, BEAR: Revolutionizing Service Domain Knowledge Graph Construction with LLM, in: F. Monti, S. Rinderle-Ma, A. Ruiz Cortés, Z. Zheng, M. Mecella (Eds.), Service-Oriented Computing, volume 14419, Springer Nature Switzerland, Cham, 2023, pp. 339–346. URL: https://link.springer.com/10.1007/978-3-031-48421-6_23. doi:10.1007/978-3-031-48421-6_23, series Title: Lecture Notes in Computer Science.

[12] J. Omeliyanenko, A. Zehe, A. Hotho, D. Schlör, CapsKG: Enabling Continual Knowledge Integration in Language Models for Automatic Knowledge Graph Completion, in: T. R. Payne, V. Presutti, G. Qi, M. Poveda-Villalón, G. Stoilos, L. Hollink, Z. Kaoudi, G. Cheng, J. Li (Eds.), The Semantic Web – ISWC 2023, volume 14265, Springer Nature Switzerland, Cham, 2023, pp. 618–636. URL: https://link.springer.com/10.1007/978-3-031-47240-4_33. doi:10.1007/978-3-031-47240-4_33, series Title: Lecture Notes in Computer Science.

[13] B. Veseli, S. Singhania, S. Razniewski, G. Weikum, Evaluating Language Models for Knowledge Base Completion, in: C. Pesquita, E. Jimenez-Ruiz, J. McCusker, D. Faria,

M. Dragoni, A. Dimou, R. Troncy, S. Hertling (Eds.), The Semantic Web, volume 13870, Springer Nature Switzerland, Cham, 2023, pp. 227–243. URL: https://link.springer.com/10.1007/978-3-031-33455-9_14. doi:10.1007/978-3-031-33455-9_14, series Title: Lecture Notes in Computer Science.

[14] L. Yao, J. Peng, C. Mao, Y. Luo, Exploring Large Language Models for Knowledge Graph Completion (2023). URL: https://arxiv.org/abs/2308.13916. doi:10.48550/ARXIV.2308.13916, publisher: arXiv Version Number: 4.

[15] S. Carta, A. Giuliani, L. Piano, A. S. Podda, L. Pompianu, S. G. Tiddia, Iterative Zero-Shot LLM Prompting for Knowledge Graph Construction (2023). URL: https://arxiv.org/abs/2307.01128. doi:10.48550/ARXIV.2307.01128, publisher: arXiv Version Number: 1.

[16] J. Chen, L. Ma, X. Li, N. Thakurdesai, J. Xu, J. H. D. Cho, K. Nag, E. Korpeoglu, S. Kumar, K. Achan, Knowledge Graph Completion Models are Few-shot Learners: An Empirical Study of Relation Labeling in E-commerce with LLMs (2023). URL: https://arxiv.org/abs/2305.09858. doi:10.48550/ARXIV.2305.09858, publisher: arXiv Version Number: 1.

[17] W. Wei, X. Ren, J. Tang, Q. Wang, L. Su, S. Cheng, J. Wang, D. Yin, C. Huang, LLMRec: Large Language Models with Graph Augmentation for Recommendation (2023). URL: https://arxiv.org/abs/2311.00423. doi:10.48550/ARXIV.2311.00423, publisher: arXiv Version Number: 6.

[18] Y. Zhu, X. Wang, J. Chen, S. Qiao, Y. Ou, Y. Yao, S. Deng, H. Chen, N. Zhang, LLMs for Knowledge Graph Construction and Reasoning: Recent Capabilities and Future Opportunities (2023). URL: https://arxiv.org/abs/2305.13168. doi:10.48550/ARXIV.2305.13168, publisher: arXiv Version Number: 2.

[19] Y. Zhang, Z. Chen, W. Zhang, H. Chen, Making Large Language Models Perform Better in Knowledge Graph Completion (2023). URL: https://arxiv.org/abs/2310.06671. doi:10.48550/ARXIV.2310.06671, publisher: arXiv Version Number: 1.

[20] L.-P. Meyer, C. Stadler, J. Frey, N. Radtke, K. Junghanns, R. Meissner, G. Dziwis, K. Bulert, M. Martin, LLM-assisted Knowledge Graph Engineering: Experiments with ChatGPT (2023). URL: https://arxiv.org/abs/2307.06917. doi:10.48550/ARXIV.2307.06917, publisher: arXiv Version Number: 1.

[21] Y. Li, R. Zhang, J. Liu, G. Liu, An Enhanced Prompt-Based LLM Reasoning Scheme via Knowledge Graph-Integrated Collaboration (2024). URL: https://arxiv.org/abs/2402.04978. doi:10.48550/ARXIV.2402.04978, publisher: arXiv Version Number: 1.

[22] K. Wang, Y. Xu, Z. Wu, S. Luo, LLM as Prompter: Low-resource Inductive Reasoning on Arbitrary Knowledge Graphs (2024). URL: https://arxiv.org/abs/2402.11804. doi:10.48550/ARXIV.2402.11804, publisher: arXiv Version Number: 1.

[23] Y. Wen, Z. Wang, J. Sun, MindMap: Knowledge Graph Prompting Sparks Graph of Thoughts in Large Language Models (2023). URL: https://arxiv.org/abs/2308.09729. doi:10.48550/ARXIV.2308.09729, publisher: arXiv Version Number: 4.

[24] F. Moiseev, Z. Dong, E. Alfonseca, M. Jaggi, SKILL: Structured Knowledge Infusion for Large Language Models (2022). URL: https://arxiv.org/abs/2205.08184. doi:10.48550/ARXIV.2205.08184, publisher: arXiv Version Number: 1.