Foundations of Trustworthy AI – Integrating Reasoning, Learning and Optimization
TAILOR
Grant Agreement Number 952215

AI-Driven Collaboration Tools v.1 Report

Document type (nature)	Report
Deliverable No	D9.1
Work package number(s)	WP9
Date	Due M24
Responsible Beneficiary	UNIBRIS, ID #16
Author(s)	Peter Flach and Miquel Perello Nieto
Publicity level	Public
Short description	This deliverable provides a public demonstrator of the SubSift matching algorithm.

History			
Revision	**Date**	**Modification**	**Author**
1.0	14/10/2022	First version	PF, MPN

Document Review		
Reviewer	**Partner ID / Acronym**	**Date of report approval**
Gabriel Gonzalez Castane	4, UCC	22/11/2022
Silke Balzert-Walter	26, DFKI	22/11/2022

Table of Contents

Summary of the report

This report describes the SubSift matching algorithm as an AI-tool that can be used to facilitate the collaboration and networking of researchers. This tool allows to rank a set of people by expertise with respect to a provided text. This can be used to find potential collaborations that were previously unknown, or even to match people with similar interests. In preparation for this deliverable we have applied the algorithm successfully in two TAILOR events, which is described in their respective deliverables and summarised in this report. In this deliverable we provide an online public demonstrator v.1.0 of the tool, which can be used by researchers to find possible collaborators among the TAILOR partner representatives. We expect TAILOR and non-TAILOR members to use the tool to explore the expertise of the TAILOR partners to potentially start new collaborations. Furthermore, in the appendix we present information about a workshop organised by TAILOR which provides additional AI-tools to facilitate the collaboration of AI researchers.

Organisation

The following people have been involved in the Deliverable:

Partner ID / Acronym	Name	Role
ID #16, UNIBRIS	Miquel Perello Nieto	Researcher
ID #16, UNIBRIS	Peter Flach	WP Lead

Introduction

Task 9.1 of Work Package 9 concerns developing AI-powered tools facilitating collaboration between AI researchers. SubSift [1] is a matching tool that was developed in Bristol since 2009 and has been brought into the TAILOR project as background IP. SubSift has been used in the TAILOR network to match people with similar interests at the second TAILOR Summer School (Barcelona, June 2022) and the second TAILOR conference (Prague, September 2022). We have adapted it here for a demonstrator, showing how arbitrary text can be matched to people in the TAILOR network. The aim of this public demonstrator is to familiarise people with the tool and collect feedback. We provide details about the underlying pipeline, the used algorithm, and describe all the functionalities of the website. Finally, we draw some conclusions on the experience of using the tool and the next steps in preparation to deliverable D9.2.

Summary of use cases

The tools described in this report have been used in two previous deliverables in order to test their feasibility and enhance the collaboration among attendees of the 2nd TAILOR Summer School (Deliverable 9.8, Appendix 3: SubSift Summary) and the 2nd TAILOR Conference (Deliverable 1.4, Appendix 2: Group Work). A detailed description of the methodology, and overall impressions by the participants and organisers can be found in these deliverable reports. In this Section, we provide a short overview for completeness.

In preparation of both events, WP9 has explored AI-tools to enhance the networking and collaboration experiences of attendees. The selected two tools are SubSift [1] and SynTeam [2], from partners UNIVBRIS and IIIA-CSIC respectively.

SubSift is a tool to match profiles represented as a bag of words, for example a personal profile, a bibliography, or a document. During the 2nd TAILOR Summer School in June 2022 it was used to create groups of people with similar research interests (16 teams from 70 participants). For the 2nd TAILOR Conference in September 2022 we additionally matched each team's expertise with the most appropriate task in a group exercise (7 teams from 52 participants). Figure 1 shows the two steps in which SubSift was used during the conference. In both occasions, a large majority of the participants felt that their teams shared similar interests.

SynTeam is a tool based on the Post-Jungian Personality theory for team composition that generates synergistic teams based on four personality traits [3]. The resulting groups have been shown to work better together on solving specific tasks [2-9]. We collected the answers to 20 multiple choice questions from the participants and used SynTeam to divide them into teams.

In both events, the overall experience was positive with 12 Summer School students (out of 23 that provided written feedback) considering future collaborations with other students in their teams. Furthermore, after the 2nd TAILOR Conference ~96% of the respondents indicated that they expanded their professional network.

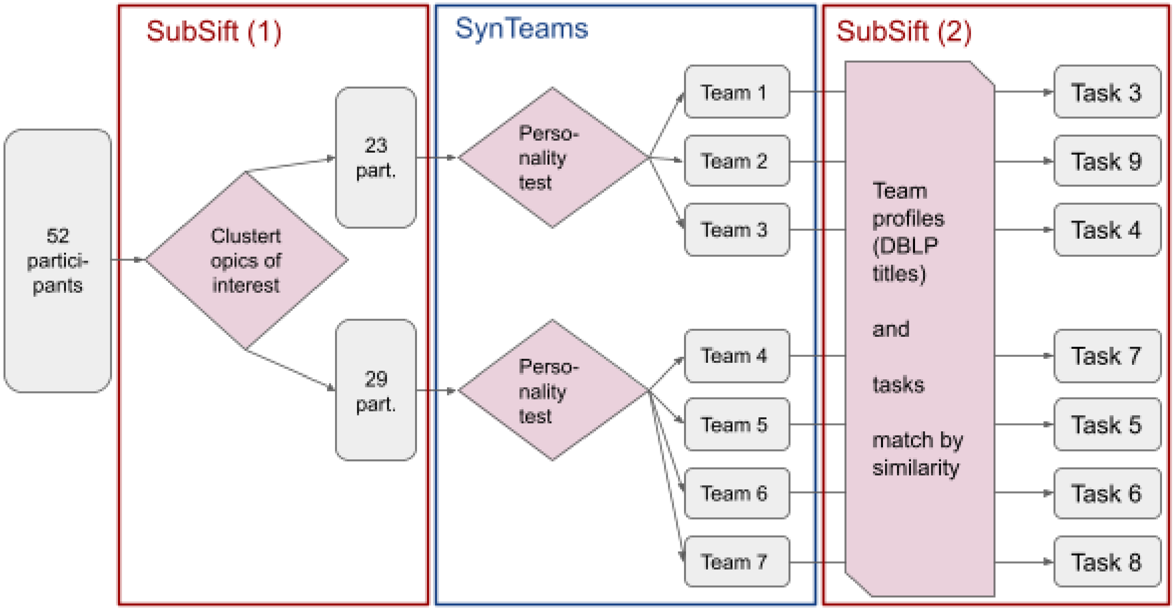


Fig. 1: Diagram of the pipeline used during the 2nd TAILOR Conference which shows how SubSift (red rectangles) was used in two steps: (1) to match participants of similar interests, and (2) to allocate the teams to the task descriptions.

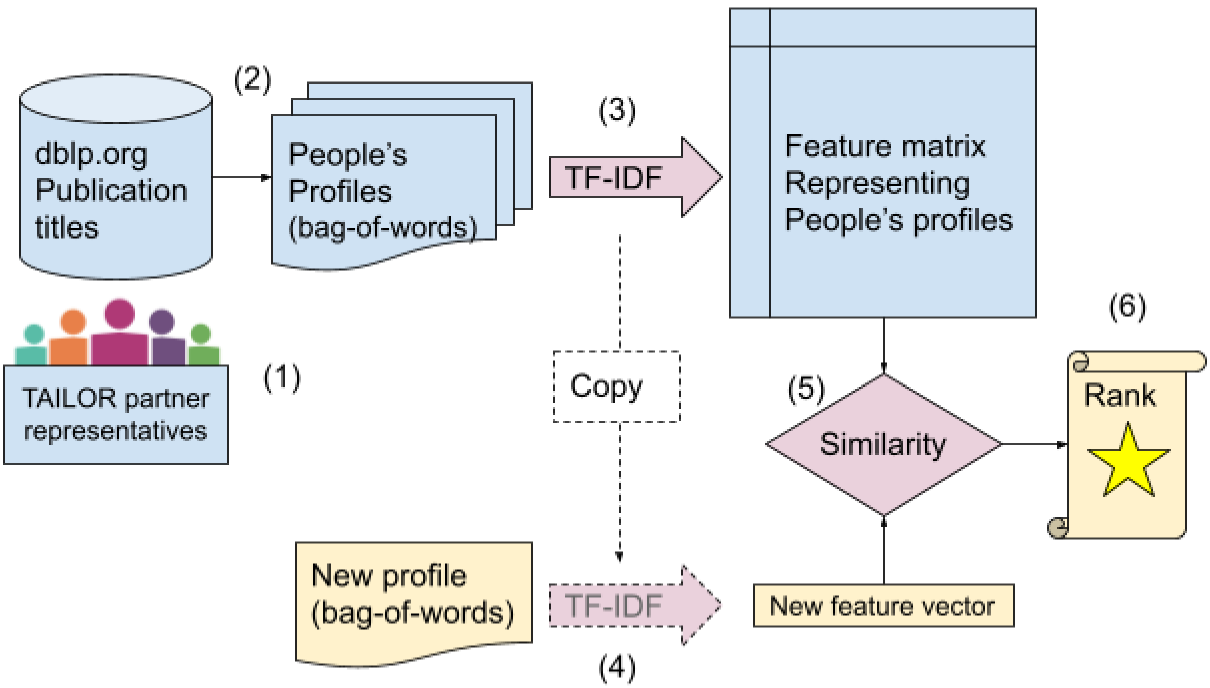


Fig. 2: Diagram of the pipeline used in the demonstrator. See text for details.

Methodology

In this section we describe the full process to create the demonstrator, from the selection of the individuals to the answers provided after a user has introduced a query. It contains a description of the following points: **(1)** Selection of people, **(2)** creation of their profiles, **(3)** creation of a feature representation that best discriminates their profiles, **(4)** transforming a new profile into the learned representation, **(5)** calculating the best matches between the provided profile and the people, **(6)** presentation of the answer, explainability and transparency.

Selection of people. With the objective of facilitating the collaboration within the TAILOR network of excellence, we decided to create a demo involving one representative of each partner as stated on the TAILOR website. This could be extended in future, but it should be noted that the process of creating profiles is not currently fully automated as the translation from names to DBLP IDs is not standardised.

Creation of people's profiles. SubSift requires a machine-understandable feature representation of each profile. One such representation is the term frequency-inverse document frequency (TF-IDF) that can be learned from profiles represented as a bag of words. Given that most of the representatives of the TAILOR partners have a list of previous publications, we collected their titles from the computer science bibliography website dblp.org. The generated profiles were manually inspected for possible errors introduced by the automatic matching.

Creation of a feature representation that best discriminates people's profiles. With a bag of words per person we can calculate the TF-IDF with a set of hyperparameters that best fits the discrimination of profiles. This includes the removal of English stop words (common words that are non-informative about a person's profile like the article "the"), and including n-grams for n up to 3, among others. Once the terms have been generated their TF-IDF is calculated and only the most informative terms are kept with their corresponding weight (importance). This step also allows the generation of a machine-understandable feature representation of each person's profile.

Transforming a new profile into the learned representation. Once the TF-IDF has been computed, it is possible to extract the bag of words of an arbitrary text and transform it into the learned representation. This enables the comparison between any of the profiles using different measurements.

Calculating the best match between the new profile and the people. It is possible to use several functions between the feature representations to obtain a similarity score. This demonstrator uses cosine similarity as it considers the direction of the feature representations while ignoring their magnitude. This is commonly used as a similarity measure for text analysis. We then compute the similarity of the provided new profile against all the precomputed profiles and sort them by score being 1 the most similar and -1 the most dissimilar.

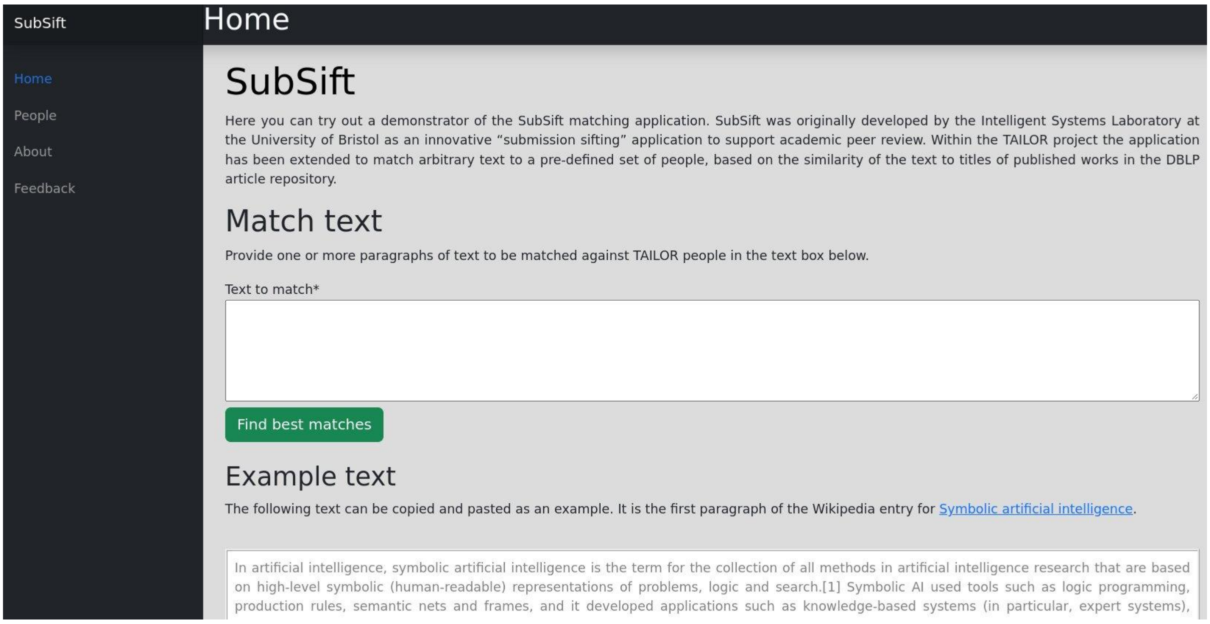
Presentation of the answer, explainability and transparency. Once the matching scores are obtained, the demonstrator sorts the profiles that yielded a positive similarity score in decreasing order. Each score is shown next to the name of the person, and the seven top terms that contributed to the score with their respective percentage contribution (among the seven terms).

Further details

The demonstrator can be accessed at the following links:
- subsift-webapp.herokuapp.com/subsift/ (this link will only work until November 28 when Heroku's free tier is no longer available);
- subsift-webapp.onrender.com/subsift/ (this link will remain available but may take up to one minute to launch initially).

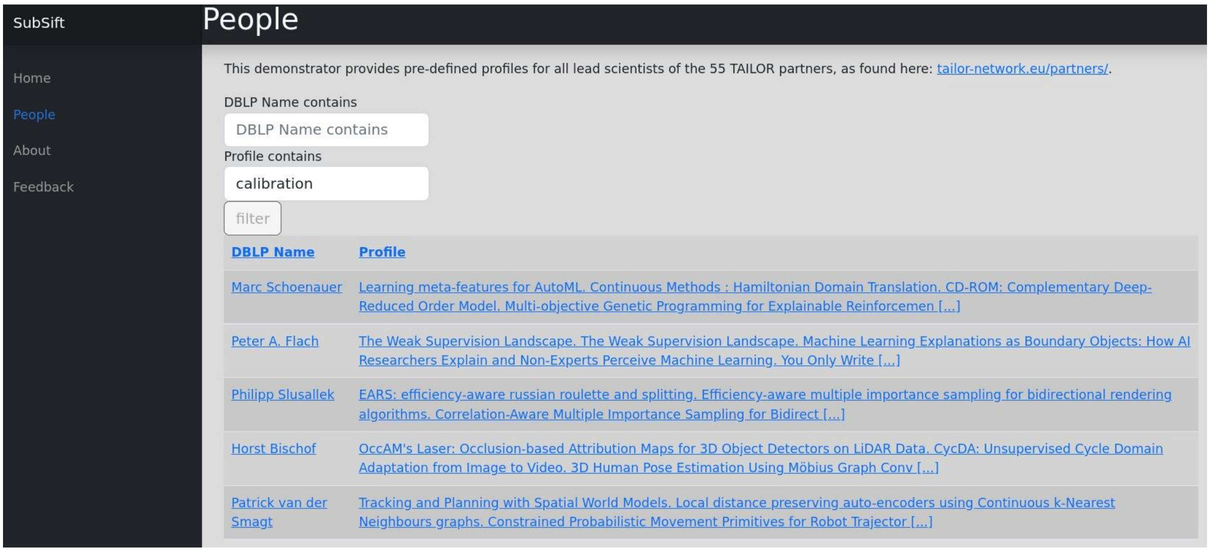
Using the demonstrator

Using the demonstrator is self-explanatory as shown in the screenshots below. The front page includes a short description of the demonstrator and provides a text field that can be filled with arbitrary text to be matched with the pre-populated representatives of the TAILOR partners. It also includes a text example that can be copied and pasted in the text field.



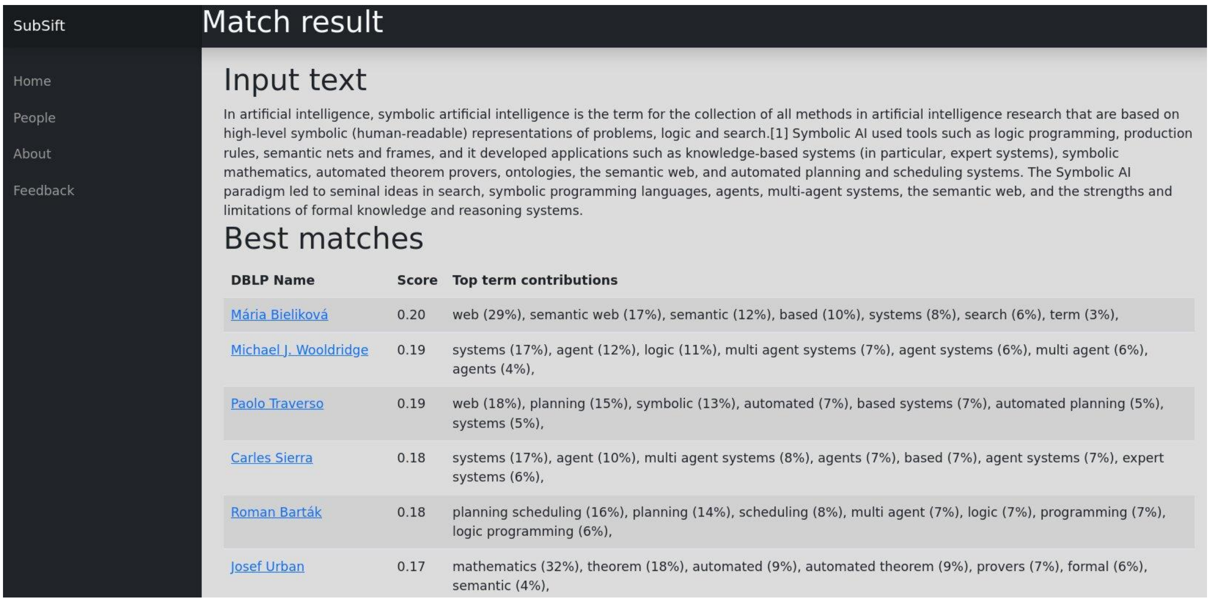
The list of people to match against is currently fixed as the lead scientists with DBLP profiles of the 54 TAILOR network partners. This allows us to pre-populate their profiles in this demo. In a future version we could expand the set of researchers either on demand or in batch; this would require (i) provision of their DBLP IDs, and (ii) re-calculation of the dictionary and the profiles.

Clicking on a person's name takes you to their DBLP page; clicking on the abbreviated profile allows you to see all publication titles as harvested from DBLP. These titles are combined in a single bag of words to match against the text provided by the user.



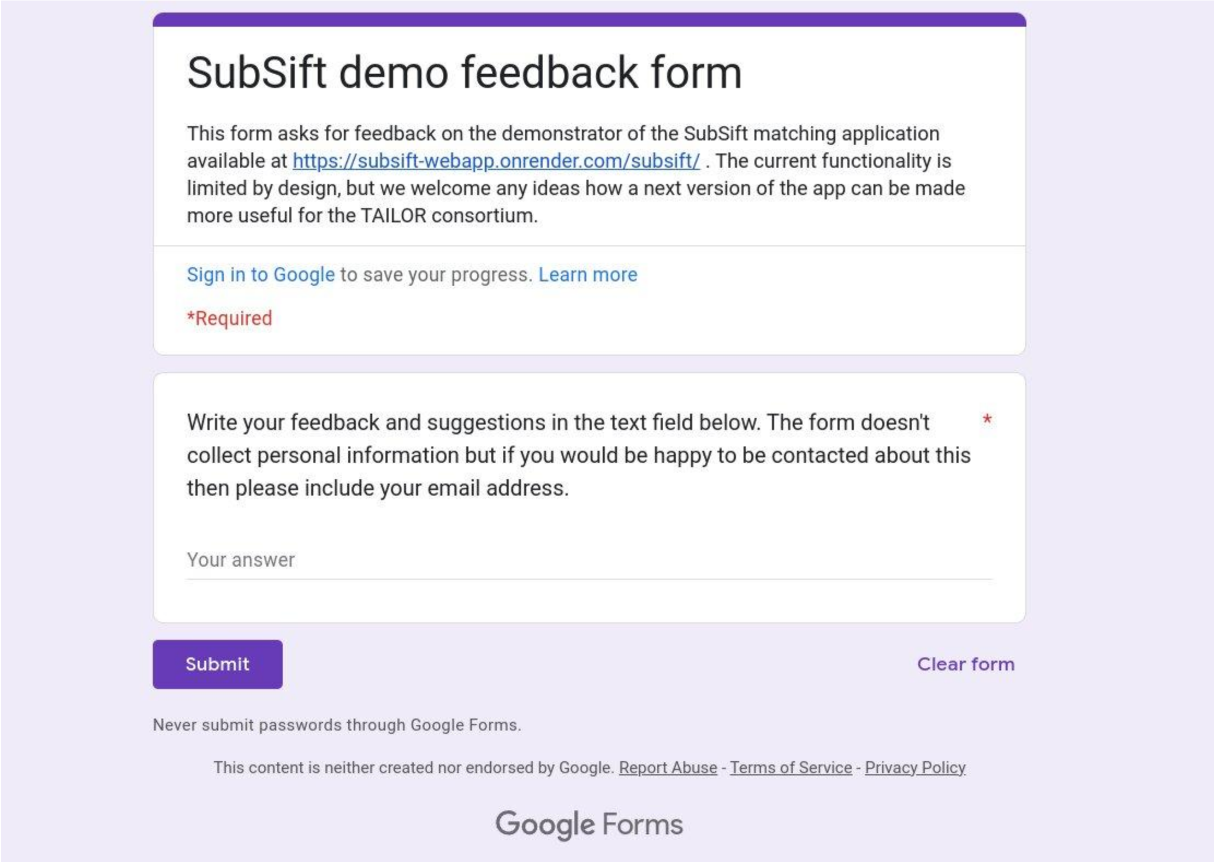
Example usage

Below we show an example of the top matches when using the first paragraph of https://en.wikipedia.org/wiki/Symbolic_artificial_intelligence as input text.
The results are ranked on decreasing match score, and the top contributing terms to the score are also indicated.



Providing feedback

The feedback link points to a simple Google form for collecting suggestions:


SubSift demo feedback form

This form asks for feedback on the demonstrator of the SubSift matching application available at https://subsift-webapp.onrender.com/subsift/ . The current functionality is limited by design, but we welcome any ideas how a next version of the app can be made more useful for the TAILOR consortium.

Sign in to Google to save your progress. Learn more

*Required

Write your feedback and suggestions in the text field below. The form doesn't collect personal information but if you would be happy to be contacted about this then please include your email address. *

Your answer

Submit

Clear form

Never submit passwords through Google Forms.

This content is neither created nor endorsed by Google. Report Abuse - Terms of Service - Privacy Policy

Google Forms

Lessons learnt

During the two use cases we were able to obtain feedback from the participants, as well as our own experience utilising the tool in real scenarios. Similarly, during the development of the demonstrator we have found some situations that may require further work in future iterations. SubSift requires a set of profiles in the form of bag-of-words in order to do the matching. For researchers that have a list of publications it is possible to obtain such bag-of-words in a semi-automatic way from the titles of their publications. On the other hand, some participants (e.g. during the Summer School) did not have previous publications. This made it necessary to require a free text per participant indicating their topics of interest. We adopted this approach for the 2nd TAILOR Conference as well to ensure that everybody could be matched. However, the provided free text contains usually less than 20 words, which makes the characterisation of each participant difficult for the matching (See Figure 3). Similarly, the task descriptions to match teams with people would also benefit from longer texts.

Another consideration is the fact that the automatic acquisition of a profile based on a name can be problematic, as people may share the same full name in databases like dblp.org which needs to be disambiguated. Or some people may use a full name and a short version in different contexts which makes it difficult to automatically match. For this demonstrator we decided to manually check all profiles to ensure their correctness.

The overall impressions received from the feedback provided by all the participants has been positive. We got multiple people encouraging us to improve and provide the same experiences in future events, or adapting the tool for their own use cases. Some of those improvements will be part of the later deliverable D9.2.

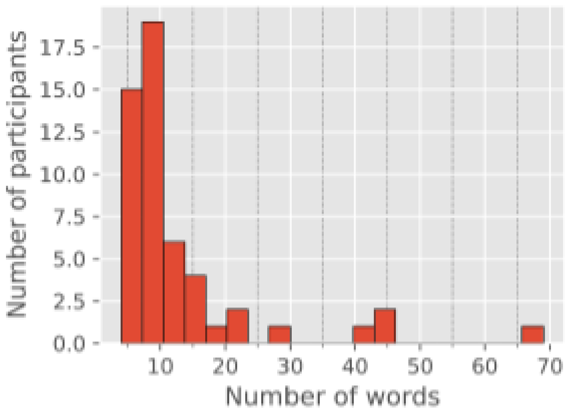


Fig. 3: Number of words provided as a free text for topics of interest.

Next steps

The positive feedback provided by the summer school and conference indicates that the proposed idea has potential for future events. The size of conferences and their number of attendees has been growing exponentially. This makes it difficult to identify potential researchers with similar interests and start a collaboration. Having a matching tool when attending an event with hundreds of participants may be a great asset to the event. We envision that the next TAILOR events can be an example to follow by other big conferences. We plan to continue working on additional use cases, and will release a new version as part of deliverable D9.2 towards the end of the TAILOR project. Some potential features are the creation of user accounts, the manual introduction of people and profiles by the final user, being able to save the match results, and better semi-automation of the profile acquisition.

References

[1] Simon Price, Peter A Flach, and Sebastian Spiegler. SubSift: a novel application of the vector space model to support the academic research process. In Tom Diethe, Nello Cristianini, and John Shawe-Taylor, editors, Proceedings of the First Workshop on Applications of Pattern Analysis, volume 11 of Proceedings of Machine Learning Research, pages 20–27, Cumberland Lodge, Windsor, UK, 2010. PMLR.

[2] Ewa Andrejczuk, Juan A Rodríguez-Aguilar, Carles Sierra, Carme Roig, and Yolanda Parejo-Romero. Don’t Leave Anyone Behind: Achieving Team Performance Through Diversity. In 2018 IEEE Frontiers in Education Conference (FIE), pages 1–9, 2018.
[3] John H Bradley and Frederic J Hebert. The effect of personality type on team performance. Journal of Management Development, 16(5):337–353, jan 1997.
[4] J. Arnold and R. Randall. Work psychology. Pearson Education Limited., Harlow, England, 2010.
[5] Larry K Michaelsen, Dean X Parmelee, Kathryn K McMahon, Ruth E Levine, and Diane M Billings. Team-based learning for health profes- sions education: A guide to using small groups for improving learning. 2008.
[6] M. K. Mount, M. R. Barrick, and G. L. Stewart. Five-factor model of personality and performance in jobs involving interpersonal interactions. Human Performance, (11):145–65, 1998.
[7] M. A. West. Effective Teamwork: Practical Lessons Learned from Organizational Research. Wiley-Blackwell, West Sussex, 2012.
[8] K. B. White. Mis project teams: An investigation of cognitive style implications. MIS Quarterly, (8(2)):95–101, 1984.
[9] D.J. Wilde. Post-Jungian Personality Theory for Individuals and Teams. SYDROSE LP, 2013.

Appendix: New Ways of Working: Tools for 21C Research

In preparation for this deliverable, WP9 partners organised an online workshop in September 2021 to discuss the tools and workflows used by TAILOR member researchers to streamline their research output. Further details of the workshop can be found online[1]. The following tools were presented at the workshop:

- **SWISH for SWI Prolog and CPLINT (Fabrizio Riguizzi, University of Ferrara).** SWISH [10] is a web front-end that integrates SWI Prolog on the server side and JavaScript code on the client side to be able to run Prolog code on the web. SWI Prolog is a web server that makes the execution of Prolog code possible. This allows the creation of interactive learning material of logic programming or other artificial intelligence concepts.
- **Creating Taxonomies and Ontologies with SONNET (Romy van Drie, P39-TNO).** Developing ontologies is an expensive and intensive task. The goal of SONNET (Semantic Ontology Engineering Toolset) [11-13] is to assist in developing taxonomies or ontologies. SONNET contains algorithms to kickstart the ontology development using a set of relevant documents. The talk showcased an algorithm to generate key concepts, as well as one to generate a first version of an ontology.
- **Managing Machine Learning Experiments with OpenML (Joaquin Vanschoren, P12-TUE).** OpenML [14] is an open-source initiative to create a global platform for sharing machine learning datasets, models, and reproducible experiments in a frictionless way. All data is open and accessible through APIs, and it is readily integrated into popular machine learning tools to allow easy sharing of models and experiments. This openness also allows a budding ecosystem of automated processes to scale up machine learning further.
- **You Only Write Thrice [15]: Adapting and Extending the Jupyter Book Ecosystem to Bespoke Presentation Needs (Kacper Sokol, P16-UNIVBRIS).** This work concerns various customisations and extensions of the Jupyter Book ecosystem that allows the creation of bespoke online learning materials. In particular, UNIVBRIS built plugins to embed non-standard interactive code boxes (SWI Prolog, CPLINT, ProbLog), and linked exercise–solution blocks. UNIVBRIS also explored the reveal.js open source library to generate slides from Markdown sources intended as online articles, and experimented with the RISE Jupyter Notebook plugin that enables creation of interactive slide shows. All of these tools come together to create a suite of engaging online learning materials.

[1] https://tailor-uob.github.io/deliverables/events/2021-09-new-ways-of-working

Appendix: References

[10] Jan Wielemaker, Fabrizio Riguzzi, Bob Kowalski, Torbjörn Lager, Fariba Sadri and Miguel Calejo, "Using SWISH to realise interactive web based tutorials for logic based languages", arXiv:1808.08042 [cs], Jan. 2020.
[11] Joaquin Vanschoren, Jan N. van Rijn, Bernd Bischl, and Luis Torgo. 2014. OpenML: networked science in machine learning. SIGKDD Explor. Newsl. 15, 2 (December 2013), 49–60. https://doi.org/10.1145/2641190.2641198
[12] M.H.T. de Boer and J.P.C. Verhoosel (2019), "Creating Data-Driven Ontologies - An Agriculture Use Case", Proc. ALLDATA 2019: The Fifth International Conference on Big Data, Small Data, Linked Data and Open Data, ISBN: 978-1-61208-700-9, pp. 52 - 57.
[13] M.H.T. De Boer and J.P.C. Verhoosel (2019), "Creating and Evaluating Data-Driven Ontologies", International Journal on Advances in Software, Q4 2019, pp 300-306.
[14] M.H.T. De Boer and J.P.C. Verhoosel (2020), "Towards Data-driven Ontologies: a Filtering Approach using Keywords and Natural Language Constructs", Language Resources and Evaluation Conference (LREC), Marseille, May 2020.
[15] Kacper Sokol and Peter Flach, "You Only Write Thrice: Creating Documents, Computational Notebooks and Presentations From a Single Source", arXiv:2107.06639 [cs], Jul. 2021.