A Hybrid Human-AI Tool for Scientometric Analysis

António Correia*[1,2], Andrea Grover[2], Shoaib Jameel[3], Daniel Schneider[4], Pedro Antunes[5], and Benjamim Fonseca[1]

[1] INESC TEC and University of Trás-os-Montes e Alto Douro, UTAD, Quinta de Prados, Apartado 1013, Vila Real, Portugal, [2] College of Information Science & Technology, University of Nebraska at Omaha, Omaha NE 68182, USA, [3] University of Southampton, SO17 1BJ Southampton, UK,

[4] Tércio Pacitti Institute of Computer Applications and Research (NCE), Federal University of Rio de Janeiro, Rio de Janeiro, Brazil, [5] LASIGE and University of Lisbon, Lisbon 1749-016, Portugal

[*] Corresponding author. Email address: antonio.g.correia@inesctec.pt

Abstract. Solid research depends on systematic, verifiable and repeatable scientometric analysis. However, scientometric analysis is difficult in the current research landscape characterized by the increasing number of publications per year, intersections between research domains, and the diversity of stakeholders involved in research projects. To address this problem, we propose SciCrowd, a hybrid human-AI mixed-initiative system, which supports the collaboration between Artificial Intelligence services and crowdsourcing services. This work discusses the design and evaluation of SciCrowd. The evaluation is focused on attitudes, concerns and intentions towards use. This study contributes a nuanced understanding of the interplay between algorithmic and human tasks in the process of conducting scientometric analysis.

Keywords: Artificial Intelligence, Bibliometric-enhanced Information Retrieval, Crowdsourcing, Human-AI Interaction, Reinforcement Learning from Human Feedback, Scientometrics.

1 Introduction

This paper applies Information Systems (IS) theories and methods to understand how hybrid human-Artificial Intelligence (AI) systems can assist researchers in doing scientometric analysis, i.e., measuring and analyzing scholarly literature (Garfield, 1979). Scientometric analysis is of utmost importance in the field of AI because it helps understand how scientific knowledge is generated, disseminated, and used across domains, paradigms, geographies, and times (De la Vega Hernández et al., 2022). In this regard, tools supporting scientometric analysis can help researchers better position and communicate about their studies. However, multidimensional scientometric analysis is difficult in the current research landscape (Ferrara and Salini, 2012). The remarkable growth in the number of publications per year (Fortunato et al., 2018), increasing intersections between research domains (e.g., technology and law, digital economy, biomedical engineering), and the diversity of stakeholders involved in research projects (e.g., universities, government, industry, non-profits, funding agencies, and citizen scientists), increase the complexity of scientometric analysis.

In the long run, the capacity of AI to automatically analyze large bibliometric data sets can help expedite literature reviews (Wang et al., 2022). However, without proper collaboration between researchers and AI, the outcomes may lack focus, accuracy, diversity, and adequate interpretation (Correia et al., 2020). In the pursuit of these and other ends, researchers have tried to find novel ways of amplifying literature-based discovery (Thilakaratne et al., 2019) through the use of AI techniques able to digest huge quantities of data while automating aspects of scientific activity with the ultimate goal of finding complex patterns and novel correlations and hypotheses (Waltz and Buchananm, 2009; Evans and Rzhetsky, 2010; Gil et al., 2014). For example, transformer-based large language models (LLMs) such as SciBERT (Beltagy et al., 2019) and GPT-3 (Floridi and Chiriatti, 2020) have gained a lot of traction in the last few years due to their modelling and predictive capabilities. However, they fail to capture contextual insights into the structure, dynamics, and implications of scientific activity since they have limited ability to reason, interpret, and contextualize research outputs. Moreover, LLMs also depend on the availability and quality of the underlying pre-trained data (Dhamala et al., 2021), which may lead to over-trust and overreliance on the model, incorrect answers (hallucination), and biased outputs.

While acknowledging the many constraints of AI-infused systems to cope with scientific complexity and the challenges that come with analyzing bibliographic data, interdisciplinary research efforts in the development of human-powered AI algorithms have been responsible for breaking down the boundaries of disciplines by accelerating the discovery of unrevealed relationships and observable properties that were not previously apparent within disciplinary silos (Hope et al., 2022). However, notwithstanding the rapid advances in AI-based functionalities for supporting theory building (Berente et al., 2019) and descriptive synthesis (Nakagawa et al., 2019; Schmiedel et al., 2019; Wagner et al., 2022), current algorithms involve expensive training and are usually prone to errors, failing to capture complex knowledge representations that may range from matrices of similarity between topics and other entities (Wang and Koopman, 2017) to the early detection of technology opportunities and threats based on overlapping words (Zhang and Yu, 2020). That is, the use of AI systems for examining patterns and correlating evidence from literature is still a long way from being a fully-fledged solution. Thus, human cognition plays a central role in the improvement of AI models when dealing with scientometric workflows.

In this study, we address these problems by integrating AI and crowdsourcing in the support of scientometric analysis through a Reinforcement Learning from Human Feedback (RLHF)-based model (Knox and Stone, 2009) intended to support bibliometric-enhanced Information Retrieval (IR) and large-scale knowledge base construction. For a survey of deep reinforcement approaches, the reader is referred to Du and Ding (2021). Building on the notion of using crowdsourcing in science as an information quality research frontier (Lukyanenko et al., 2020), along with the vast repertoire of empirical studies demonstrating that well-specified online tasks and narrowly focused projects do not typically require domain expertise to be competently performed (Rosser and Wiggins, 2019), the RLHF-based model proposed grasps elements of human-based decision-making by learning from crowd behavior with the ultimate goal of generating insights that are not immediately apparent to humans while keeping them in control by allowing to review the actions taken by the algorithm and providing feedback on its accuracy. By integrating research on IS and Human-Computer Interaction (HCI) for AI-based relational algorithms with a scientific basis, this work explores the use of crowdsourcing as a valid strategy to integrate inputs from a diverse pool of science contributors (Franzoni and Sauermann, 2014; Beck et al., 2022). While AI brings processing power and speed, crowdsourcing brings diverse and careful interpretations to the scientometric analysis process.

We propose a hybrid human-AI system, named SciCrowd, where AI-driven data discovery is complemented by crowdsourced interpretations. We suggest that *the integration between AI-driven data discovery and human-driven crowdsourced interpretations enables a more differentiated assessment of scientific production*. In particular, we focus on the collaborative development of semi-automated classification schemes, filters and dimensions of analysis. From a methodological standpoint, we develop SciCrowd using Design Science Research (DSR) (Hevner et al., 2004), which is a popular IS research paradigm focused on the construction of innovative IS artifacts. Following the DSR paradigm, we evaluate SciCrowd considering both feasibility (prototype construction) and utility (attitudes, concerns, and intention to use).

The remainder of the article is structured as follows. In Section 2, we overview some key issues and enabling technologies related to hybrid human-AI systems. In Section 3, we discuss DSR and position our study according to DSR. In Section 4, we address the problems and needs that exist in reality, while Section 5 describes the objectives for a solution that combines human-AI interaction for overcoming the challenges identified during the problem identification stage. In Section 6, we discuss the design and evaluation of SciCrowd. In Section 7, we provide a detailed discussion of the results obtained from a qualitative study with domain experts. We conclude the article by discussing a set of challenges and research implications for hybrid human-AI scientometric analysis.

2 Related Work

There has been a rapid increase of studies considering hybrid forms of interaction between humans and AI, where AI is often seen as a “teammate” (Bansal et al., 2019; Seeber et al., 2020; Zhang et al., 2021). Several challenges have been identified in this type of interaction, considering in particular the lack of control by humans (Yang et al., 2020), lack of transparency in AI agency (Liu, 2021; Vössing et al., 2022), and lack of trust in AI (Jorge et al., 2022). Nevertheless, researchers have also been studying how human-AI interactions can be of value to individuals, organizations and society in general. For instance, Schroder et al. (2022) found benefits related to the standardization of outputs, norms, skills, and knowledge (but no benefits related to mutual adjustment and supervision). An extensive body of literature highlights tensions around explainability, information overload, bias, and algorithmic fairness in human-AI settings (Ehsan et al., 2021). Increased creativity has also been associated with human-AI interaction (Karimi et al., 2020; Micchi et al., 2021), although some studies reported reductions in the users' sense of ownership over their artifacts when the AI system acts as a "co-creative partner" (Gero and Chilton, 2019; Biermann et al., 2022). As previously pointed out by Suh and colleagues (2021), AI can be used as a "helpful assistant" or a (third) "collaborator" with the ability to change the roles played by users during a collaborative activity by interactively offering suggestions to finalize incomplete work or co-creating novel content through prompt-based interfaces and generative mechanisms.

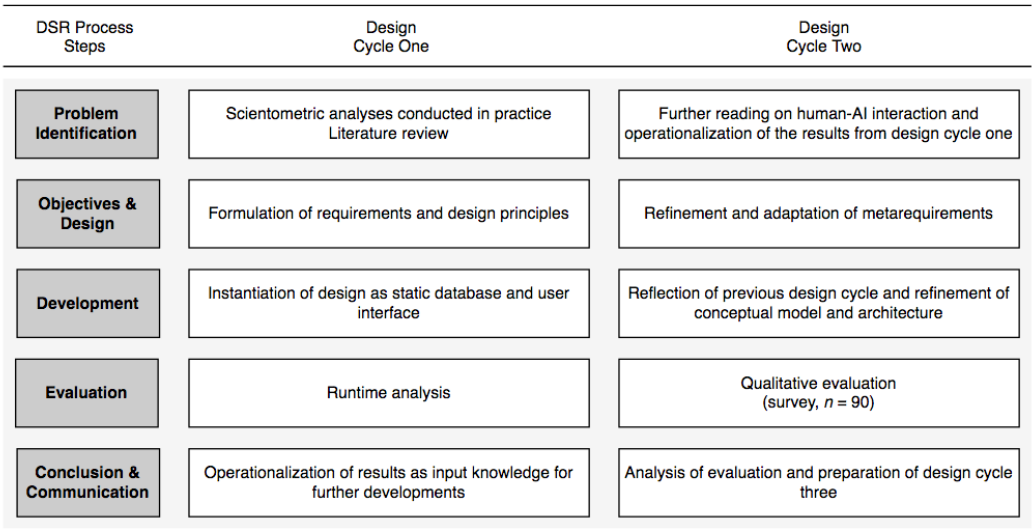
In regard to our study, researchers have proposed human-AI interaction in a variety of research undertakings. For instance, Johnson and associates (2022) note that most current approaches concern the early stages of the research process, including database searches, literature reviews, and summarizing papers. Antunes and co-authors (2022) develop semi-automated support for scoping reviews, which summarize the breadth of knowledge on a particular topic. Moreover, Jiang et al. (2021) consider a later step in the research process, where researchers synthesize their findings. Here, human-AI interaction can assist in developing deep insights from data. Accordingly, Feuston and Brubaker (2021) stressed the potential application of algorithmic approaches in data sampling and coding while augmenting (and increasing the scale of) qualitative data analysis by highlighting patterns and gaps from human analytic practices. As the authors noted, innovative tools are needed to overcome the limitations of overreliance on purely statistical indicators based on high-frequency distributions. As a brief example, quantifying the individual scientific output without making a distinction between self-citations and external citations can lead to misleading results and unfair decisions by hiring and tenure committees in contexts where research evaluation is purely based on performance indicators (Vincent-Lamarre and Larivière, 2023). Thus, designing trustworthy human-centered AI systems also capable of capturing infrequent occurrences of high value can improve the exploration experience by means of more contextual and qualitative aspects of data analysis.

To uncover the enduring aspects of the peer-review process, Price and Flach (2017) identify several parts that can benefit from human-AI interaction, including expert finding, assignment of reviewers, and scoring. The authors identify two human-AI interaction mechanisms, one that simplifies tasks (in particular human labor) through computational power (e.g., search engines), and another that augments the researcher's capabilities in the conduct of research (e.g., recommendation engines). The combination of AI and crowdsourcing can augment the researcher's capabilities in the conduct of research by leveraging the collective power of the crowd (Correia et al., 2020). At this level, several studies point out the potential benefits of crowdsourcing in research, e.g., supporting distributed knowledge discovery, data collection, and literature reviews (e.g., Krivosheev et al., 2018; Noel-Storr et al., 2021). However, few studies have investigated the combination of human-AI interaction and crowdsourcing using a RLHF-based model, as a novel approach to augment the researchers' capabilities in conducting scientometric analysis. This motivated us to develop SciCrowd.

3 Methodology

We adopt the DSR paradigm to develop SciCrowd. DSR is a popular IS research paradigm, which is centered on problem-solving, and the design and evaluation of innovative and useful IS artifacts (technological or not) (Hevner et al., 2004). This research paradigm is adequate to the current study because 1) the study involves problem-solving; 2) considering the complexity of the problem, the proposed solution is exploratory; and 3) the research is focused on artifact development, i.e., a software tool supporting researchers and other science stakeholders worldwide conducting scientometric analysis.

The DSR paradigm is applied to the study using the methodology proposed by Peffers et al. (2007), which suggests a logical organization of the research in the following steps (Figure 1): identify the problem and motivate, define objectives of a solution, design and develop an artifact, and demonstrate the suitability of the artifact and/or evaluate the artifact's utility.



DSR Process Steps	Design Cycle One	Design Cycle Two
Problem Identification	Scientometric analyses conducted in practice Literature review	Further reading on human-AI interaction and operationalization of the results from design cycle one
Objectives & Design	Formulation of requirements and design principles	Refinement and adaptation of metarequirements
Development	Instantiation of design as static database and user interface	Reflection of previous design cycle and refinement of conceptual model and architecture
Evaluation	Runtime analysis	Qualitative evaluation (survey, $n = 90$)
Conclusion & Communication	Operationalization of results as input knowledge for further developments	Analysis of evaluation and preparation of design cycle three

Fig. 1. SciCrowd development process

The design and development of SciCrowd are organized in two cycles (Figure 1). The first cycle considers initial development and demonstration of viability, while the second cycle considers refinement and validation. Next, we further discuss the problem identification, objectives of the solution, and artifact design and development.

4 Problem Identification

Every DSR process starts with a relevance cycle grounded in a problem identification stage (Hevner et al., 2004; Peffers et al., 2007). This research endeavor contributes to raising awareness of the problem to be investigated and justifies the need and value of a solution in terms of relevance. The main problem addressed by this research lies in how researchers can perform scientometric analysis in the current research landscape, which is characterized by the increasing number of publications, intersections between research domains, and diversity of stakeholders. This problem can be seen as a wicked problem, since it relates to a variety of other problems, including the way scientific knowledge is searched and identified, search process quality, variety of methods and methodologies, purposes of the researcher, quantitative and qualitative lenses, variety and quality of sources, etc. The solution to the identified problem has also an open, exploratory nature, as it combines the roles of individual researchers, support technologies, and the role of the crowd in supporting and collaborating with individual researchers. Finally, the method leading from problem to solution is a problem in itself, because the method needs to simultaneously focus on people and technology, and validation criteria can vary significantly. Thus, the wicked nature of the problem makes it an adequate candidate for applying design science in its resolution.

Similarly to other research projects described in the literature (e.g., Wiethof and Bittner, 2022), the problem formulation in this work is inspired by the gaps found in previous studies, including the difficulty of getting fine-grained and high-quality multidimensional data able to match particular information-seeking requirements when dealing with the specificity of scientometric analysis. At a higher level, the reader is invited to see Correia and co-authors (2020) for further comparison of the features of current scientometric tools and the needs that they cannot meet. In particular, obtaining broad-scale multidisciplinary views is critical when mapping the structure and evolution of a scientific domain (De la Vega Hernández et al., 2022), as many stakeholders increasingly need to make informed decisions and remain aware of the developments across fields and disciplines. In addition, a socio-technical solution is required to overcome bibliometric-enhanced IR problems such as information overload, data errors, ambiguity, and vocabulary mismatch (Ma et al., 2020). Consequent upon this, novel and alternative ways of computationally modelling scientometric data are needed to address the overarching problems outlined above.

5 Objectives of the Solution

Design science aims to refine theory by seeking answers to questions regarding the circumstances under which a solution or artifact may or may not be effective, including the potential reasons behind it (Engström et al., 2020). To account for this, design science requires that solutions should be challenging to achieve, should have an abstract nature, and should be positioned in relation to existing knowledge, both regarding existing and new knowledge contributions. The proposed solution exhibits these characteristics. The main challenge is to integrate AI-driven data discovery and crowdsourcing in support of scientometric analysis. The abstract aspect of the solution is the integration architecture, which structures several services, including classification, human verification, and user interface. Thus, the goal of this research is to build a solution that uses an interactive, mixed-initiative approach with humans-in-the-loop for scientometric data curation, fusion, and augmentation. Regarding knowledge contributions, the focus is on expository instantiation, i.e., a working prototype and its validation. The validation is conducted with domain experts to find out the extent of the level of feasibility of the solution by integrating the determinants perceived by potential users. In this way, we can obtain a better understanding of their intention to use the system taking into account its utility and effectiveness in addressing their particular expectations and needs.

6 Design and Development

One foundational aspect of DSR is that artifacts have a socio-technical nature (Rhode et al., 2009; Iivari, 2017), combining technical (tool) with informational (data model), and social perspectives (situated practices related to artifact use). Therefore, SciCrowd has to be designed taking into consideration the three perspectives. Next, we discuss these perspectives in detail.

6.1 SciCrowd: Technical Artifact

SciCrowd helps find patterns, relationships and associations that can be time-consuming and difficult to identify without tool support (e.g., Tchoua et al., 2017; Chan et al., 2018). In line with this, SciCrowd provides a set of services that help search and analyze large, heterogeneous volumes of bibliometric and altmetric data (Bornmann, 2014). It also integrates a set of crowdsourcing services that enable researchers to collaborate with volunteers (i.e., citizen scientists) and online paid workers on certain tasks, including additional searches, quality control (e.g., data correctness), and pattern finding. From the point of view of volunteers' engagement as citizen scientists, the reader is invited to see Correia and co-authors (2021) for a detailed description of the experimental work conducted in the evaluation of the usability of AuthCrowd, a submodule of SciCrowd intended to support author name disambiguation and entity matching. Here, a total of 24 volunteers took part in the experiment, and the overall results demonstrated accuracy values higher than 75% in the eight microtasks performed by volunteers. Moreover, the prototype's usefulness and ease of use were positively perceived by the majority of participants in the study.

On a technical level, the combination of AI and crowdsourcing services helps explore the breadth and depth of publication metadata, and identify and create new semantically meaningful linkages between bibliometric entities (e.g., theories, methods, studies, findings, and concepts). This is achieved through an option that allows users to classify these entities, while the system dynamically updates bibliometric information (e.g., citation indicators) from multiple sources. Rather than relying solely on the outcomes from either the human or the model, a crowd-AI combined approach is employed. In fact, this strategy has consistently demonstrated effectiveness in achieving higher levels of accuracy in decision-making (Singh et al., 2023). SciCrowd adopts a process architecture considering data extraction (with AI only), classification and verification (AI and crowdsourcing), storage, and visualization (Figure 2). Data extraction is performed by an automated crawler that runs an algorithm to extract metadata and metrics from selected databases using a simple XML query API. More information on our extraction framework is mentioned in Ley (2009). The extracted elements are then organized into an evolutionary taxonomy. The interactions between the crowd and AI can help co-create the algorithmic knowledge base and manipulate this evolutionary taxonomy in a hybrid fashion. Details about data visualization are discussed in relation to the social artifact and aligned with Shneiderman's guidelines on "overview first, zoom and filter, then details on demand". Additional details about SciCrowd, including source code, are available on GitHub[1].

[1] https://github.com/trrproject/SciCrowd

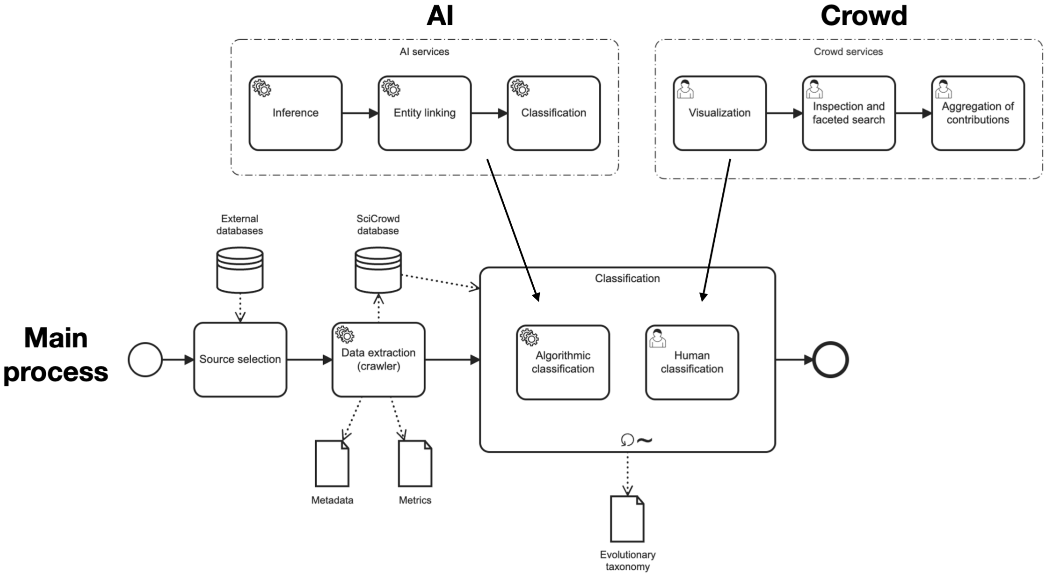


Fig. 2. SciCrowd architecture

In machine learning, a computational model is developed that is characterized by certain model parameters that are updated during the model training process. The same model also comprises a certain number of hyperparameters that control the overall training process, e.g., a simple hyperparameter is the number of iterations that the model must run. Examples of parameters include mean and variance, in the case of a simple Gaussian model. In our setting, the AI framework uses the evolutionary taxonomy, in combination with metadata and metrics, to train the model parameters. The adopted human-in-the-loop approach (Tokarchuk et al., 2012), which combines AI with the crowd, helps to reduce errors generated by the AI model. For instance, if the AI model makes a classification mistake, a crowd worker or volunteer can help correct that mistake. This correction can then be fed back to the AI model to improve future predictive performance. This feedback loop is popular in reinforcement learning (Kaelbling et al., 1996). Feedback also helps humans to understand the shortcomings of the AI model.

Relying on the crowd can also alleviate the need for a certain level of expertise when conducting scientometric analysis. Using the Swanson and Smalheiser's (1997) approach for informing the design of literature-based discovery support systems based on the lessons learned from internal and cross-domain studies conducted in different contexts (e.g., Correia et al., 2018; Correia and Lindley, 2022), a RLHF-based scientometric model interactively supports literature-based discovery processes in SciCrowd through a modular framework where human inspection plays a fundamental role in improving the general functioning of the system. To shed light on this point, an evaluation of the SciCrowd's system prototype in terms of performance is thoroughly described by Correia and associates (2019). The paper focuses on the system design and implementation, including a set of performance tests carried out with different servers to evaluate the robustness of the crawler.

6.2 SciCrowd: Information Artifact

SciCrowd adopts a uniform bibliometric data structure (Figure 3). All bibliometric elements support one-to-one and one-to-many relationships with other elements. For instance, a publication can be linked to one or multiple authors, institutional affiliations, countries, keywords, funding agencies, etc. That is, all the elements can be linked so that we can correlate metadata and specific properties of the taxonomic schema (e.g., concepts, sample characteristics) with bibliometric and altmetric indicators (e.g., citations, downloads, views, and social media mentions). As a result, SciCrowd provides fine-grained representations based on descriptive statistics resulting from complex relationships among several variables (Doré et al., 2000). By resorting to multidimensional analysis (Frame, 1984), SciCrowd allows users to explore a vast array of phenomena and perspectives such as highly-cited topics by country or ratio between self and external citations in the scientific production of a particular university.

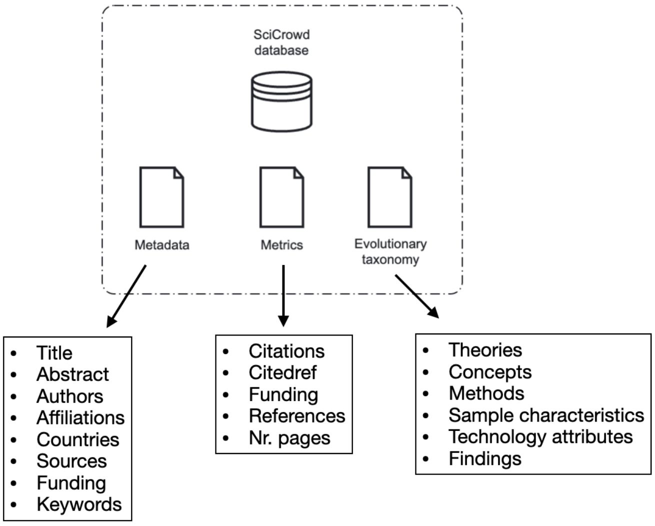


Fig. 3. SciCrowd database schema

The hybrid algorithmic-crowdsourcing approach supports processes like author name disambiguation (Sanyal et al., 2021) along with the enlargement of the database through iterative taxonomy generation and improvement (Chilton et al., 2013). In such a scenario, a collaborative system introduced in the literature can be classified according to the specific attributes of a classification scheme (e.g., Mittleman et al., 2008). This task is particularly difficult for a single human to perform at a massive scale since it involves qualitative assessment of large amounts of publications. By incorporating crowd inputs into the RLHF-based model, the system learns from these interactions in a way that allows improving the database in a co-evolving manner and thus facilitates a knowledge-guided interactive faceted search and visualization. Figure 4 illustrates the designed system framework and workflow based on an integrative view of the database schema and the role of the diversity of knowledge providers in the classification and overall improvement of data quality for scientometric analysis purposes.

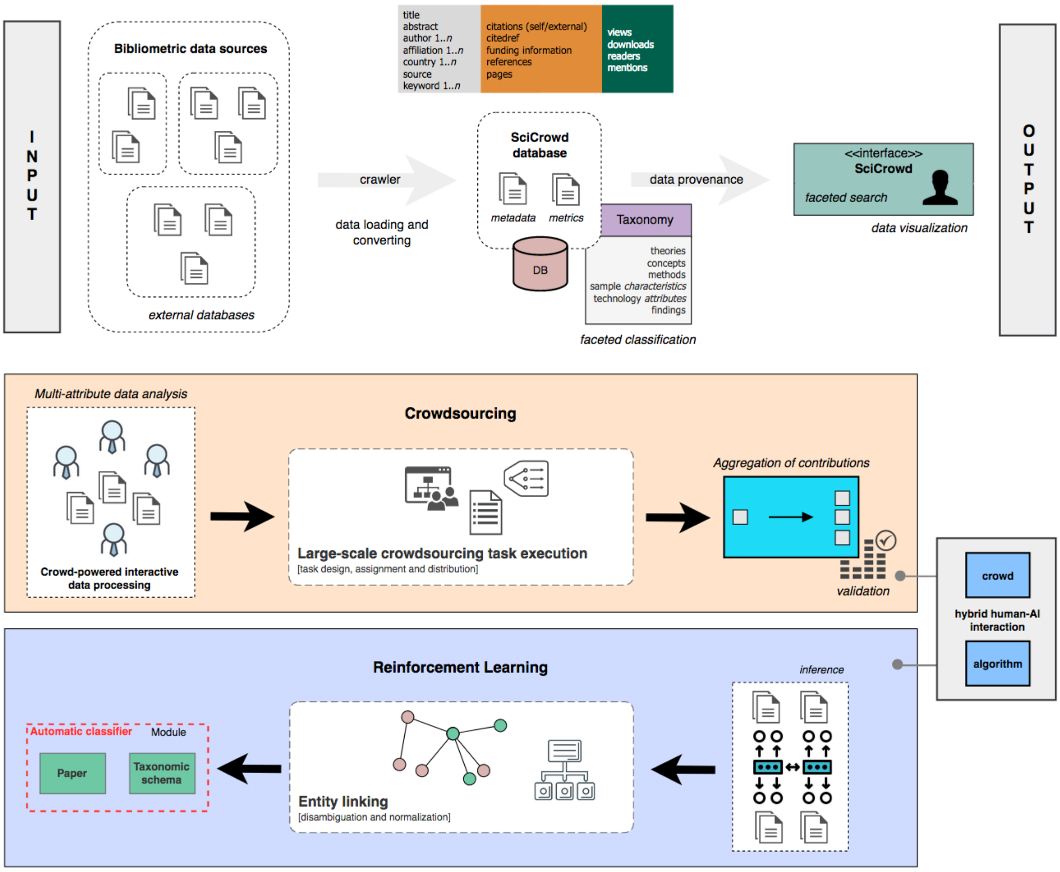


Fig. 4. Hybrid human-AI scientometric workflow that formally represents the system's capacity to co-evolve from crowd-algorithmic interactions

6.3 SciCrowd: Social Artifact

Understanding the social artifact is an essential aspect of SciCrowd design. In the second design cycle, we asked researchers in the field of crowdsourcing to suggest the most important features of a hybrid human-AI system for scientometric analysis. A brief overview of the questionnaire, findings and characteristics of the respondents is provided in Tables 2 and 3 in the Appendix. Table 3 also identifies a list of design requirements extracted from this exercise. Here, we detail two important requirements addressed by the social artifact: interactive and personalized faceted search, and crowd participation.

Interactive and personalized faceted search (Koren et al., 2008) was identified as particularly relevant for enhancing information seeking. Faceted search is a popular technique in Information Retrieval, which retrieves content on a particular topic or domain (Armentano et al., 2014). We developed a set of faceted search filters, which allow users to specify and refine attributes, categories and relationships between bibliometric elements in a dynamic fashion. The developed filters help users make questions like: “What are the most addressed <<**design principles**>> for implementing <<**emergency response systems**>> as mentioned in the <<**AI**>> literature?”; “How many <<**US health institutions**>> have introduced <<**human-in-the-loop machine learning algorithms**>> in their workflows from <<**2018 to 2022**>>?”; or “What <<**methods**>> have been commonly reported in empirical studies intended to examine the relationship between <<**medication non-adherence**>> and <<**health literacy**>> in <<**Asian**>> countries?”. Figure 5 presents two processes, related to the social artifact. The crowdsourcing process deals with recruitment and task assignment. The human classification process deals with the classification tasks done by the crowd. Figure 6 presents the implemented social artifact with faceted search filters.

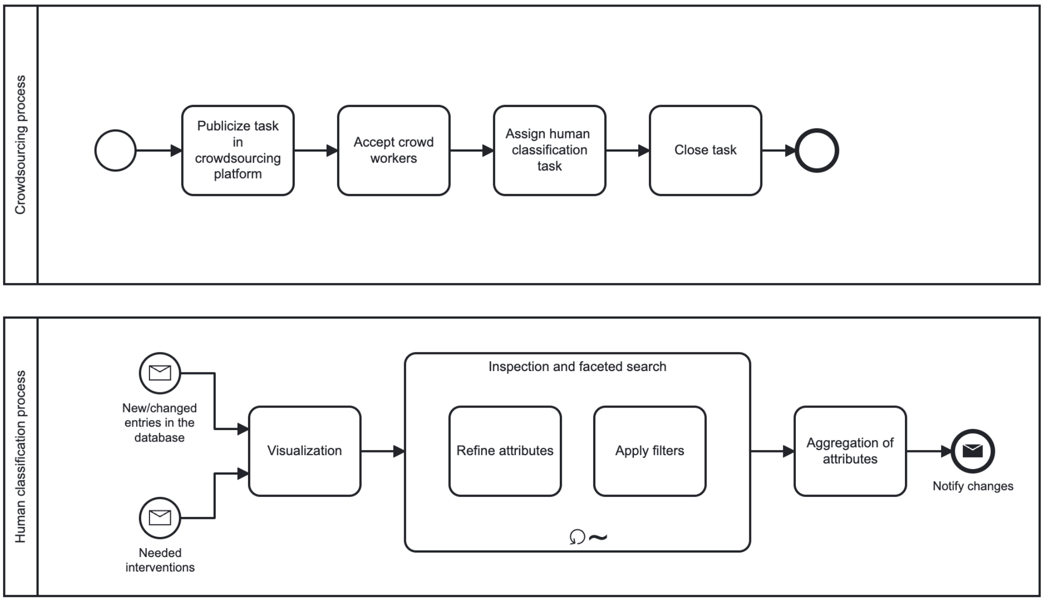
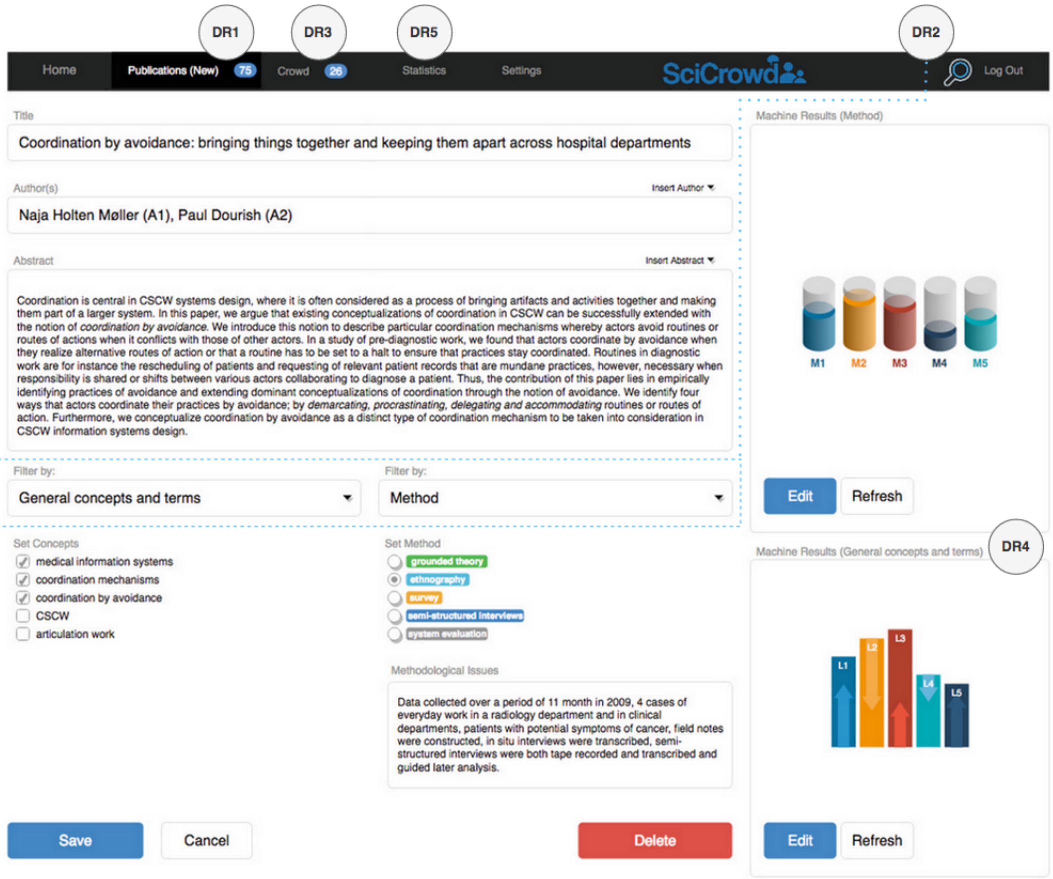

Crowdsourcing process
Publicize task in crowdsourcing platform
Accept crowd workers
Assign human classification task
Close task
Human classification process
New/changed entries in the database
Needed interventions
Visualization
Inspection and faceted search
Refine attributes
Apply filters
Aggregation of attributes
Notify changes

Fig. 5. Processes related to the social artifact



Title: Coordination by avoidance: bringing things together and keeping them apart across hospital departments

Author(s): Naja Holten Møller (A1), Paul Dourish (A2)

Abstract: Coordination is central in CSCW systems design, where it is often considered as a process of bringing artifacts and activities together and making them part of a larger system. In this paper, we argue that existing conceptualizations of coordination in CSCW can be successfully extended with the notion of *coordination by avoidance*. We introduce this notion to describe particular coordination mechanisms whereby actors avoid routines or routes of actions when it conflicts with those of other actors. In a study of pre-diagnostic work, we found that actors coordinate by avoidance when they realize alternative routes of action or that a routine has to be set to a halt to ensure that practices stay coordinated. Routines in diagnostic work are for instance the rescheduling of patients and requesting of relevant patient records that are mundane practices, however, necessary when responsibility is shared or shifts between various actors collaborating to diagnose a patient. Thus, the contribution of this paper lies in empirically identifying practices of avoidance and extending dominant conceptualizations of coordination through the notion of avoidance. We identify four ways that actors coordinate their practices by avoidance; by *demarcating, procrastinating, delegating and accommodating* routines or routes of action. Furthermore, we conceptualize coordination by avoidance as a distinct type of coordination mechanism to be taken into consideration in CSCW information systems design.

Filter by: General concepts and terms

Filter by: Method

Set Concepts
☑ medical information systems
☑ coordination mechanisms
☑ coordination by avoidance
□ CSCW
□ articulation work

Set Method
○ grounded theory
● ethnography
○ survey
○ semi-structured interviews
○ system evaluation

Methodological Issues: Data collected over a period of 11 month in 2009, 4 cases of everyday work in a radiology department and in clinical departments, patients with potential symptoms of cancer, field notes were constructed, in situ interviews were transcribed, semi-structured interviews were both tape recorded and transcribed and guided later analysis.

Fig. 6. SciCrowd web interface (cf. design requirements in Table 3 in the Appendix)

6.4 Evaluation

The main goal of DSR is to demonstrate the utility of designed artifacts. Addressing this concern, the SciCrowd evaluation is focused on identifying the strengths and weaknesses of SciCrowd's support for scientometric analysis. The evaluation was based on the revised Unified Theory of Acceptance and Use of Technology (UTAUT) (Dwivedi et al., 2019). This model was selected because it models the artifact's utility by theorizing about the individuals' intention to use it. Furthermore, the model also helps identify enablers (facilitators) and blockers (inhibitors) behind the intention to use SciCrowd. Next, we briefly discuss the evaluation data collection, followed by the results, which are divided between intention to use, and enablers and blockers.

6.4.1 Data Collection

We adopted the online survey to collect evaluation data. The survey was piloted with a small number of participants and then reviewed and validated by two expert researchers. Several revisions were suggested by the experts, which focused on the overall structure, comprehensiveness and understandability of the survey. The final survey contained 19 questions[2].

We selected the crowdsourcing research community as the target population for the survey, as SciCrowd targets that community. A purposive sampling strategy was applied to recruit the participants. The participants were identified by carrying out searches in several outlets (SpringerLink, ACM Digital Library, and IEEE Xplore) using the term "crowdsourcing", restricted to publications from 2006 till today, as 2006 was the year the crowdsourcing concept was popularized (Howe, 2006). Information regarding corresponding authors was extracted and the authors were contacted by email with an invitation to participate in the survey. No incentives were provided, and all survey participants gave informed consent.

Ninety people with valid responses participated in the survey. 75.6% were male and 24.4% were female. 60% of participants had more than eight years of research experience. 47.8% of participants indicated they perform literature searches on a daily basis. Table 1 summarizes the demographic profile of the participants.

Table 1. Demographic profile of the survey participants (*n* = 90)

Feature	Category	Count	%
Gender	Male	68	75.6
	Female	22	24.4
Experience	1-2 years	3	3.3
	3-4 years	10	11.1
	5-6 years	12	13.3
	7-8 years	11	12.2
	More than 8 years	54	60
Usage frequency	Daily	43	47.8
	Weekly	38	42.2
	Monthly	9	10
Affiliation continent	Europe	34	37.8
	North America	32	35.6
	South America	5	5.6
	Africa	2	2.2
	Oceania	4	4.4
	Asia	13	14.4

We used qualitative coding to analyze the participants' responses (Corbin and Strauss, 1990) and coded the survey data using NVivo. We also used a directions matrix to analyze the directions of the responses (Karunagaran et al., 2019) as positive [+], negative [-] or neutral [|].

6.4.2 Results: Intention to Use

The results indicate a generally positive attitude (*n* = 41; 45.6%) towards using SciCrowd. This is exemplified in a quote from a participant: "I am really interested in a human-machine intelligence hybrid system for scientific purposes. It will be the future!" Another respondent elaborated that "crowd-computing can fill the gap that AI cannot address (yet)" by combining the best of both worlds.

[2] The questionnaire is available at: https://forms.gle/hQaPLMo1PDZWqCU46.

Moreover, a participant emphasized the innovative nature of SciCrowd in the following terms: “I have not found any tool that I feel makes it easy to find good research papers from the past one or two years, or to get an overview of what is currently happening in a particular field.” About one-third of participants were neutral about SciCrowd (*n* = 29; 32.2%), while no particular sentiment (positive, negative or neutral) was detected in the remaining respondents (*n* = 15; 16.7%).

Few respondents (*n* = 5; 5.6%) demonstrated clear skepticism towards SciCrowd. A participant wrote: “I did my Ph.D. on mixed crowd/machine learning (ML) systems, and I honestly don’t think this would be a good approach for making sense of large bodies of research. Making sense of research papers requires lots of expert knowledge. [...] What I think would be useful is an expert system that can help a single researcher or a small team of researchers to make sense of large collections of documents”. The same participant also noted, “[h]owever, I don’t see how this can be built today, [since] most research is behind paywalls.”

6.4.3 Results: Enablers and Blockers

We summarize the most relevant enablers and blockers identified by the participants. Table 3 in the Appendix provides a more detailed list of results.

Considering enablers, the contextualization of search results to the specific needs of the researcher was mentioned by multiple participants. As one informant said, “I tend to seek out and explore papers that help me to understand a scientific space or craft a specific argument. So, I don’t often engage with the overwhelming totality of a scientific field or discipline, but rather with fairly constrained subsets.” Automatic identification of semantic linkages between bibliometric items was also seen by participants as a key success factor. This was exemplified by an informant who called for “more connections (links) between the data, the published papers, the search (query) terms that led a user to the data, how the data were used and/or analyzed.” Finally, the capacity to reuse bibliometric data across multiple searches was also considered an enabler.

Regarding blockers, the participants identified difficulties avoiding data overload, issues related to data incompleteness and noise, poor matching results, poor similarity detection and duplicate identification, limited access to some databases, and lack of replicability. Another blocker concerns the time-consuming task of identifying which publications are most relevant from the search results. Finally, lack of transparency was also seen as potentially problematic. As noted by one participant, transparency can add significant value, if SciCrowd allows users “to keep track of all versions of the information and knowledge capture, so that the state of knowledge can be rewound and replayed from any point in time to any other point in time.”

7 Discussion

At its core, SciCrowd automatically takes bibliographic information from various sources and processes it combining AI and crowdsourcing services. The interactions between AI and crowdsourcing services help harness the best of both human and AI abilities to perform complex collaborative tasks. Scientometric analysis then emerges as a collaborative process founded on hybrid and collective intelligence (Lukyanenko et al., 2020; Blesik et al., 2021; Peeters et al., 2021). We now draw several insights from the design and evaluation of SciCrowd.

7.1 Collaboration

The participants in the survey pointed out that the ability to search scientific literature by combining automated crawling and faceted search is a positive aspect of SciCrowd. On the other hand, the participants also identified some potential inhibitors, in particular lack of contextualization in data analysis. As we pursue more contextualized analyses, crowdsourcing may bring the benefit of the crowd (Luz et al., 2015). Effective collaboration between systematic, large-scale, fast-responding AI-assisted search services, and diversified and attentive members of the crowd, can provide better support to scientometric analysis than single-user approaches. The formation of expert groups to oversee mundane bibliometric analyses can be problematic, if not unfeasible (e.g., Vinella et al., 2022). The integration with crowdsourcing platforms offered by SciCrowd represents a simple strategy for procuring experts who are willing to collaborate on scientometric analyses in a repeated way. At the same time, SciCrowd is an adaptive system that provides an environment in which non-paid contributors can implicitly interact with scientists and the machine as "teammates" to gain insight into the underlying dynamics of knowledge production by using stigmergy (Thomas and Zaytseva, 2016) and other self-organizing processes as a basis for scaling out scientometric activities.

Looking at the feedback provided by the participants in the evaluation, feedback generated by the crowd in the form of comments, questions, and expert insights, contributes to perceiving SciCrowd as a useful tool. In future work, SciCrowd tasks focused on analyses of specific domains could be used as a training and discovery tool for graduate students and subject specialist librarians. For instance, consider a real-world application where a computer science student is using SciCrowd as an assistance tool to narrow down her literature search. She starts by entering a few keywords related to her topic, such as "human-centered AI" and "mental health". As soon as the system processes her query, a set of publications matching her criteria can be suggested and filtered by intervention strategy, mental health effects, and clinical implications. The results are displayed with relevant metadata alongside each record, including publication dates, author names, and citation counts. As she explores the filtered results, she can view citation and author topic-related networks for the chosen dataset. This helps her to gauge the relevance and impact of the research. Throughout her interaction with the SciCrowd system, the scientometric assistance tool not only assists her in finding relevant literature but also enhances her understanding of the research landscape by providing comprehensive and filtered information. Instead of being the primary instrument for analysis, SciCrowd can assume various roles (such as "assistant", "facilitator", "mediator", and "partner") in the analytical process. This demonstrates how SciCrowd can bring attention to diverse perspectives in data analysis, akin to the tasks that a human collaborator might undertake.

7.2 Collective governance

One relevant issue arising from hybrid human-AI interaction concerns governance, i.e., who has control over the process. Governance rules are required to ensure the quality and integrity of the process. From a regulatory perspective, it is critical to establish rules regarding the interaction between humans and AI, the collaboration between researchers and crowd workers, and the overall quality of shared services (by humans and AI).

7.3 Participation

SciCrowd depends on the contributions of the crowd to deliver the intended benefits. As such, it seems critical to consider the willingness of the crowd to participate; this is also an ethical consideration, as it prioritizes the autonomy of participants. Incentivizing and sustaining scientometric analysis as a repeatable practice is essential, which is a concern of most citizen science and crowdsourcing initiatives (Jackson et al., 2015). Crowdsourcing tasks should be clear, focused, time-boxed, and easy to engage by the crowd (Gadiraju et al., 2015). On the other hand, tasks should also be stimulating, in terms of knowledge contributions. However, unlike many common crowdsourcing tasks (e.g., classifying pictures), which emphasize intensive labor, SciCrowd depends on knowledge workers to perform knowledge-intensive tasks. As noted by a participant in the evaluation, the crowd "must have a certain level of understanding otherwise somebody can't reach the right conclusions." This has ethical implications as well, since the crowd workers for SciCrowd leverage their professional expertise, which is usually compensated. However, the voluntary nature of participation typically mitigates concerns about exploitation.

One positive aspect of using the crowd for scientometric analysis is that the crowd can bring openness and inclusivity to the process. Some participants in the evaluation noted the importance of openness and inclusivity when dealing with research: "we should […] engage diverse voices but that is challenging, particularly if [research] designers are not representative of the diversity in education and income levels, race, gender, etc., themselves.” The crowd can also contribute “[…] by providing ontologies from where to select classes, descriptions, and metadata such as self-assessed confidence which is important to evaluate the trust of the data.” On the other hand, we should not neglect some negative aspects, such as extreme overfitting and crowd bias. The latter is assumed to occur in different ways in crowdsourcing, such as when workers allow themselves to be led by the crowd’s behavior instead of their own way of thinking (Eickhoff, 2018). To overcome these issues, a variety of quality assurance and quality control techniques are commonly applied (Daniel et al., 2018). In addition, crowdsourcing services should be extended with services (which could use either AI or the crowd) to oversee the crowd. A successful example of this approach is the CrowdScape system (Rzeszotarski and Kittur, 2012), which can infer negative behaviors from crowd workers.

7.4 Dynamic process

The increasing volume, flux, diversity, and multidisciplinarity of scientific publications often make it challenging to identify and integrate multiple viewpoints. SciCrowd combines AI and crowdsourcing services to support frequent updates and dynamic changes to scientometric analysis. The combination of automated searches with distributed crowd work helps reduce the costs of keeping literature reviews up to date, continuously checking for new and emerging trends, and serendipity and exploration.

8 Conclusion

Solid research depends on systematic, verifiable and repeatable scientometric analysis. However, we are still only beginning our journey to overcome the limitations of current approaches to scientometric analysis. This work addresses the collaboration between AI and humans in supporting scientometric analysis, where human support is crowdsourced.

In this study, we report the design and evaluation of SciCrowd, a tool supporting scientometric analysis through the combination of AI and crowdsourcing services. The study contributes design knowledge in the form of tool instantiation. Moreover, the tool evaluation contributes knowledge regarding the intention to use and perceived enablers and blockers.

The evaluation suggests that SciCrowd can be valuable for researchers to dynamically develop and continuously update their scientometric analyses with contributions from the crowd. Furthermore, SciCrowd offers a self-adapting system, which allows AI algorithms to adapt to human requests.

Nevertheless, some complexities remain unsolved, which are left for future work. These include human-AI interaction in relation to the quality control of scientometric data, and a more sophisticated human-AI workflow in regard to data analysis and visualization.

Acknowledgements

This research was mainly performed during an internship of António Correia at Microsoft Research, Cambridge, UK. The work was supported in part by the Portuguese Foundation for Science and Technology (FCT), national funding through the individual research grant SFRH/BD/136211/2018. The authors would like to thank Siân Lindley from Microsoft Research for the important role in understanding and modifying the human-AI scientometric workflow that supports the SciCrowd system, as well as Jorge Santos for the help while building the necessary infrastructure. Our thanks extend to Hugo Paredes for the helpful discussions and valuable insights in the early stages of this work.

Declarations

Conflict of Interest. The authors of this manuscript have no conflicts of interest or competing interests to declare.

References

Antunes, P., Johnstone, D., Hoang Thuan, N., & de Vreede, G. J. (2022). Delivering evidence-based management services: Rising to the challenge using design science. *Knowledge Management Research & Practice*, pp. 1–16.

Armentano, M. G., Godoy, D., Campo, M., & Amandi, A. (2014). NLP-based faceted search: Experience in the development of a science and technology search engine. *Expert Systems with Applications*, 41(6), pp. 2886–2896.

Bansal, G., Nushi, B., Kamar, E., Weld, D. S., Lasecki, W. S., & Horvitz, E. (2019). Updates in human-AI teams: Understanding and addressing the performance/compatibility tradeoff. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(1), pp. 2429–2437.

Beck, S., Brasseur, T. M., Poetz, M., & Sauermann, H. (2022). Crowdsourcing research questions in science. *Research Policy*, 51(4), 104491.

Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 3613–3618.

Berente, N., Seidel, S., & Safadi, H. (2019). Data-driven computationally intensive theory development. *Information Systems Research*, vol. 30, no. 1, pp. 50–64.

Biermann, O. C., Ma, N. F., & Yoon, D. (2022). From tool to companion: Storywriters want AI writers to respect their personal values and writing strategies. In *Proceedings of the Designing Interactive Systems Conference*, pp. 1209–1227.

Blesik, T., Bick, M., & Kummer, T. F. (2021). A conceptualisation of crowd knowledge. *Information Systems Frontiers*, pp. 1–19.

Bornmann, L. (2014). Do altmetrics point to the broader impact of research? An overview of benefits and disadvantages of altmetrics. *Journal of Informetrics*, 8(4), pp. 895–903.

Chan, J., Chang, J. C., Hope, T., Shahaf, D., & Kittur, A. (2018). SOLVENT: A mixed initiative system for finding analogies between research papers. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW), pp. 1–21.

Chilton, L. B., Little, G., Edge, D., Weld, D. S., & Landay, J. A. (2013). Cascade: Crowdsourcing taxonomy creation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1999–2008.

Corbin, J. M., & Strauss, A. (1990). Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative Sociology*, 13(1), pp. 3–21.

Correia, A., Paredes, H., & Fonseca, B. (2018). Scientometric analysis of scientific publications in CSCW. *Scientometrics*, 114(1), pp. 31–89.

Correia, A., Fonseca, B., Paredes, H., Schneider, D., & Jameel, S. (2019). Development of a crowd-powered system architecture for knowledge discovery in scientific domains. In *Proceedings of the 2019 IEEE International Conference on Systems, Man, and Cybernetics*, pp. 1372–1377.

Correia, A., Jameel, S., Schneider, D., Paredes, H., & Fonseca, B. (2020). A workflow-based methodological framework for hybrid human-AI enabled scientometrics. In *Proceedings of the 2020 IEEE International Conference on Big Data*, pp. 2876–2883.

Correia, A., Guimarães, D., Paulino, D., Jameel, S., Schneider, D., Fonseca, B., & Paredes, H. (2021). AuthCrowd: Author name disambiguation and entity matching using crowdsourcing. *In Proceedings of the IEEE 24th International Conference on Computer Supported Cooperative Work in Design*, pp. 150–155.

Correia, A., & Lindley, S. (2022). Collaboration in relation to human-AI systems: Status, trends, and impact. In *Proceedings of the 2022 IEEE International Conference on Big Data*, pp. 3417–3422.

Daniel, F., Kucherbaev, P., Cappiello, C., Benatallah, B., & Allahbakhsh, M. (2018). Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. *ACM Computing Surveys*, 51(1), pp. 1–40.

De la Vega Hernández, I. M., Urdaneta, A. S., & Carayannis, E. (2023). Global bibliometric mapping of the frontier of knowledge in the field of artificial intelligence for the period 1990–2019. *Artificial Intelligence Review*, vol. 56, no. 2, pp. 1699–1729.

Dhamala, J., Sun, T., Kumar, V., Krishna, S., Pruksachatkun, Y., Chang, K. W., & Gupta, R. (2021). Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 862–872.

Doré, J. C., Dutheuil, C., & Miquel, J. F. (2000). Multidimensional analysis of trends in patent activity. *Scientometrics*, 47(3), pp. 475–492.

Du, W., & Ding, S. (2021). A survey on multi-agent deep reinforcement learning: From the perspective of challenges and applications. *Artificial Intelligence Review*, vol. 54, no. 5, pp. 3215–3238.

Dwivedi, Y. K., Rana, N. P., Jeyaraj, A., Clement, M., & Williams, M. D. (2019). Re-examining the unified theory of acceptance and use of technology (UTAUT): Towards a revised theoretical model. *Information Systems Frontiers*, 21(3), pp. 719–734.

Ehsan, U., Liao, Q. V., Muller, M., Riedl, M. O., & Weisz, J. D. (2021). Expanding explainability: Towards social transparency in AI systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–19.

Eickhoff, C. (2018). Cognitive biases in crowdsourcing. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pp. 162–170.

Engström, E., Storey, M. A., Runeson, P., Höst, M., & Baldassarre, M. T. (2020). How software engineering research aligns with design science: A review. *Empirical Software Engineering*, 25, pp. 2630–2660.

Evans, J. A., & Rzhetsky, A. (2010). Machine science. *Science*, vol. 329, no. 5990, pp. 399–400.

Ferrara, A., & Salini, S. (2012). Ten challenges in modeling bibliographic data for bibliometric analysis. *Scientometrics*, vol. 93, no. 3, pp. 765–785.

Feuston, J. L., & Brubaker, J. R. (2021). Putting tools in their place: The role of time and perspective in human-AI collaboration for qualitative analysis. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2), pp. 1–25.

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, vol. 30, pp. 681–694.

Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., Petersen, A. M., Radicchi, F., Sinatra, R., Uzzi, B., Vespignani, A., Waltman, L., Wang, D., & Barabási, A.-L. (2018). Science of science. *Science*, 359(6379), eaao0185.

Frame, J. D. (1984). Multidimensionality is alive and well in applied statistics. *Scientometrics*, vol. 6, no. 2, pp. 97–101.

Franzoni, C., & Sauermann, H. (2014). Crowd science: The organization of scientific research in open collaborative projects. *Research Policy*, 43(1), pp. 1–20.

Gadiraju, U., Demartini, G., Kawase, R., & Dietze, S. (2015). Human beyond the machine: Challenges and opportunities of microtask crowdsourcing. *IEEE Intelligent Systems*, 30(4), pp. 81–85.

Garfield, E. (1979). Scientometrics comes to age. *Current Contents*, vol. 46, pp. 5–10.

Gero, K. I., & Chilton, L. B. (2019). Metaphoria: An algorithmic companion for metaphor creation. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pp. 1–12.

Gil, Y., Greaves, M., Hendler, J., & Hirsh, H. (2014). Amplify scientific discovery with artificial intelligence. *Science*, vol. 346, no. 6206, pp. 171–172.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, pp. 75–105.

Hope, T., Downey, D., Etzioni, O., Weld, D. S., & Horvitz, E. (2022). A computational inflection for scientific discovery. *arXiv preprint arXiv:2205.02007*.

Howe, J. (2006). The rise of crowdsourcing. *Wired Magazine*, 14(6), pp. 1–4.

Iivari, J. (2017). Information system artefact or information system application: That is the question. *Information Systems Journal*, 27(6), pp. 753–774.

Jackson, C. B., Østerlund, C., Mugar, G., Hassman, K. D., & Crowston, K. (2015). Motivations for sustained participation in crowdsourcing: Case studies of citizen science on the role of talk. In *Proceedings of the 48th Hawaii International Conference on System Sciences*, pp. 1624–1634.

Jiang, J. A., Wade, K., Fiesler, C., & Brubaker, J. R. (2021). Supporting serendipity: Opportunities and challenges for human-AI collaboration in qualitative analysis. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1), pp. 1–23.

Johnsson, M., Gustafsson, C., & Johansson, P. E. (2022). *Disrupting the research process through artificial intelligence: Towards a research agenda*. Artificial Intelligence and Innovation Management, pp. 161–183.

Jorge, C. C., Tielman, M. L., & Jonker, C. M. (2022). Artificial trust as a tool in human-AI teams. In *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 1155–1157.

Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 4, pp. 237–285.

Karimi, P., Rezwana, J., Siddiqui, S., Maher, M. L., & Dehbozorgi, N. (2020). Creative sketching partner: An analysis of human-AI co-creativity. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pp. 221–230.

Karunagaran, S., Mathew, S. K., & Lehner, F. (2019). Differential cloud adoption: A comparative case study of large enterprises and SMEs in Germany. *Information Systems Frontiers*, 21(4), pp. 861–875.

Knox, W. B., & Stone, P. (2009). Interactively shaping agents via human reinforcement: The TAMER framework. In *Proceedings of the 5th International Conference on Knowledge Capture*, pp. 9–16.

Koren, J., Zhang, Y., & Liu, X. (2008). Personalized interactive faceted search. In *Proceedings of the 17th International Conference on World Wide Web*, pp. 477–486.

Krivosheev, E., Casati, F., Baez, M., & Benatallah, B. (2018). Combining crowd and machines for multi-predicate item screening. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW), pp. 1–18.

Kuechler, B., & Vaishnavi, V. (2008). On theory development in design science research: anatomy of a research project. *European Journal of Information Systems*, 17(5), pp. 489–504.

Ley, M. (2009). DBLP: Some lessons learned. *Proceedings of the VLDB Endowment*, 2(2), pp. 1493–1500.

Liu, B. (2021). In AI we trust? Effects of agency locus and transparency on uncertainty reduction in human–AI interaction. *Journal of Computer-Mediated Communication*, 26(6), pp. 384–402.

Lukyanenko, R., Wiggins, A., & Rosser, H. K. (2020). Citizen science: An information quality research frontier. *Information Systems Frontiers*, 22(4), pp. 961–983.

Luz, N., Silva, N., & Novais, P. (2015). A survey of task-oriented crowdsourcing. *Artificial Intelligence Review*, vol. 44, no. 2, pp. 187–213.

Ma, S., Zhang, C., and Liu, X. (2020). A review of citation recommendation: From textual content to enriched context. *Scientometrics*, 122(3), pp. 1445–1472.

Micchi, G., Bigo, L., Giraud, M., Groult, R., & Levé, F. (2021). I keep counting: An experiment in human/AI co-creative songwriting. *Transactions of the International Society for Music Information Retrieval*, 4(1), pp. 263–275.

Mittleman, D. D., Briggs, R. O., Murphy, J., & Davis, A. (2008). Toward a taxonomy of groupware technologies. In *Proceedings of the 14th International Workshop on Groupware: Design, Implementation, and Use*, pp. 305–317.

Nakagawa, S., Samarasinghe, G., Haddaway, N. R., Westgate, M. J., O'Dea, R. E., Noble, D. W., & Lagisz, M. (2019). Research weaving: Visualizing the future of research synthesis. *Trends in Ecology & Evolution*, vol. 34, no. 3, pp. 224–238.

Noel-Storr, A. H., Redmond, P., Lamé, G., Liberati, E., Kelly, S., Miller, L., Dooley, G., Paterson, A., & Burt, J. (2021). Crowdsourcing citation-screening in a mixed-studies systematic review: A feasibility study. *BMC Medical Research Methodology*, 21(1), pp. 1–10.

Peeters, M. M., van Diggelen, J., Van Den Bosch, K., Bronkhorst, A., Neerincx, M. A., Schraagen, J. M., & Raaijmakers, S. (2021). Hybrid collective intelligence in a human–AI society. *AI & Society*, vol. 36, no. 1, pp. 217–238.

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, 24(3), pp. 45–77.

Price, S., & Flach, P. A. (2017). Computational support for academic peer review: A perspective from artificial intelligence. *Communications of the ACM*, 60(3), pp. 70–79.

Rohde, M., Stevens, G., Brödner, P., & Wulf, V. (2009). Towards a paradigmatic shift in IS: Designing for social practice. In *Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technology*, pp. 1–11.

Rosser, H., & Wiggins, A. (2019). Crowds and camera traps: Genres in online citizen science projects. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*, pp. 5289-5298.

Rzeszotarski, J., & Kittur, A. (2012). CrowdScape: Interactively visualizing user behavior and output. In *Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology*, pp. 55–62.

Sanyal, D. K., Bhowmick, P. K., & Das, P. P. (2021). A review of author name disambiguation techniques for the PubMed bibliographic database. *Journal of Information Science*, 47(2), pp. 227–254.

Schmiedel, T., Müller, O., & Vom Brocke, J. (2019). Topic modeling as a strategy of inquiry in organizational research: A tutorial with an application example on organizational culture. *Organizational Research Methods*, vol. 22, no. 4, pp. 941–968.

Schroder, A., Constantiou, I., Tuunainen, V. K., & Austin, R. D. (2022). Human-AI collaboration – Coordinating automation and augmentation tasks in a digital service company. In *Proceedings of the 55th Hawaii International Conference on System Sciences*, pp. 206–215.

Seeber, I., Bittner, E., Briggs, R. O., de Vreede, T., de Vreede, G. J., Elkins, A., Maier, R., Merz, A. B., Oeste-Reiß, S., Randrup, N., Schwabe, G., & Söllner, M. (2020). Machines as teammates: A research agenda on AI in team collaboration. *Information & Management*, 57(2), 103174.

Shneiderman, B. (1996). The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings of the 1996 IEEE Symposium on Visual Languages*, pp. 336–343.

Singh, S., Jain, S., & Jha, S. S. (2023). On subset selection of multiple humans to improve human-AI team accuracy. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, pp. 317–325.

Suh, M., Youngblom, E., Terry, M., & Cai, C. J. (2021). AI as social glue: Uncovering the roles of deep generative AI during social music composition. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–11.

Swanson, D. R., & Smalheiser, N. R. (1997). An interactive system for finding complementary literatures: A stimulus to scientific discovery. *Artificial Intelligence*, vol. 91, no. 2, pp. 183–203.

Tchoua, R. B., Chard, K., Audus, D. J., Ward, L. T., Lequieu, J., De Pablo, J. J., & Foster, I. T. (2017). Towards a hybrid human-computer scientific information extraction pipeline. In *Proceedings of the 2017 IEEE 13th International Conference on e-Science*, pp. 109–118.

Thilakaratne, M., Falkner, K., & Atapattu, T. (2019). A systematic review on literature-based discovery: General overview, methodology, & statistical analysis. *ACM Computing Surveys*, vol. 52, no. 6, pp. 1–34.

Thomas, J., & Zaytseva, A. (2016). Mapping complexity/human knowledge as a complex adaptive system. *Complexity*, 21(S2), pp. 207–234.

Tokarchuk, O., Cuel, R., & Zamarian, M. (2012). Analyzing crowd labor and designing incentives for humans in the loop. *IEEE Internet Computing*, 16(5), pp. 45–51.

Vincent-Lamarre, P., & Larivière, V. (2023). Are self-citations a normal feature of knowledge accumulation?. *arXiv preprint arXiv:2303.02667*.

Vinella, F. L., Hu, J., Lykourentzou, I., & Masthoff, J. (2022). Crowdsourcing team formation with worker-centered modeling. *Frontiers in Artificial Intelligence*, 102.

Vössing, M., Kühl, N., Lind, M., & Satzger, G. (2022). Designing transparency for effective human-AI collaboration. *Information Systems Frontiers*, pp. 1–19.

Wagner, G., Lukyanenko, R., & Paré, G. (2022). Artificial intelligence and the conduct of literature reviews. *Journal of Information Technology*, vol. 37, no. 2, pp. 209–226.

Waltz, D., & Buchanan, B. G. (2009). Automating science. *Science*, vol. 324, no. 5923, pp. 43–44.

Wang, S., & Koopman, R. (2017). Clustering articles based on semantic similarity. *Scientometrics*, vol. 111, no. 2, pp. 1017–1031.

Wang, W., Jiang, X., Tian, S., Liu, P., Dang, D., Su, Y., Lookman, T., & Xie, J. (2022). Automated pipeline for superalloy data by text mining. *npj Computational Materials*, 8(1), pp. 1–12.

Wiethof, C., & Bittner, E. A. (2022). Toward a hybrid intelligence system in customer service: Collaborative learning of human and AI. In *Proceedings of the 30th European Conference on Information Systems*, 66.

Yang, Q., Steinfeld, A., Rosé, C., & Zimmerman, J. (2020). Re-examining whether, why, and how human-AI interaction is uniquely difficult to design. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–13.

Zhang, J., & Yu, W. (2020). Early detection of technology opportunity based on analogy design and phrase semantic representation. *Scientometrics*, vol. 125, no. 1, pp. 551–576.

Zhang, R., McNeese, N. J., Freeman, G., & Musick, G. (2021). "An Ideal Human": Expectations of AI teammates in human-AI teaming. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW3), pp. 1–25.

Appendix

Table 2. Excerpt from the survey with participants' quotes

Question	Participant quote
How willing do you think a citizen scientist is motivated or interested to participate in crowd-driven research by collecting or analyzing the type of data that you work with?	"*I study engagement and motivation in this kind of work through games and play.*"
In your opinion, which forms of crowdsourcing are best suited for research? What are the best practices and the main drawbacks of using crowdsourcing for scientific purposes?	"*Crowdsourcing is best when you can chunk data analysis into bite-sized problems that can be solved quickly. This is also a drawback as you only get micro data this way. Also, motivation can be low as it seems like it just offloading the monotonous and not allowing the "citizen" to be engaged and doing the science.*" "*Best practice would be to use the crowd to create or validate observations of targets of interest for research by providing ontologies from where to select classes, descriptions, and metadata such as self assessed confidence which is important to evaluate trust of the data.*"
Integration of human input into AI systems offers great promise for the development of practical applications. What kind of features and/or forms of intelligence would you like to be added to those platforms to understand scientific data at large scale?	"*Using AI to automatically doing quality check would be great. Poor work and poor workers should be labeled so to intervene.*" "*Automatic reasoning for large scale scientific data should be widely applied nowadays. As far as I know, astronomers rely on machine learning techniques to process tons of data collected by radio telescope arrays. Another integration I think would be impactful is to let machine teach crowds to participate in highly convoluted tasks.*"
What do you think about a hybrid, mixed-initiative crowd-computing system combining machine and human intelligence to overcome the shortcomings of existing AI and crowd-powered systems for scientific purposes? What are the implications for design?	"*I think machine learning algorithms could learn from user/volunteer responses in applications where the computer classification skill is below a desirable threshold.*" "*It sounds good, crowd-computing can fill the gap that AI cannot address (yet).*" "*Error tolerance is one problem, but it will never eliminate the ethical dilemmas of introducing AI.*" "*Such designs exist, usually the crowd trains the machine. Care has to be taken that this does not turn the other way round, i.e. the machine imposing (wrong) interpretations.*"

Table 3. Overview of coded elements from the survey and derived requirements

1st-order elements	2nd-order elements	Aggregate dimensions	E / B*	DR
Wider retrieval coverage allowing to search studies over multiple sources from different application domains	Comprehensiveness	Information retrieval	E	DR1, DR2
Low precision and (…) noise effect during retrieval	Retrieval performance		B	DR1
Seeking out and exploring papers that help to understand a scientific space or craft a specific argument by narrowing in on specific research areas instead of engaging with the overwhelming totality of a scientific field or discipline	Specificity		E	DR2, DR5
Filtering data according to the needs	Filtering (faceted search)		E	DR2
Providing the ability to search for specific questions through content- and keyword-based queries	Query syntax		E	DR2
Search query language is limited			B	DR2
Lack of support for semantic search			B	DR2, DR3
Difficult to find specific research topics			B	DR2
Automatic identification of semantic links among scientific data	(Semantic) linkages	Data	E	DR3, DR4
Lack of information on relationships (connections) and differences between documents			B	DR3, DR5
Keeping accessible and usable notes for further reuse	Data reuse		E	DR5
Ability to export in multiple formats (e.g., BibTex)	Format		E	DR1
Poor performance in entity matching (e.g., authorship attribution)	Entity matching		B	DR3, DR4, DR5
Too much information available	Overload		B	DR1, DR5
Provenance is key to keep track of all versions of the information and knowledge capture, so that the state of knowledge can be rewound and replayed from (and to) any point in time	Provenance		E	DR1
Data quality can be an issue and the information obtained may not be useful	Quality		B	DR1
Reliability and authenticity of information and sources must be ensured for data analysis and query purposes	Reliability (validity-as-credibility)		E	DR1, DR3, DR5
Time consuming to check whether a paper is relevant and why it is cited	Relevance		B	DR5
Ranking features for tracking progress and publication relevance	Rank (score)		E	DR5
Results may be not replicable and dramatically vary depending on a number of factors hard to control	Replicability		B	DR3
Need for data at large spatial and temporal scales	Scale		E	DR1, DR2, DR5
Metadata is often incomplete or malformed (e.g., missing citations)	Sparsity (incompleteness)		B	DR1, DR5
Standardized input to alleviate the problem of incompatible formats	Standardization		E	DR1
Transparency about the use of data	Transparency		E	DR3
Proprietary access (documents behind paywalls) and copyrighted material	Access and availability		B	DR3
All sharing of data (e.g., sensitive content) must soon be assessed against the general data protection regulation (GDPR)	Security & privacy (regulation)		E	DR1, DR3
Keeping track of the latest works on a specific topic by following certain conferences, journals, authors, keywords, etc.	Traceability (notification/alert mechanisms)		E	DR3

* Potential enablers (**E**) and blockers (**B**) for the use of SciCrowd system.

1st-order elements *(cont.)*	2nd-order elements	Aggregate dimensions	E / B	DR
Graph of publications based on the impact of the publications	Visualization (impact)	Data *(cont.)*	E	DR5
Citation count and aggregation with emphasis on the publications citing the work of interest	Visualization (networked structure)		E	DR5
Quick overview on a topic of interest (e.g., experimental results)	Visualization (summarization)		E	DR5
Highlighting trends and differences among entities (e.g., authors and institutions)	Visualization (trend analysis)		E	DR5
Elements of gamification and good interface development should be employed to provide a pleasant and engaging experience	Gamification and user interface design	Platform	E	DR3
Improving accessibility is crucial to support fair and equitable engagements	Accessibility		E	DR3
None of current literature search systems allow to search and store, annotate, and summarize literature	Comprehensiveness		B	DR1, DR2, DR5
Considerable effort for experiment setup (limited configuration capabilities)	Configurability		B	DR3
Difficulty to develop and maintain dedicated, specific crowdsourcing systems	Development and maintenance		B	DR3
The platform must be easy to use	Ease of use		E	DR3
Discrimination by the platform	Equality and discrimination		B	DR3
Limited interoperability capabilities	Interoperability		B	DR3
Systems which require sign in or push to invite other users tend to be more obtrusive to the end-users	Authentication		B	DR3
There must be comprehension capability (automated reasoning) of the AI system to understand large-scale scientific data	Automated reasoning (logical inference)	AI-driven interaction	E	DR4
Automatic recognition of outliers, trends and correlations	Data mining and pattern recognition		E	DR4
Elaborate explanations of what is discovered	Explainability		E	DR4
User-friendly AI algorithms	User-friendliness		E	DR4
Using AI to automatically doing quality check would be beneficial	Quality control		E	DR4
Human oversight (e.g., vigilance against bots) is necessary in some cases	Supervision		B	DR3
Extreme overfitting on training data and thus be unable to generalize concepts	Training		B	DR3
Ethical dilemmas of introducing AI	Ethical issues		B	DR4
Overfitting to crowd workers and bias may be problematic	Bias	Citizen science & paid crowd work	B	DR3
Attribution and credit for work done by making sure humans see their own contributions as part of a whole (value human input)	Credit (recognition)		E	DR3
Crowd behavior analysis leads to better knowledge/information about the workers or volunteers and thus improve the general outcomes (e.g., task performance) and user experience	Crowd behavior		E	DR3
Relying on the citizen science community to challenge ideas, highlight the most relevant publications, etc.	Diversity of views		E	DR3
Crowd must have certain level of understanding otherwise somebody can’t reach to right conclusions	Expertise		B	DR3
Make research work open for all individuals so that we engage diverse voices, irrespective of their education and income levels, race, gender, etc.	Openness and inclusivity		E	DR3
Lack of responses	Participation		B	DR3

1st-order elements *(cont.)*	2nd-order elements	Aggregate dimensions	E / B	DR
Poor work and careless users must be labeled so to intervene	Quality control (worker assessment)	Citizen science & paid crowd work *(cont.)*	E	DR3
Reaching audiences that might be suitable for research as much as possible since general crowds might not be able to perform highly specialized tasks like scientific analysis	Suitability		E	DR3
Designing well-defined microtasks which can be performed without deep contextual knowledge	Task design (decomposition)		E	DR3
Suggesting a task design depending on the nature of the problem and the individual characteristics (e.g., personal interests and capabilities) of each member of the crowd	Task design (personalization)		E	DR3
Contributors may feel to be observed	Security & privacy (monitoring and control)		B	DR3
Keep questions/observations simple to get the crowd to engage in data collection and processing	Motivation & engagement (clarity/simplicity)		E	DR3
Incentivizing people and (…) sustaining participation is challenging	Motivation & engagement (incentives)		B	DR3
Using the crowd to create or validate observations of targets of interest for research	Reliability & trust (authenticity)		E	DR3, DR5
Possibility of confidence assessment	Reliability & trust (confidence)		E	DR3
Collaborative crowdsourcing is appropriate for scientific research	Collaboration		E	DR3
Ability to recommend items (e.g., papers)	Recommendation		E	DR3
Software and inputs of the specific study should be shareable	Shared artifacts (data exchange)		E	DR3
Feedback and comments might be very valuable by engaging participants in social discussion of their results, questions, and activities	Social discussion (feedback)		E	DR3
Let contributors get feedback from scientists using the data	Social discussion (feedback)		E	DR3
Algorithms could learn from user/volunteer responses in applications where the computer classification skill is below a desirable threshold and vice-versa	Complementariness	Human-AI interaction	E	DR4
A hybrid approach makes sure that the developed solution scales to large amounts of data and comes closer to solving real-life problems (thanks to the hybrid collective intelligence)	Complementariness		E	DR4
Human-AI interface development is a challenge	Interface		B	DR3, DR4
Task delegation by leaving easy tasks to computers and grey areas to humans	Delegation in human-AI teaming		E	DR4
Accountability for wrong outcomes (error-handling mechanisms) is one problem	Error tolerance and mitigation		B	DR4
Care has to be taken when the machine is imposing (wrong) interpretations to the crowd	Error tolerance and mitigation		B	DR4
Need of authority for managing conflicts	Conflict management		B	DR4
More aspects of the human cognitive processes should be included in AI	Cognition		E	DR4

Requirements

DR1 – Automate or semi-automate data acquisition (from multiple sources), cleaning (e.g., removal of duplicates) and integration.

DR2 – Support exploration using a variety of criteria (faceted search).

DR3 – Allow crowd users to explore data, configure searches, visualize classifications, verify data, and share information through annotations.

DR4 – Provide inferences, suggestions and interactive explanations.

DR5 – Provide a statistics panel with bibliometric indicators.