ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Distributed
Computing

ConfSearch 2020

Bachelor's Thesis

Lukas Schmid
luschmi@ethz.ch

Distributed Computing Group
Computer Engineering and Networks Laboratory
ETH Zürich

Supervisors:
Roland Schmid, Pankaj Khanchandani
Prof. Dr. Roger Wattenhofer

May 20, 2020

Acknowledgements

This paper came to life thanks to an idea of the Departement of Distributed Computing at ETH Zurich and my desire to make ConfSearch more easily accessible for both programmers and users. It was a long road of continuously improving a project built from scratch. I had many design-decisions to make and am very glad of the help Pankaj and Roland offered. Despite not being able to meet after the first few weeks, they kept providing insights remotely.

Special thanks to friends and family that helped testing the website and always had an open ear for my ideas.

Abstract

The goal of this paper is to improve the old website at confsearch.ethz.ch with a new and improved version. This required a thorough analysis of the existing website to determine what functionality was in place, how data and user input has been handled and what could be improved upon in future versions. Especially since there was no documentation or any record of the person that initially developed it.

As quoted from the task description, the goal is to help "detect and organize conferences, as it presents deadlines and other relevant metadata conveniently". A big focus here is on the automation. While we always will have to rely on the users for some inputs, there are a lot of reliable websites providing information about upcoming conferences and/or their deadlines. We will scrape some of the most recognized ones to fill the ConfSearch database with the data provided by them.

The project still has work to be done and many of the extensions from the task description could further enhance the appeal to the user. I'm happy with my work towards replicating the functionality of the old website and the transfer of the core features into a newer environment.

Contents

Chapter 1
Introduction

The department of Distributed Computing has a website to display upcoming conferences called ConfSearch. Being based on Java Server Pages with hardly any documentation or visible structure, it was decided that it needs to be rebuilt.
From the many available web-frameworks, python-based Django has been deemed the most suitable one. Similar python frameworks like Flask or Bottle didn't have the depth required and weren't as strong in both community support and projected longevity. A python framework was the preferred choice, as that is where I have the most experience in.
This paper will provide documentation and aim to describe what reasoning lies behind the decisions made in the design process. It will also provide an easier entry-point for future changes.
With the aforementioned issues, nobody knows how the old website worked and where it got the data from. That means that it was not possible to really maintain it. While the inner working of the old site weren't fully mapped out, a basic understanding of the processes behind it was achieved and worked into the development of the new page. This paper will describe all the processes within and introduce a more straight-forward version.

1.1 Related work

There have been websites not unlike ConfSearch, many of which we are pulling data from.
One that stood out as a pretty complete one was from the "Laboratoire d'informatique de l'École polytechnique"[1] hosted by Miki (Nicolas) Hermann.
An Email enquiry then revealed that the website is managed by him personally.
He updates the website daily with information he gathered himself or that he received by email from anyone that wants to submit a conference.
He listed the following reasons for the manual update:

[1] https://www.lix.polytechnique.fr/∼hermann/conf.html

- There are many people on the web who would just add junk, just to have pleasure from destroying my work
- I do not want predator conferences to be listed on my website
- I do not want other doubtful conferences listed there either

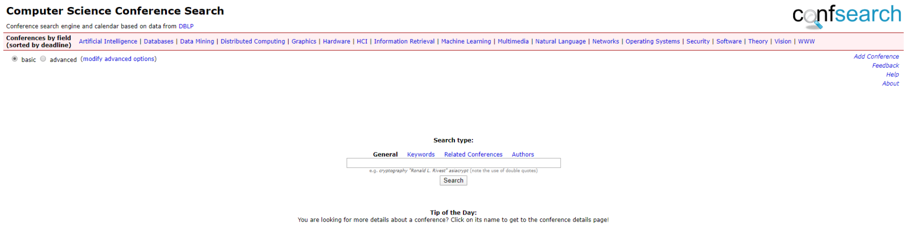
He further emphasized that the website only was supposed to enclose a very narrow range of conferences that belong to his field of study.

Chapter 2
Analysis of ConfSearch

This chapter introduces the old ConfSearch page and the technology defining it. The design will be discussed, as well as how data was processed. I'll further describe the functionality of the old site and it's datastructures.

2.1 Design

Figure 2.1: confsearch.ethz.ch as of 10.5.2020



Computer Science Conference Search

Conference search engine and calendar based on data from DBLP

Conferences by field (sorted by deadline) Artificial Intelligence | Databases | Data Mining | Distributed Computing | Graphics | Hardware | HCI | Information Retrieval | Machine Learning | Multimedia | Natural Language | Networks | Operating Systems | Security | Software | Theory | Vision | WWW

● basic ○ advanced (modify advanced options)

Search type:

General Keywords Related Conferences Authors

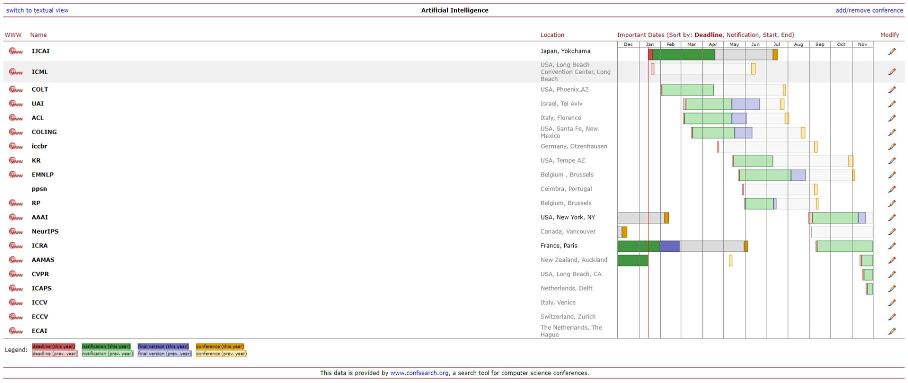
e.g. *cryptography "Ronald L. Rivest" asiacrypt* (note the use of double quotes)

Tip of the Day:
You are looking for more details about a conference? Click on its name to get to the conference details page!

As seen in the image above, the homepage is held quite simple, including mainly a search bar and the different scientific fields by which one can filter.

Figure 2.2 illustrates the main part of website, where the conferences are shown. Each colored part indicates some time constraints concerning the conference. Therefore every line between two bars represents a relevant date.
The dates indicated are in the following order:
Abstract registration, submission, notification, final version, conference start, conference end
This represents all the steps that usually need to be done going from the interest

Figure 2.2: subset of conferences as of 10.5.2020



switch to textual view

Artificial Intelligence

add/remove conference

WWW	Name	Location	Important Dates (Sort by: **Deadline**, Notification, Start, End)	Modify
	IJCAI	Japan, Yokohama		
	ICML	USA, Long Beach Convention Center, Long Beach		
	COLT	USA, Phoenix,AZ		
	UAI	Israel, Tel Aviv		
	ACL	Italy, Florence		
	COLING	USA, Santa Fe, New Mexico		
	iccbr	Germany, Otzenhausen		
	KR	USA, Tempe AZ		
	EMNLP	Belgium , Brussels		
	ppsn	Coimbra, Portugal		
	RP	Belgium, Brussels		
	AAAI	USA, New York, NY		
	NeurIPS	Canada, Vancouver		
	ICRA	France, Paris		
	AAMAS	New Zealand, Auckland		
	CVPR	USA, Long Beach, CA		
	ICAPS	Netherlands, Delft		
	ICCV	Italy, Venice		
	ECCV	Switzerland, Zurich		
	ECAI	The Netherlands, The Hague		

Legend: deadline (this year) notification (this year) final version (this year) conference (this year)
deadline (prev. year) notification (prev. year) final version (prev. year) conference (prev. year)

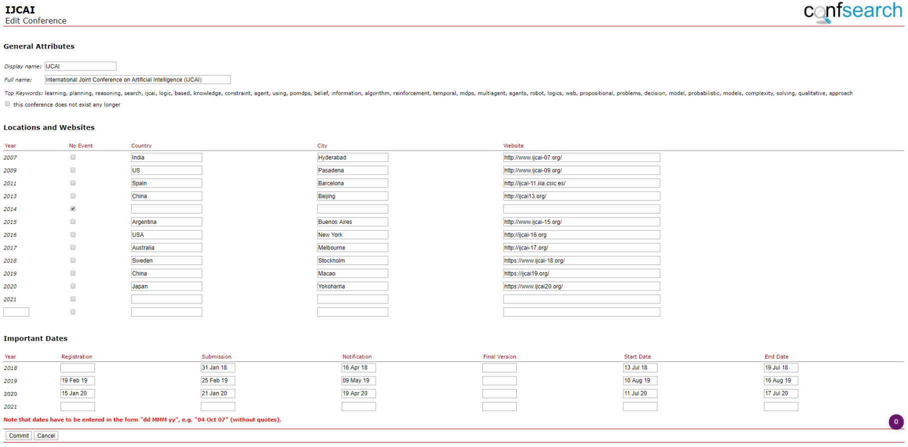
This data is provided by www.confsearch.org, a search tool for computer science conferences.

in a conference to have a paper presented there. The conferences that are more transparent don't have dates defined for this year and are therefore represented by a shadow of last year's data.

The acronym on the left is a link that leads to a more detailed page, where past conferences can be seen and where the conference can be edited by anyone. An example of such a page is shown in Figure 2.3.

The only fields that can't be edited are the Top Keywords and the Rating.

Figure 2.3: conference details as of 10.5.2020



IJCAI
Edit Conference

General Attributes

Display name: IJCAI

Full name: International Joint Conference on Artificial Intelligence (IJCAI)

Top Keywords: learning, planning, reasoning, search, ijcai, logic, based, knowledge, constraint, agent, using, pomdps, belief, information, algorithm, reinforcement, temporal, mdps, multiagent, agents, robot, logics, web, propositional, problems, decision, model, probabilistic, models, complexity, solving, qualitative, approach

☐ this conference does not exist any longer

Locations and Websites

Year	No Event	Country	City	Website
2007	☐	India	Hyderabad	http://www.ijcai-07.org/
2009	☐	US	Pasadena	http://www.ijcai-09.org/
2011	☐	Spain	Barcelona	http://ijcai-11.iiia.csic.es/
2013	☐	China	Beijing	http://ijcai13.org/
2014	☑			
2015	☐	Argentina	Buenos Aires	http://www.ijcai-15.org/
2016	☐	USA	New York	http://ijcai-16.org
2017	☐	Australia	Melbourne	http://ijcai-17.org/
2018	☐	Sweden	Stockholm	https://www.ijcai-18.org/
2019	☐	China	Macao	https://ijcai19.org/
2020	☐	Japan	Yokohama	https://www.ijcai20.org/
2021	☐			
	☐			

Important Dates

Year	Registration	Submission	Notification	Final Version	Start Date	End Date
2018		31 Jan 18	16 Apr 18		13 Jul 18	19 Jul 18
2019	19 Feb 19	25 Feb 19	09 May 19		10 Aug 19	16 Aug 19
2020	15 Jan 20	21 Jan 20	19 Apr 20		11 Jul 20	17 Jul 20
2021						

Note that dates have to be entered in the form "dd MMM yy", e.g. "04 Oct 07" (without quotes).

2.2 Data handling

The database has proven quite extensive, even if some some tables were hardly used or seem to have been there for debugging purposes.

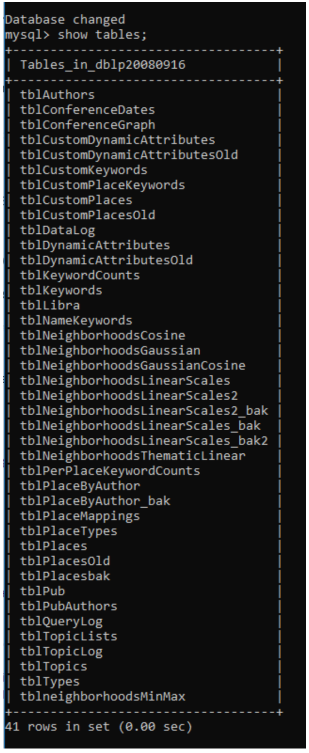
The database in figure 2.4 represents the website https://dblp.org/ from which data has been pulled and saved into a local database at some point. There was

no indication of the table being updated in any of the files found in the websites folder.
The database in figure 2.5 shows where the conference data from the website is stored. Any edits via the website's interface are written here and anything displayed is read from here. The *ConfSearch* module in the website gets it's data from this database, while the supporting modules *DBLPAutomation* and *DBLPParser* read data from local files into the database.
The main file here is *MyDBLPBhtParser.java* which parses the data provided from
DataFiles
DBLP
dblp.xml which seems itself to be replaced manually if at all. The same folder also contains static files for LibraRank, RogerRank, CiteseerRank and ExtNHRank which are all displayed on each conferences detailpage.
The module *DBLPConferencegraph* provides a minimum spanning tree over all conferences, connecting those that are closest based on common keywords. There also seems to be a GraphVisualizer which isn't connected to the website.

2.3 User functionality

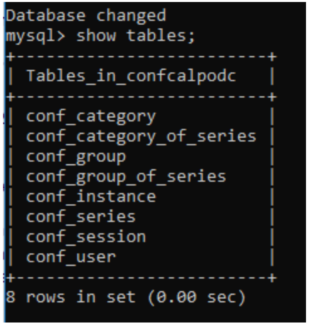
As discussed with the project supervisors who themselves are regular users of the confsearch website, the most important aspect of the site is the gantt chart. It lets users quickly identify the upcoming deadlines and the location of the conference, to help with deciding whether the conference fits the paper they want to release.

Figure 2.4: dblp database



```
Database changed
mysql> show tables;
+-----------------------------------+
| Tables_in_dblp20080916            |
+-----------------------------------+
| tblAuthors                        |
| tblConferenceDates                |
| tblConferenceGraph                |
| tblCustomDynamicAttributes        |
| tblCustomDynamicAttributesOld     |
| tblCustomKeywords                 |
| tblCustomPlaceKeywords            |
| tblCustomPlaces                   |
| tblCustomPlacesOld                |
| tblDataLog                        |
| tblDynamicAttributes              |
| tblDynamicAttributesOld           |
| tblKeywordCounts                  |
| tblKeywords                       |
| tblLibra                          |
| tblNameKeywords                   |
| tblNeighborhoodsCosine            |
| tblNeighborhoodsGaussian          |
| tblNeighborhoodsGaussianCosine    |
| tblNeighborhoodsLinearScales      |
| tblNeighborhoodsLinearScales2     |
| tblNeighborhoodsLinearScales2_bak |
| tblNeighborhoodsLinearScales_bak  |
| tblNeighborhoodsLinearScales_bak2 |
| tblNeighborhoodsThematicLinear    |
| tblPerPlaceKeywordCounts          |
| tblPlaceByAuthor                  |
| tblPlaceByAuthor_bak              |
| tblPlaceMappings                  |
| tblPlaceTypes                     |
| tblPlaces                         |
| tblPlacesOld                      |
| tblPlacesbak                      |
| tblPub                            |
| tblPubAuthors                     |
| tblQueryLog                       |
| tblTopicLists                     |
| tblTopicLog                       |
| tblTopics                         |
| tblTypes                          |
| tblneighborhoodsMinMax            |
+-----------------------------------+
41 rows in set (0.00 sec)
```

Figure 2.5: confcal database



```
Database changed
mysql> show tables;
+-------------------------+
| Tables_in_confcalpodc   |
+-------------------------+
| conf_category           |
| conf_category_of_series |
| conf_group              |
| conf_group_of_series    |
| conf_instance           |
| conf_series             |
| conf_session            |
| conf_user               |
+-------------------------+
8 rows in set (0.00 sec)
```

Chapter 3

Basics of the new website

This chapter describes the idea behind the new website, the tools used and the basis future features are built upon.

3.1 Idea

The rework should make it even easier to decide on a venue to publish on. Due to the age of the old site, newer technology should be used, while still keeping the sleek design principle and retaining the user base. With Django as the new framework, the possibility of keeping the old frontend was considered. While that would have provided some benefits, the effort that would have had to go into creating an interface between the two different system largely outweighed them. Various data sources that have been deemed reliable will provide data towards the website to both stay up-to-date and to encourage users to complete partial entries.

The user experience will also be improved in collaboration with two regular users, the supervisors Roland Schmid and Pankaj Khanchandani. Further potential improvements include recommendation methods, smart completion and user-assisted webscraping.

3.2 Software

While the development VM ran on Ubuntu 18.04, the Django application should be compatible with any system that can run Python 3.x and a compatible website hosting service.

The new website will be running on Django 3.0, the latest stable version with guaranteed long-term support.

Django is a python framework for websites, that is intuitive to code in, modular and clearly states where which part of the website is located. Being based on python, a very well-known and easy to read language, the code should be easy to understand for anyone working on this website in the future.

Both the framework and the language used provide extensive documentation and a lively community, with python being the 3rd most used tag on stackoverflow.com and Django being on rank 30 as of 18.05.2020. HTML, javascript and CSS ranking in-between the two.
With a focus on data aquisition, python provided libraries such as *requests* for HTTPRequests and the external package *beautifulsoup4* to parse HTML data.

3.3 Django Configuration

The new website lives in a django-project called *confsearch*. That relevant files in this folder are *settings.py* and *urls.py*, the other two files are relevant for running and deployment of the server at a later date.
The settings file contains the sites, which are allowed to host the website. Here we have both the IP-Address and the domain name, as well as the names for local access. Furthermore there's a list of plugins called INSTALLED_APPS which contains the default Django modules with the addition of our modules *landingpage* and *manageconf*, as well as the module *background_ task*, which manages the recurring tasks. The second list called MIDDLEWARE lists a number of default security measures that prevent various cyberattacks. The DATABASES list has the credentials for our default and only local postgres database. For security reasons, that database is not accessible from anywhere but localhost.
The urls file contains the information about where certain urls need to be redirected to. It furthermore contains the scheduled task, how often it should be recurring and the specific interval in which it gets called.
The project has two modules called *landingpage* and *manageconf* each in their respective folder. The former defines the website, while the latter holds the models and the option to extend into more detailed managing and smart proposals. The views file in the *landingpage* folder contains the views that are linked by in the urls file, those views always return a rendered template file, which usually is an HTML from the static folder.
The dataGathering folder is independent of the Django installation. It contains python modules that the recurring update uses.

3.4 Data Storage

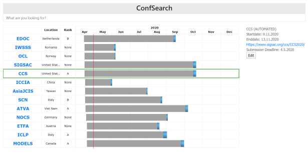
The models currently present in *manageconf* are the conference, tag and author objects. While tag and author are not fully integrated into the front-end, the conference object holds all the fields necessary for the users information.

The database connection is in the config file as mentioned above, currently running PostgreSQL 10.12. The *pg_ hba.conf* file specifies that only local connections are allowed to decrease the risk of outside entities gaining unwanted access

Figure 3.1: Conference model

title	characters	full name
acronym	characters	abbreviation
authors	ManyToMany	authors of all the papers
rank	characters	the rating by the CORE portal
tags	ManyToMany	fields of research
address	characters	physical location
link	characters	webpage
abstractHandinDate	date	date for abstract submission
submissionDate	date	date for paper submission
startDate	date	date of the start
endDate	date	date of the end
isExtracted	boolean	Whether it was scraped
lastUpdated	date	date of last change for this entry

to the database. Commands with user input still need to be sanitized, as the website backend does have ownership of the confsearchdb database. The user is called luschmi and has no other privileges over other local databases.



	Location	Rank
EDOC	Netherlands	B
IWSSS	Romania	None
OCL	Norway	None
SIGSAC	United Stat...	None
CCS	United Stat...	A
ICCIA	China	None
AsiaJCIS	Taiwan	None
SCN	Italy	B
ATVA	Viet Nam	A
NOCS	Germany	None
ETFA	Austria	None
ICLP	Italy	A
MODELS	Canada	A

CCS (AUTOMATED)
Startdate: 9.11.2020
Enddate: 13.11.2020
https://www.sigsac.org/ccs/CCS2020/
Submission Deadline: 4.5.2020

3.5 Javascript libraries

The *fuzzyset*[1] library was used to calculate the levenshtein distance between two strings to sort by similarity to the searchterm.

The *JSGantt*[2] library provided a basic gantt chart to build upon. Major modifications had to be made to this module to modify the look and behaviour according to our goals.
Among many different gantt chart libraries, this one had the most similarity with the idea we wanted to achieve and a documentation that promised a high degree of customizability.

[1] https://glench.github.io/fuzzyset.js/

[2] https://jsganttimproved.github.io/jsgantt-improved/

The *jQuery*[3] library allowd for more dynamic changes to the document and its functions.

3.6 The current state

The website currently lives on a virtual machine hosted on *ee-tik-vm012.ethz.ch* and is being run as a development server. That means it is only reachable by localhost or SSH tunnels into the virtual machine, as it is neither secure nor stable in that state.
To be made available to external users, it needs to be installed on a webserver with a supported deployment platform. One of those can be found in the Django documentation[4].

Because the scraping of the websites happens in parallel, two separate processes need to be started for the full functionality.

```
python3 manage.py runserver 0.0.0.0:8000
python3 manage.py process_tasks
```

The process_tasks command comes from an external library called Django Background Tasks[5] that allows to run commands in predefined intervals.

[3] https://jquery.com/

[4] https://docs.djangoproject.com/en/2.1/howto/deployment/

[5] https://django-background-tasks.readthedocs.io/

Chapter 4

Automated retrieval of conference data

This chapter describes the methods used to retrieve conference data from various websites. All the methods take no arguments and return a tuple containing the title of the conference, it's acronym, the CORE-rank[1], the latest submissiondate for the paper, the startdate and enddate of the conference, both the place and the country where it happens and finally the link to the conferences homepage.

4.1 Association for Computing Machinery

As is obvious from Figures 4.1[2] and 4.2[3] both websites are built very similarily, which also let's the first part of the two scraping functions look pretty similar. Both websites HTML content is requested via the python requests package. Two parameters have been identified to be relevant.

startDate0 is a date in the format YYYYMMDD and identifies the month that will be displayed.
view0 defines the format of the page which can either be day, week or month. To lessen the amount of requests required, we'll leave it at month.
eventType0 is CallsToPapers for the submissions page and Conferences for the conference-events page.

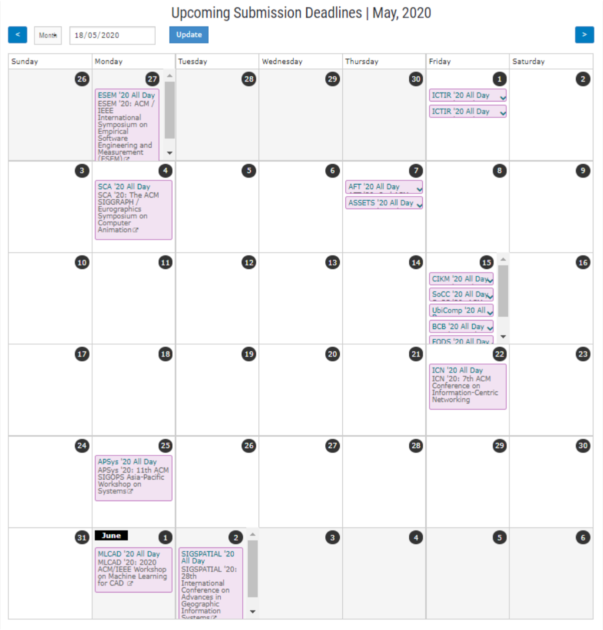
Using beautifulsoup to get a queryable object, one can easily isolate all **li** tags with the "day" class. This leaves us with a list of pure day classes and some "day other-month" classes at both ends that indicate the previous and the next month.
We get the submission date from either counting through the pure day classes

[1] http://www.core.edu.au/conference-portal

[2] https://www.acm.org/conferences/upcoming-submission-deadlines?startDate0=20200520&eventType0=CallsForPa

[3] https://www.acm.org/conferences/conference-events?startDate0=20200520&eventType0=Conferences

Figure 4.1: ACM's submission page



Upcoming Submission Deadlines | May, 2020

< Month 18/05/2020 Update >

Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
26	27 ESEM '20 All Day ESEM '20: ACM / IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)	28	29	30	1 ICTIR '20 All Day ICTIR '20 All Day	2
3	4 SCA '20 All Day SCA '20: The ACM SIGGRAPH / Eurographics Symposium on Computer Animation	5	6	7 AFT '20 All Day ASSETS '20 All Day	8	9
10	11	12	13	14	15 CIKM '20 All Day SoCC '20 All Day UbiComp '20 All BCB '20 All Day FODS '20 All Day	16
17	18	19	20	21	22 ICN '20 All Day ICN '20: 7th ACM Conference on Information-Centric Networking	23
24	25 APSys '20 All Day APSys '20: 11th ACM SIGOPS Asia-Pacific Workshop on Systems	26	27	28	29	30
31	June 1 MLCAD '20 All Day MLCAD '20: 2020 ACM/IEEE Workshop on Machine Learning for CAD	2 SIGSPATIAL '20 All Day SIGSPATIAL '20: 28th International Conference on Advances in Geographic Information Systems	3	4	5	6

or reading the number in the field, we can then extract all events by looking for links as **a** tags within the **li** objects we found. Those links directly point to the individual conference sites, which we can read directly from the href attribute. Furthermore we can split the title attribute at the ":" to get both the acronym and the conference title.

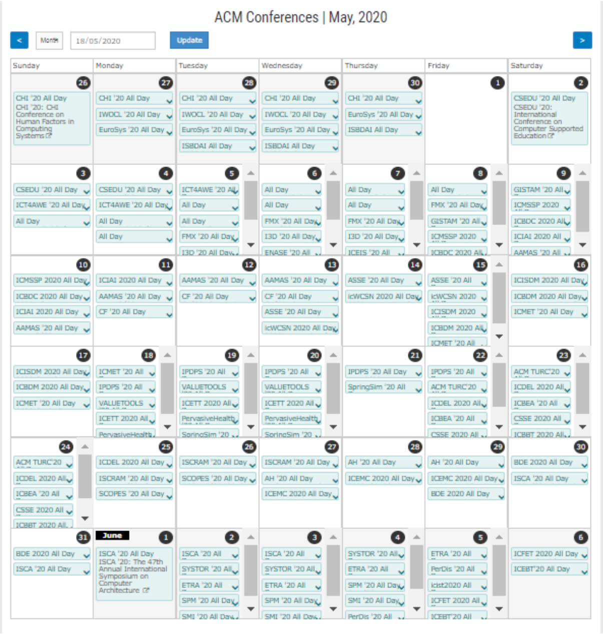
For the conference website we need to add a post-processing step, that detects consecutive days with the same conference, so they can be merged into one database entry.

4.2 Guide2research

This page[4] has an even stronger partition into list elements, but does feature more condensed data which we need to separate.
To get all the rows we can iterate over all the beautifulsoup objects that admit to the type *div* and the class "grey myshad" which is the specific coloring of the list elements. We do have to start from the second element though, as the first one contains the headings.
Using a RegEx-Search we can identify the title and split it into the acronym and

[4] http://www.guide2research.com/conferences/page-1

Figure 4.2: ACM's conference page



ACM Conferences | May, 2020

< Month 18/05/2020 Update >

Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
26 CHI '20 All Day CHI '20: CHI Conference on Human Factors in Computing Systems	27 CHI '20 All Day IWOCL '20 All Day EuroSys '20 All Day	28 CHI '20 All Day IWOCL '20 All Day EuroSys '20 All Day ISBDAI All Day	29 CHI '20 All Day IWOCL '20 All Day EuroSys '20 All Day ISBDAI All Day	30 CHI '20 All Day EuroSys '20 All Day ISBDAI All Day	1	2 CSEDU '20 All Day CSEDU '20: International Conference on Computer Supported Education
3 CSEDU '20 All Day ICT4AWE '20 All Day All Day	4 CSEDU '20 All Day ICT4AWE '20 All Day All Day All Day	5 ICT4AWE '20 All All Day All Day FMX '20 All Day I3D '20 All Day	6 All Day All Day FMX '20 All Day I3D '20 All Day ENASE '20 All	7 All Day All Day FMX '20 All Day I3D '20 All Day ICEIS '20 All	8 All Day FMX '20 All Day GISTAM '20 All ICMSSP 2020 ICBDC 2020 All	9 GISTAM '20 All ICMSSP 2020 ICBDC 2020 All ICIAI 2020 All AAMAS '20 All
10 ICMSSP 2020 All Day ICBDC 2020 All Day ICIAI 2020 All Day AAMAS '20 All Day	11 ICIAI 2020 All Day AAMAS '20 All Day CF '20 All Day	12 AAMAS '20 All Day CF '20 All Day	13 AAMAS '20 All Day CF '20 All Day ASSE '20 All Day icWCSN 2020 All Day	14 ASSE '20 All Day icWCSN 2020 All Day	15 ASSE '20 All icWCSN 2020 ICISDM 2020 ICBDM 2020 All ICMET '20 All	16 ICISDM 2020 All Day ICBDM 2020 All Day ICMET '20 All Day
17 ICISDM 2020 All Day ICBDM 2020 All Day ICMET '20 All Day	18 ICMET '20 All IPDPS '20 All VALUETOOLS ICETT 2020 All PervasiveHealth	19 IPDPS '20 All VALUETOOLS ICETT 2020 All PervasiveHealth SpringSim '20	20 IPDPS '20 All VALUETOOLS ICETT 2020 All PervasiveHealth SpringSim '20	21 IPDPS '20 All Day SpringSim '20 All	22 IPDPS '20 All ACM TURC'20 ICDEL 2020 All ICBEA '20 All CSSE 2020 All	23 ACM TURC'20 ICDEL 2020 All ICBEA '20 All CSSE 2020 All ICBBT 2020 All
24 ACM TURC'20 ICDEL 2020 All ICBEA '20 All CSSE 2020 All ICBBT 2020 All	25 ICDEL 2020 All Day ISCRAM '20 All Day SCOPES '20 All Day	26 ISCRAM '20 All Day SCOPES '20 All Day	27 ISCRAM '20 All Day AH '20 All Day ICEMC 2020 All Day	28 AH '20 All Day ICEMC 2020 All Day	29 AH '20 All Day ICEMC 2020 All Day BDE 2020 All Day	30 BDE 2020 All Day ISCA '20 All Day
31 BDE 2020 All Day ISCA '20 All Day	June 1 ISCA '20 All Day ISCA '20: The 47th Annual International Symposium on Computer Architecture	2 ISCA '20 All SYSTOR '20 All ETRA '20 All SPM '20 All Day SMI '20 All Day	3 ISCA '20 All SYSTOR '20 All ETRA '20 All SPM '20 All Day SMI '20 All Day	4 SYSTOR '20 All ETRA '20 All SPM '20 All Day SMI '20 All Day PerDis '20 All	5 ETRA '20 All PerDis '20 All icist2020 All ICFET 2020 All ICEBT'20 All	6 ICFET 2020 All Day ICEBT'20 All Day

the conference-name. For the submissiondate on the right, the *datetime.strptime* function "%a %d %b %Y" can translate weekday, date, month and year into a date object, which we then write to the database. Using another RegEx-Search, we split the conference startdate, enddate and city into three elements, parsing the dates with the same formula as above. Another RegEx search then results in the country and therefore the final field from this page.
Since this website has had malformed dates before, we make some sanity checks to verify they have the correct order and the conference has a reasonable length.

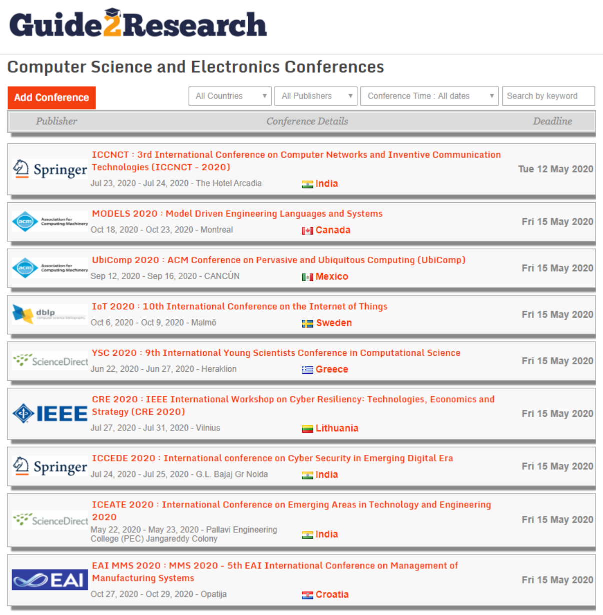
4.3 IEEE

While IEEE's Computational Intelligence Society[5] doesn't provide submission dates for their papers, they still had a lot of information about conferences and the links to the webpages. Even though for the current extent of the website, this data isn't that useful, it can still be completed by findings of other websites or user input.
The list consists of *div* elements with the class *conf-full-item*, which allows for easy extraction. Each row is formatted as "Conference (Acronym)" with it's location in a *div* element and the dates in a span element. The dates are in the Unix Epoch[6] format and therefore require a different parsing function.

[5] https://cis.ieee.org/conferences/conference-calendar

[6] Number of seconds since 00:00:00 UTC on 1 January 1970

Figure 4.3: Guide2Research's conference page



Guide2Research

Computer Science and Electronics Conferences

Publisher	Conference Details	Deadline
Springer	ICCNCT : 3rd International Conference on Computer Networks and Inventive Communication Technologies (ICCNCT - 2020) Jul 23, 2020 - Jul 24, 2020 - The Hotel Arcadia India	Tue 12 May 2020
Association for Computing Machinery	MODELS 2020 : Model Driven Engineering Languages and Systems Oct 18, 2020 - Oct 23, 2020 - Montreal Canada	Fri 15 May 2020
Association for Computing Machinery	UbiComp 2020 : ACM Conference on Pervasive and Ubiquitous Computing (UbiComp) Sep 12, 2020 - Sep 16, 2020 - CANCÚN Mexico	Fri 15 May 2020
dblp	IoT 2020 : 10th International Conference on the Internet of Things Oct 6, 2020 - Oct 9, 2020 - Malmö Sweden	Fri 15 May 2020
ScienceDirect	YSC 2020 : 9th International Young Scientists Conference in Computational Science Jun 22, 2020 - Jun 27, 2020 - Heraklion Greece	Fri 15 May 2020
IEEE	CRE 2020 : IEEE International Workshop on Cyber Resiliency: Technologies, Economics and Strategy (CRE 2020) Jul 27, 2020 - Jul 31, 2020 - Vilnius Lithuania	Fri 15 May 2020
Springer	ICCEDE 2020 : International conference on Cyber Security in Emerging Digital Era Jul 24, 2020 - Jul 25, 2020 - G.L. Bajaj Gr Noida India	Fri 15 May 2020
ScienceDirect	ICEATE 2020 : International Conference on Emerging Areas in Technology and Engineering 2020 May 22, 2020 - May 23, 2020 - Pallavi Engineering College (PEC) Jangareddy Colony India	Fri 15 May 2020
EAI	EAI MMS 2020 : MMS 2020 - 5th EAI International Conference on Management of Manufacturing Systems Oct 27, 2020 - Oct 29, 2020 - Opatija Croatia	Fri 15 May 2020

4.4 Call For Papers Wiki

The HTML elements on this page[7] are nearly indistinguishable, so I had to use some form of reliable index. Our assumption here is that there will always be 5 conference links in the Popular CFPs, so the sixth conference link can be used to identify the table relevant to our query.

After getting the grand-grand-parent of that element, the rows have to be parsed pairwise, as each differently shaded row consists of two HTML-rows. The first yields both the acronym and the name of the conference, the second one provides start date, end date and the location.

4.5 CORE Conference Portal

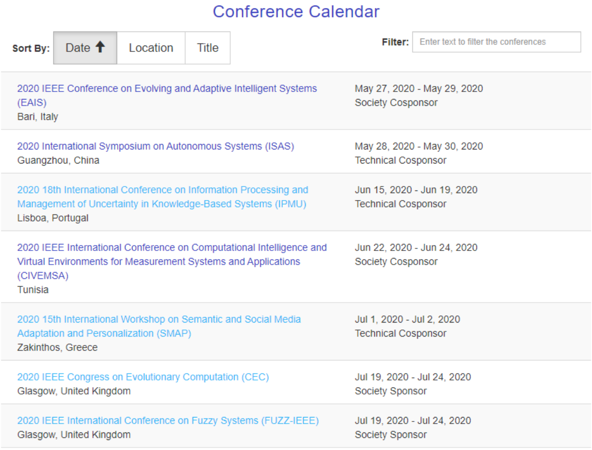
This page[8] provides the rankings displayed on the new website. While all the other pages had a limit on how far into the future we'd parse them, this page gets read fully every time. The page indicators at the bottom serve as indicator how many pages we'll have to loop through.

Iterating through all table rows except the header proves easy, as the website is very minimalistic, the rows of interest are the acronym used to match to the conferences already in our database and the rank. The rank is checked against it's

[7] http://www.wikicfp.com/cfp/

[8] http://portal.core.edu.au/conf-ranks/?search=by=all

Figure 4.4: IEEE's conference page



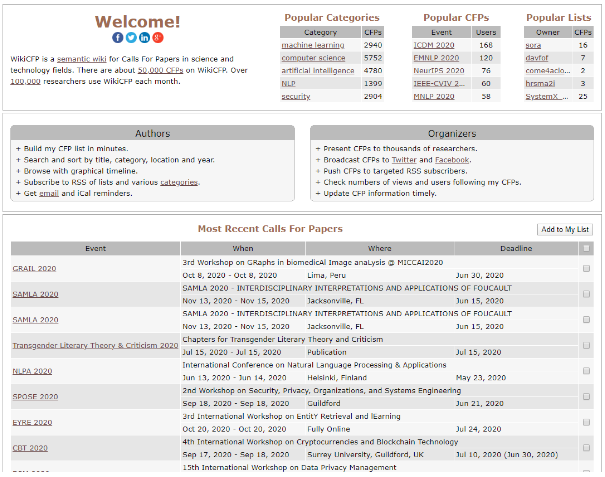
Conference Calendar

Sort By: Date ↑ | Location | Title

Filter: Enter text to filter the conferences

2020 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS) Bari, Italy	May 27, 2020 - May 29, 2020 Society Cosponsor
2020 International Symposium on Autonomous Systems (ISAS) Guangzhou, China	May 28, 2020 - May 30, 2020 Technical Cosponsor
2020 18th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU) Lisboa, Portugal	Jun 15, 2020 - Jun 19, 2020 Technical Cosponsor
2020 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA) Tunisia	Jun 22, 2020 - Jun 24, 2020 Society Cosponsor
2020 15th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP) Zakinthos, Greece	Jul 1, 2020 - Jul 2, 2020 Technical Cosponsor
2020 IEEE Congress on Evolutionary Computation (CEC) Glasgow, United Kingdom	Jul 19, 2020 - Jul 24, 2020 Society Sponsor
2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE) Glasgow, United Kingdom	Jul 19, 2020 - Jul 24, 2020 Society Sponsor

Figure 4.5: WikiCFP's conference page



Welcome!

WikiCFP is a semantic wiki for Calls For Papers in science and technology fields. There are about 50,000 CFPs on WikiCFP. Over 100,000 researchers use WikiCFP each month.

Popular Categories

Category	CFPs
machine learning	2940
computer science	5752
artificial intelligence	4780
NLP	1399
security	2904

Popular CFPs

Event	Users
ICDM 2020	168
EMNLP 2020	120
NeurIPS 2020	76
IEEE-CVIV 2...	60
MNLP 2020	58

Popular Lists

Owner	CFPs
sora	16
davfof	7
come4aclo...	2
hrsma2i	3
SystemX ...	25

Authors

+ Build my CFP list in minutes.
+ Search and sort by title, category, location and year.
+ Browse with graphical timeline.
+ Subscribe to RSS of lists and various categories.
+ Get email and iCal reminders.

Organizers

+ Present CFPs to thousands of researchers.
+ Broadcast CFPs to Twitter and Facebook.
+ Push CFPs to targeted RSS subscribers.
+ Check numbers of views and users following my CFPs.
+ Update CFP information timely.

Most Recent Calls For Papers

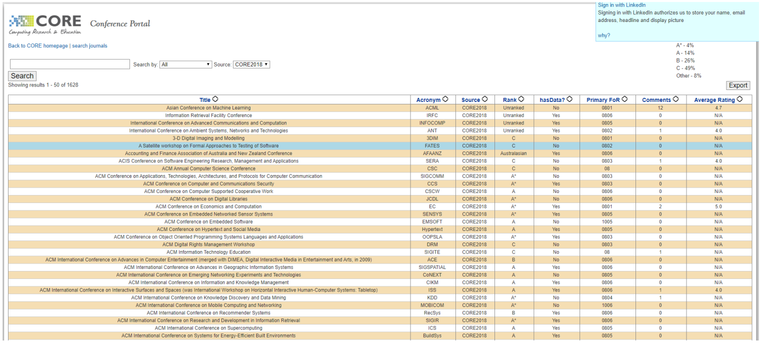
Add to My List

Event	When	Where	Deadline	□
GRAIL 2020	3rd Workshop on GRaphs in biomedicAl Image anaLysis @ MICCAI2020 Oct 8, 2020 - Oct 8, 2020	Lima, Peru	Jun 30, 2020	□
SAMLA 2020	SAMLA 2020 - INTERDISCIPLINARY INTERPRETATIONS AND APPLICATIONS OF FOUCAULT Nov 13, 2020 - Nov 15, 2020	Jacksonville, FL	Jun 15, 2020	□
SAMLA 2020	SAMLA 2020 - INTERDISCIPLINARY INTERPRETATIONS AND APPLICATIONS OF FOUCAULT Nov 13, 2020 - Nov 15, 2020	Jacksonville, FL	Jun 15, 2020	□
Transgender Literary Theory & Criticism 2020	Chapters for Transgender Literary Theory and Criticism Jul 15, 2020 - Jul 15, 2020	Publication	Jul 15, 2020	□
NLPA 2020	International Conference on Natural Language Processing & Applications Jun 13, 2020 - Jun 14, 2020	Helsinki, Finland	May 23, 2020	□
SPOSE 2020	2nd Workshop on Security, Privacy, Organizations, and Systems Engineering Sep 18, 2020 - Sep 18, 2020	Guildford	Jun 21, 2020	□
EYRE 2020	3rd International Workshop on EntitY Retrieval and lEarning Oct 20, 2020 - Oct 20, 2020	Fully Online	Jul 24, 2020	□
CBT 2020	4th International Workshop on Cryptocurrencies and Blockchain Technology Sep 17, 2020 - Sep 18, 2020	Surrey University, Guildford, UK	Jul 10, 2020 (Jun 30, 2020)	□
	15th International Workshop on Data Privacy Management			□

length, as there are four kinds of unranked conferences, including Australasian, Unranked, National and Regional.

This function returns the acronym and the rank of the conference.

Figure 4.6: CORE's conference page



Sign in with LinkedIn
Signing in with LinkedIn authorizes us to store your name, email address, headline and display picture

why?

Back to CORE homepage | search journals

A* - 4%
A - 14%
B - 26%
C - 49%
Other - 8%

Search by: All ▾ Source: CORE2018 ▾

Showing results 1 - 50 of 1628

Title	Acronym	Source	Rank	hasData?	Primary FoR	Comments	Average Rating
Asian Conference on Machine Learning	ACML	CORE2018	Unranked	No	0801	12	4.7
Information Retrieval Facility Conference	IRFC	CORE2018	Unranked	Yes	0806	0	N/A
International Conference on Advanced Communications and Computation	INFOCOMP	CORE2018	Unranked	Yes	0805	0	N/A
International Conference on Ambient Systems, Networks and Technologies	ANT	CORE2018	Unranked	Yes	0802	1	4.0
3-D Digital Imaging and Modelling	3DIM	CORE2018	C	No	0801	0	N/A
A Satellite workshop on Formal Approaches to Testing of Software	FATES	CORE2018	C	No	0802	0	N/A
Accounting and Finance Association of Australia and New Zealand Conference	AFAANZ	CORE2018	Australasian	Yes	0806	0	N/A
ACIS Conference on Software Engineering Research, Management and Applications	SERA	CORE2018	C	No	0803	1	4.0
ACM Annual Computer Science Conference	CSC	CORE2018	C	No	08	0	N/A
ACM Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication	SIGCOMM	CORE2018	A*	No	0803	0	N/A
ACM Conference on Computer and Communications Security	CCS	CORE2018	A*	Yes	0803	0	N/A
ACM Conference on Computer Supported Cooperative Work	CSCW	CORE2018	A	No	0806	0	N/A
ACM Conference on Digital Libraries	JCDL	CORE2018	A*	No	0806	0	N/A
ACM Conference on Economics and Computation	EC	CORE2018	A*	Yes	0801	2	5.0
ACM Conference on Embedded Networked Sensor Systems	SENSYS	CORE2018	A*	Yes	0805	0	N/A
ACM Conference on Embedded Software	EMSOFT	CORE2018	A	No	1005	0	N/A
ACM Conference on Hypertext and Social Media	Hypertext	CORE2018	A	Yes	0805	0	N/A
ACM Conference on Object Oriented Programming Systems Languages and Applications	OOPSLA	CORE2018	A*	Yes	0803	0	N/A
ACM Digital Rights Management Workshop	DRM	CORE2018	C	No	0803	0	N/A
ACM Information Technology Education	SIGITE	CORE2018	C	No	08	1	N/A
ACM International Conference on Advances in Computer Entertainment (merged with DIMEA, Digital Interactive Media in Entertainment and Arts, in 2009)	ACE	CORE2018	B	No	0806	0	N/A
ACM International Conference on Advances in Geographic Information Systems	SIGSPATIAL	CORE2018	A	Yes	0806	0	N/A
ACM International Conference on Emerging Networking Experiments and Technologies	CoNEXT	CORE2018	A	No	0805	0	N/A
ACM International Conference on Information and Knowledge Management	CIKM	CORE2018	A	Yes	0806	0	N/A
ACM International Conference on Interactive Surfaces and Spaces (was International Workshop on Horizontal Interactive Human-Computer Systems: Tabletop)	ISS	CORE2018	A	Yes	0806	1	4.0
ACM International Conference on Knowledge Discovery and Data Mining	KDD	CORE2018	A*	No	0804	1	N/A
ACM International Conference on Mobile Computing and Networking	MOBICOM	CORE2018	A*	No	1006	0	N/A
ACM International Conference on Recommender Systems	RecSys	CORE2018	B	Yes	0806	0	N/A
ACM International Conference on Research and Development in Information Retrieval	SIGIR	CORE2018	A*	Yes	0806	0	N/A
ACM International Conference on Supercomputing	ICS	CORE2018	A	Yes	0805	0	N/A
ACM International Conference on Systems for Energy-Efficient Built Environments	BuildSys	CORE2018	A	Yes	0805	0	N/A

Chapter 5
Design Decisions

This chapter will talk about the higher-level decisions made during developement.

5.1 Logical Separation

To increase modularity of the code, every step from the conference data to the display on the webpage has it's own space.
In the folder called dataGathering, the conferences are read from their respective sites and coerced into the required datatypes and one global format, which will later allow us to iterate over all the generators in one loop. If some type of data is not provided from a specific source, the values will be returned as None, while still maintaining the agreed upon format.

Figure 5.1: Scraped conference data

	G2R	ACM	IEEE	WikiCFP	CORE
title	x	x	x	x	
acronym	x	x	x	x	x
rating					x
submissionDate	x	x		x	
startDate	x	x	x	x	
endDate	x	x	x	x	
location details	x		x	x	
country	x		x	x	
link	x	x	x	x	

The data then gets processed by the urls.py file in the main confsearch folder. This is the place where all the routing happens. It's an interface to the outside world that defines where HTTP GET requests for the page go and where the

data gathered from external websites goes.
It's folder contains all the configuration for the website in the settings.py file.

All parts concerning the actual website are located in the landingpage folder. The admin and models file import the models from the manageconf folder, where the model definitions are located.
The views.py file defines all direct interactions between our webpage and the database. It also renders the webpage by replacing all Django level variables and code with the specified data. Since there is only one page, there is only one template in the templates folder, which defines the HTML and Javascript needed to display the data given by it's view.

5.2 Visuals

The visuals were consciously kept simple without colorful CSS frameworks. The display focus is on Desktop PCs, no testing was done on smartphones.
For the table and Gantt chart within, the JSGantt[1] open source library was used. It was heavily modified during development to adhere to our vision of the website. Only the rows crucial for making decisions about the conference were kept with some additional data displaying once the users hovers or clicks the conference. These were the location and rank, along with the acronym and timeline.
The acronym links to the webpage of the conference to allow the users to verify the dates and quickly submit their paper. In case that no link was found, the users will be redirected to google, automatically searching for the conference in the current year.
The look of the gantt bar was adjusted to mimic that of the original confsearch site.

[1] https://jsganttimproved.github.io/jsgantt-improved/

Chapter 6

Handling user input

Since the automated data collection will likely not catch every information there is about a conference, the website allows for users to complete existing entries with additional data.

6.1 Editing conferences

Each entry has a "Edit" button, that is accessible to any user on the website.
The assumption is that people using the website will be willing to contribute to it, so the community can improve the amount of information available. If that functionality gets abused, a login system should be introduced to restrict access to verified users that can be held accountable for their edits.
Editing a conference will directly write the data to the database and notify the user that the operation was successful. Due to the current status of the gantt table, the website needs to be reloaded to display the changed data.
Once a conference has been edited, it will not get updated by the scripts anymore. Any conferences in their original form which get found with new data will get overwritten.

6.2 Creating conferences

The website currently doesn't allow the creation of new conferences. If any noticeable conferences are missing, they can be created via the admin page, requiring a admin issued account.
Having all the biggest sites as data sources, we assume that all reputable conferences already exist and only need to be edited.

Chapter 7

Discussion

The goal was to create a website that improved upon the old one in terms of technology, presentation and data acquisition. Django introduced a more recent, more readable framework and language. It also allowed to strictly separate the different layers of application, so we have clear interfaces and every part can be changed independently of everything else.

The removal of the individual pages for each conference made the experience more predictable, but also reduced the amount of information gained. This was necessary to improve the efficiency of the users search.

Tags and the ability to search by field has temporarily been removed, but should be re-introduced if the need arises in the future.

7.1 The Gantt library

The javascript library *JSGantt* used in the website was responsible for both the table with the acronym, the location and the rating of the conference, as well as the Gantt chart in the right-hand side of the table. It worked in such a way, that the left and right table were disconnected and only seemed connected through the same CSS formatting that keeps both sides aligned.

This caused various issues late in the project when trying to sort the table. Those issues were solved in two different ways. The default sorting was shifted to the Django view module that delivered the data. Since the data was already retrieved sorted, the table would be sorted too once created. For local sorting recreating the gantt object is more efficient than sorting both sides individually while trying to ensure they line up. For future changes one might consider further changing the *JSGantt* library to create a globally consistent ID for both the tables and the edit functionality, so that only the HTML elements can be reordered to improve on the performance.

Another issue here is, that the library only allows one bar per task, so the displayed conferences are actually multiple tasks grouped up on one line. The grouping works by having them after each other in the list of tasks while the later tasks

reference the parent by its ID. For the sorting to work on that level, all children have to be removed, the parents get sorted and subsequently the children of each parent get added back. This could be improved in speed by adding the same data to all children and using a stable sorting algorithm.
The data being handled this way also is a reason for the data not instantly being updated upon editing a conference. It would either have to update the javascript list entry and all its children or check which row is highlighted and translate the changes into changes there and a redrawing of the gantt bar on the right side.

With many features of the library unused and a plethora of changes made, it should be considered that writing a new library from scratch or finding one that fits this usecase better may improve future results.

Chapter 8

Conclusion

While the main goals have been achieved, there still remains some room for improvements within the website. The design is more modern and presents the data without the extra steps required in the old version.

With the various websites providing data, there is no doubt that over time new corner cases will appear and require some sort of user intervention. With the possibility that every user can identify such issues and correct them, there is a lot of potential for a clean and complete database. While manually adding conferences is the exception now, some of the websites providing data might shut down. That should not impede the data collection, but may lead to some conferences going missing. With the structure built, newly appearing sites are easy to add and can be seamlessly integrated into the existing dataflow.

This automation shows that a lot of data can be gathered without requiring user interaction and only requires minimal interaction to stay up-to-date for an arbitrarily long time.