Pilot analysis - Computer Science

Angelo Di Iorio, Kumar Guha, Silvio Peroni, Laurent Romary, Thanasis Vergoulis

Pilot analysis - Computer Science

Authors: Angelo Di Iorio (Unibo), Kumar Guha (Inria), Silvio Peroni (Unibo), Laurent romary (Inria), Thanasis Vergoulis (Athena)

1. State of affairs

a. Open science and/or research assessment

The present pilot deals with the computer science domain which covers a wide span of research areas ranging from highly theoretical topics (e.g. complexity, graph theory etc.) to practical or experimental areas, often in a multidisciplinary context (e.g. computational biology, natural language processing or digital humanities). By construction, computer science productions go beyond the sole publication of scholarly papers, with the strong importance of software and correlatively of datasets.

Computer Science has a long-standing tradition of openness, not the least because the production of software, even at the small scale of research projects, requires opening and sharing software components and libraries so that others can test, reuse and eventually contribute to its development, as demonstrated by the early history of the unix operating system (Bretthauer, 2001). This has led to a whole range of contributions around Open Source Software (OSS), with technical facilities such as forges to collectively maintain contributions, versions and dissemination of software, as well as a solid cumulative experience in the domain of licensing in an open context. It should be noted that forge platforms have also been a systematic host for various types of other objects, for instance datasets, either in relation to specific pieces of software (parameters, regression tests etc.) or in isolation for various data modelling or production projects[1].

Beside software development itself, the CS domain has been an early adopter of technologies and platforms facilitating the dissemination of research results. For instance CS researchers have adopted the arXiv publication repository since its onset in 1991 and have become the main contributing domain as shown in Figure 1. From a wider perspective, a profound culture of disseminating preprints has pervaded the CS domain (Lin et al., 2020), with some communities such as cryptology[2] deploying their own preprint servers.

[1] For instance the Text Encoding Initiative, which is a standardisation consortium producing guidelines for representing digital texts of all sorts, is entirely maintained on the GitHub forge (https://github.com/TEIC).

[2] Cryptology ePrint Archive: https://eprint.iacr.org/

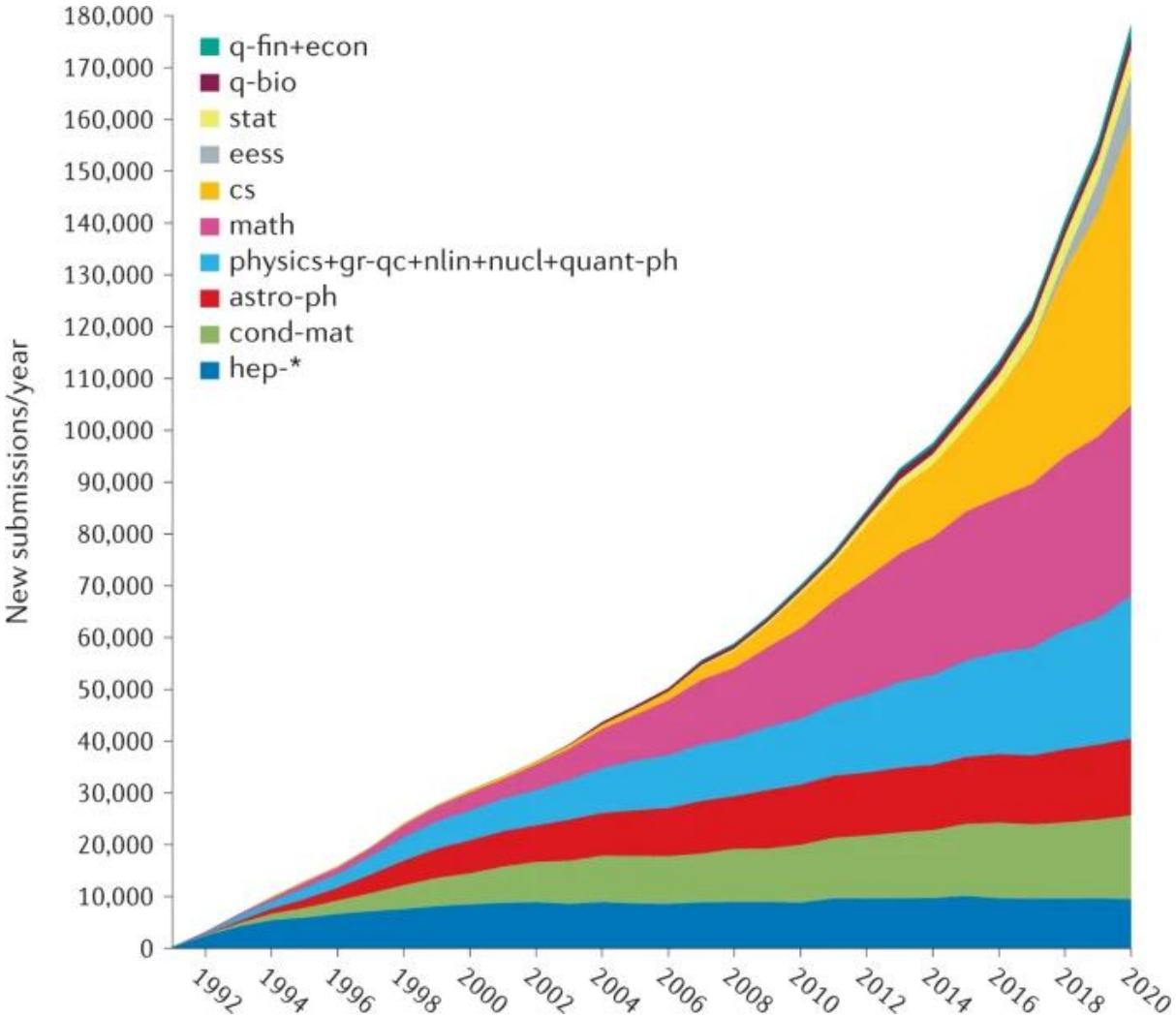


Figure 1: Overview of yearly submission over the course of arXiv activity (source: Ginsparg 2021)

Another important feature of publication practices in the CS domain is the essential role of workshops and conferences, which is unique within academic communities. Not only do conferences and workshops represent a higher share of published articles within computer science, but also the quality peer-review that they bear makes them serve as important contributions of CS researchers, carrying significant weight in research assessment processes, thereby attracting a majority of citations (Vrettas and Sanderson, 2014). They are thus recognised as a central place for quality publication, although not always acknowledged within other fields, for instance in the context of multidisciplinary recruitment or promotion panels. As a consequence, the coverage of computer science publications is lower in commercial bibliographic databases (Kuserow & Groppe, 2014) and there are less reliable infrastructures for the provision of authority data for conferences as would be needed for the field.

From a pure open access perspective, as we have already observed, the CS domain has a long-standing involvement in opening up publications. Some institutions have even taken the lead at national level, with Inria reaching 89% of open access publications for 2022, essentially in the French national repository HAL and in arXiv. According to the FoS classes[3] provided by OpenAIRE as, there are two main related fields:

- Computer & Information Sciences:

[3] Fields of Science, see https://explore.openaire.eu/fields-of-science

- 98,078 research products
- 65,265 of them (66.5%) being open

- Electrical Engineering, Electronic Engineering & Information Engineering:
  - 2,026,965 research products
  - 820,858 of them (40.5%) being open

If we make a quick comparison with other disciplines, we can see (Figure 2, from the French open science monitor[4]) that Computer science, together with Mathematics, which is closely related, has a specific profile of favouring so-called green open access (i.e. author's manuscripts deposited in a publication repository) and have a low level of presence in gold open access journals (i.e. with author-pays fees, aka APC - Article Processing Charges).

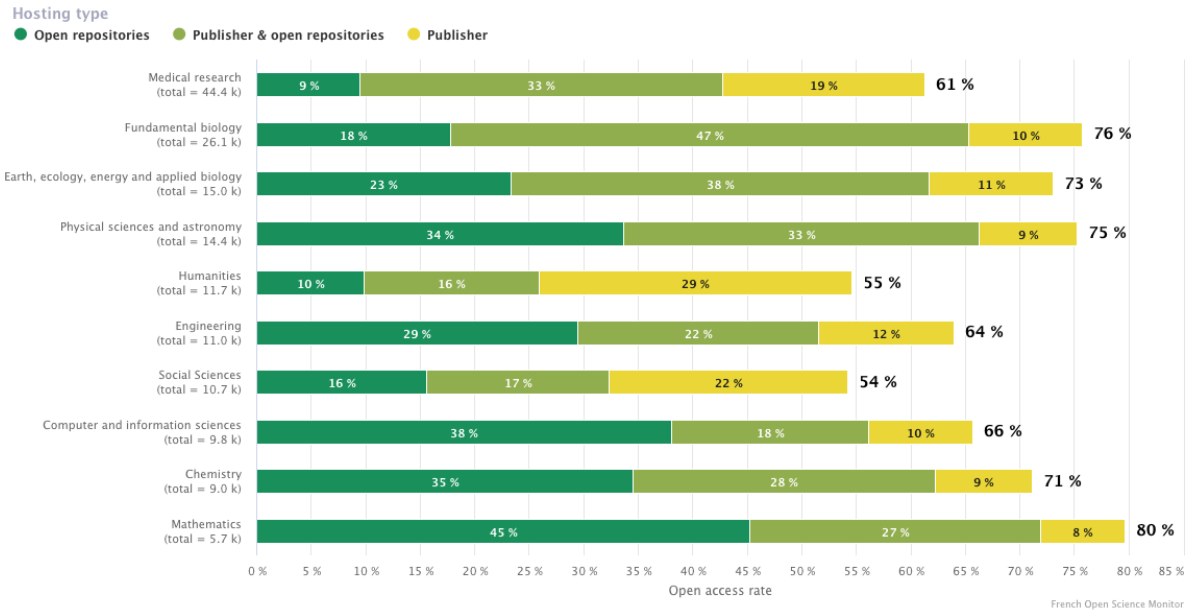


Figure 2: Distribution of publications in France, with a Crossref DOI, by opening route for each discipline (publications of 2021)

From the point of view of research assessment, there is a general international movement towards a more qualitative oriented assessment of research that expressed itself through the DORA declaration[5] and more recently with the setting up of CoARA, the Coalition for Advancing Research Assessment. In particular, the urge to take a wider range of research outputs in consideration[6], is in phase with the expectations of the computer science field. It is thus no surprise that several ICT institutions[7] have joined the CoARA membership.

The participation of ICT institutions combined with the strength of a consortium such as GraspOS, could be the basis for the setting up of recommendations related to research software

[4] https://frenchopensciencemonitor.esr.gouv.fr/
[5] https://sfdora.org
[6] The Agreement on Research Assessment that lead to the creation of COARA explicitly refers to "the full range of research outputs, such as scientific publications, data software, models, methods, theories, algorithms, protocols, workflows, exhibitions, strategies, policy contributions, etc.".
[7] Inria, Inesc TEC, IBICT, GRIN, SCIE, DARIAH, see https://coara.eu/coalition/membership/

when assessing researchers, on the basis of seminal works such as (Alliez et al. 2019; Canteaut et al. 2021), which has been carried out within Inria.

b. Tools, services and data used

The computer science research community already benefits from a wealth of generic or thematic infrastructures at the service of the open and free dissemination of their research outputs. We have already seen how computer science has been among the early adopters of arXiv and represents a topical majority of its content.

One of the essential infrastructure available to the CS domain is of course Software Heritage, which is a universal archive of open source code. Software Heritage harvests all (GitHub, Gitlab etc.) open software forges, with now more than 2 billion source files integrated in the archive. Constituted as a foundation in collaboration with Unesco, it aims at being a stable service that we can use to register, identify or document research software code. For instance, Software Heritage associates a single intrinsic identifier (SWHId) to each single file that allows anyone to reference and disambiguate its software, for instance in the context of an assessment report. Experiments have also been carried out to relate Software Heritage entries with publication repositories. For instance there is now a full cross-referencing workflow between the French national publication repository HAL and Software Heritage. Software Heritage is also a central contributor to ongoing international efforts within EOSC and RDA to compile best practices, guidelines or standards for software documentation and citation.

In terms of bibliographic databases and digital libraries, there are several prominent initiatives to host and disseminate content related to the computer science domain:

- DBLP (dblp: computer science bibliography) is a computer science bibliographic database set up at Trier University (Germany). Initiated in 1993 as a web technology test, its corpus of bibliographic references holds now over 6M records, more than half of which being conference papers. It is now operated by Schloss Dagstuhl – Leibniz Center for Informatics (Germany). Its content is open source and it is used for different purposes, including bibliometrics. However, person name disambiguation is stated as the main challenge (Ley M. 2009), imported data goes through both algorithmic and manual process of curation for disambiguation (Kim 2018);
- CiteSeerX (https://citeseerx.ist.psu.edu) is a bibliographic database in the domains of computer and information sciences, maintained at the College of Information Sciences and Technology at Pennsylvania State University. Initiated as an indexing and citation tracking portal, it also links indexed documents to other sources of metadata such as DBLP and the ACM Portal;
- CEUR (https://ceur-ws.org/) is a free open-access publication service hosted by the RWTH Aachen University in Germany. Recognized by the community, and established as an ISSN publication series, the project started in 1995 and the number of published volumes increased constantly up to about 250 per year in the recent past. As of June 2023, it contains 3423 volumes, including PDF articles and covering heterogeneous topics in Computer Science;
- The ACM, one of the two main scholarly associations in computer science, has set up a digital library with all the published content from its journals but also the various ACM related conferences. The ACM has also been maintaining an extensive thesaurus, the ACM Computing Classification System (https://dl.acm.org/ccs), for the computer science domain, which is fully open and usable in other services. For instance, the Inria portal of the French national repository HAL offers the possibility to index entries according to CCS. Besides, it should be noticed that ACM is planning to flip all its journals in a full open access model in the context of a full transparent disclosure of its business breakout;

- IFIP, the International Federation for Information Processing, the third scholarly societe in computer science, has set up a fully open digital library for disseminating its legacy and current proceedings (https://ifip.hal.science/). Hosted on the HAL platform with the support from Inria, it currently contains more than 18 000 documents from all the thematic domains covered by IFIP;
- Even if we have to mention IEEE, as probably the most important scholarly society in computer science, its contribution to open science has been minimal as the institution has mainly behaved as a private publisher with hard commercial practices in the recent years.

In addition, there are several other collections of open bibliographic and citation data that, even if their coverage go beyond Computer Science, also contain relevant material that can support the case studies. These are:

- OpenCitations[8], an independent, community-led, and not-for-profit Open Science infrastructure organisation, which aims at harvesting and openly publishing accurate and comprehensive metadata describing the world's academic publications and the scholarly citations that link them. It provides this information, both in human-readable form and interoperable machine-readable Linked Open Data formats, under open licences at zero cost and without restriction for third-party analysis and re-use. The two main collections it publishes are the OpenCitations Index (containing citation links) and OpenCitations Meta (containing bibliographic metadata of the bibliographic resources involved in citations). It also covers several computer science publications.
- OpenAIRE Graph, a large open scholarly record collection, which is key in fostering Open Science and establishing its practices in the daily research activities. Conceived as a public and transparent good, populated out of data sources trusted by scientists, the Graph aims at bringing discovery, monitoring, and assessment of science back in the hands of the scientific community. Imagine a vast collection of research products all linked together, contextualised and openly available. For the past years OpenAIRE has been working to gather this valuable record. It is a massive collection of metadata and links between scientific products such as articles, datasets, software, and other research products, entities like organisations, funders, funding streams, projects, communities, and data sources.
- CORE[9], a not-for-profit service dedicated to the open access mission that provides access to a large collection of open access research papers, collecting and indexing research from repositories and journals. It includes more than 300 thousand research articles in Computer Science.
- Semantic Scholar, a service that provides free, AI-driven search and discovery tools, and open resources for the global research community. It indexes over 200 million academic papers sourced from publisher partnerships, data providers, and web crawls. It contains metadata about several Computer Science papers.

c. Gaps in terms of practices, tools, services and datasets used

As we have seen, the computer science domain is already well-equipped with solid open science infrastructures for hosting and disseminating its core research productions. Still, we are at a crossroad when linking these infrastructures with general research assessment practices, as the latter still mostly rely on older generation databases and principles that have never really fulfilled the needs of computer science at large. In the context of the present project, we do not

[8] https://opencitations.net/
[9] https://www.core.edu.au/

aim at making a revolution in the domain of open science based research assessment but rather identify a few lacunae concerning which we may make progress in the mid-term. Among the missing components that we have identified while working on this computer science pilot analysis, we can put forward the following elements:

- Although we can identify a variety of stable repositories and archives for research publications, data sets and software, the capacity we have to link between those objects is still rather poor. In particular, two items could be improved: a) ensuring better cross-referencing mechanisms between publication repositories and data repositories and Software Heritage (for open software sources) and b) making progress in the automatic detection or references to data sets and software in publications, experimenting on a corpus where full texts are available;
- We need to have better coverage of conferences as an essential publication channel in computer science. In particular some insights have to be gained in having a reliable and structured authority list of conferences to control the metadata used in publication and data repositories;
- On a more prospective level, we need to see how the CS community could be in the position to express some priorities for action towards CoARA, either as a group of involved institutions, or via existing European organisations such as Informatics Europe[10] or ERCIM.

2. Evaluation context

As hinted in the preceding sections, open science is deeply entrenched in the research practices and values in computer science and in particular we have set forth the specificities of the field in terms of publications, with the central role of conferences, and other research productions, with the role of software as a major research output. This paves the way for identifying dedicated profiles for research assessment. We have also hinted at the fact that computer science does not come in isolation and is often a component within a highly multidisciplinary context.

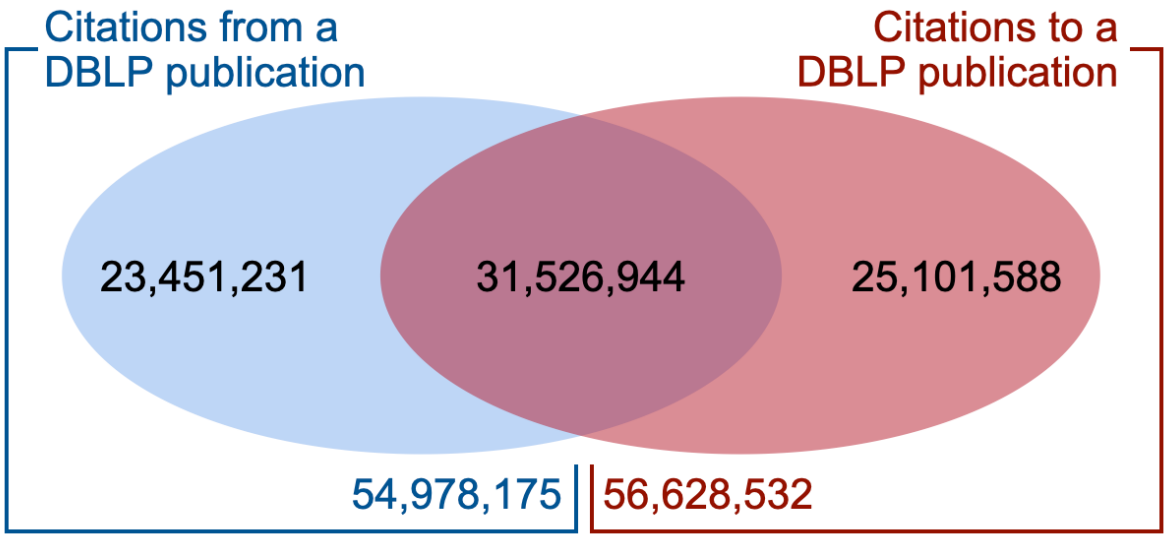
The current situation related to research assessment in computer science is clearly not limited to one type of stakeholder: if we are to make advances in this domain and make a better usage of what is offered with the widening of the adoption of open science principles, we need to ensure that it is beneficial to all levels of the HER stakeholders, institutions, laboratories, research groups, funded project and above all, researchers themselves, so that they feel part of the decision process but also that they can make use of any type of infrastructural component that could be set in place. For the sake of the present project we would actually put researchers and institutions as two priority target groups with the following use cases in mind:

- providing **researchers** with the mean to create one's own personal dashboard to simplify references when providing content for assessments or promotions purposes, but also to gain a better view of one's global production (publication, data and software);

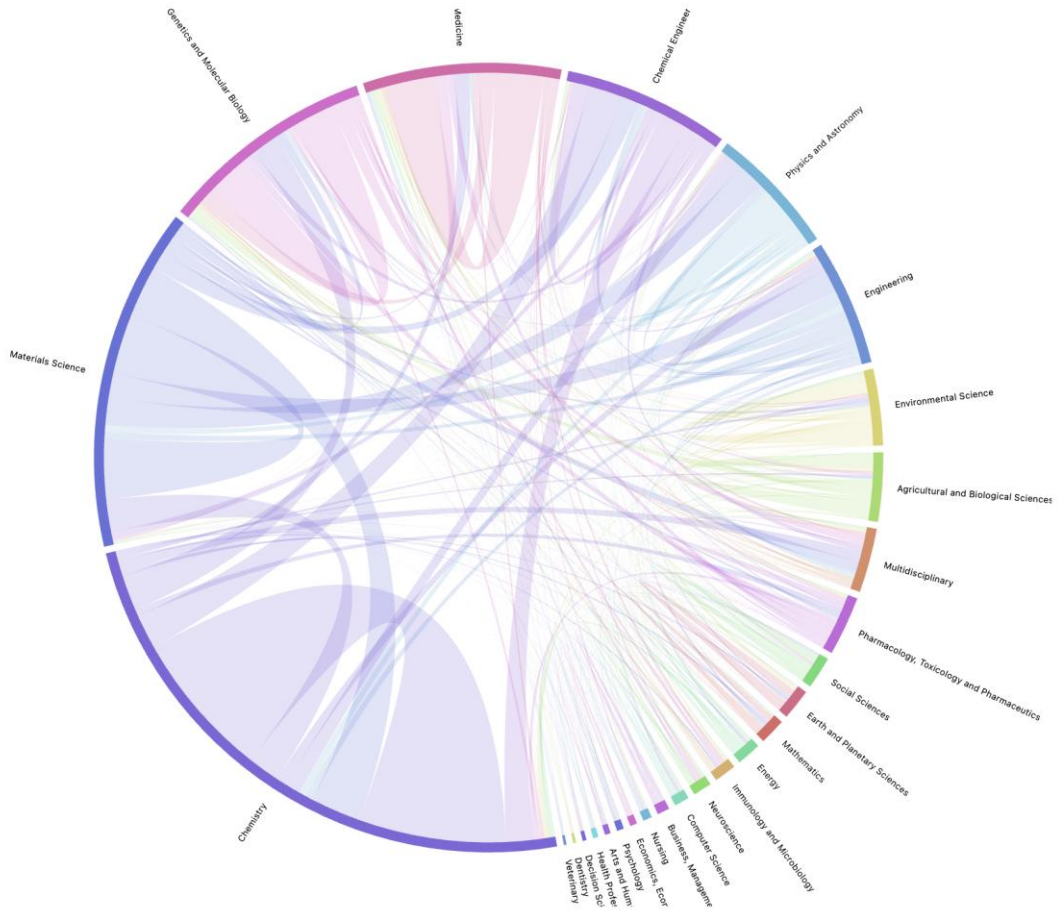
- helping **institutions** design better indicators of their research activities on the basis of high quality open data rather than privately held services which are both closed (in particular for mining and reuse purposes) and bear a poor coverage of the computer science domain. For instance, we could imagine observing patterns of international collaboration from the analysis of affiliations in publication authorships.

[10] See for instance the preliminary work accessible from https://www.informatics-europe.org/research/research-evaluation.html

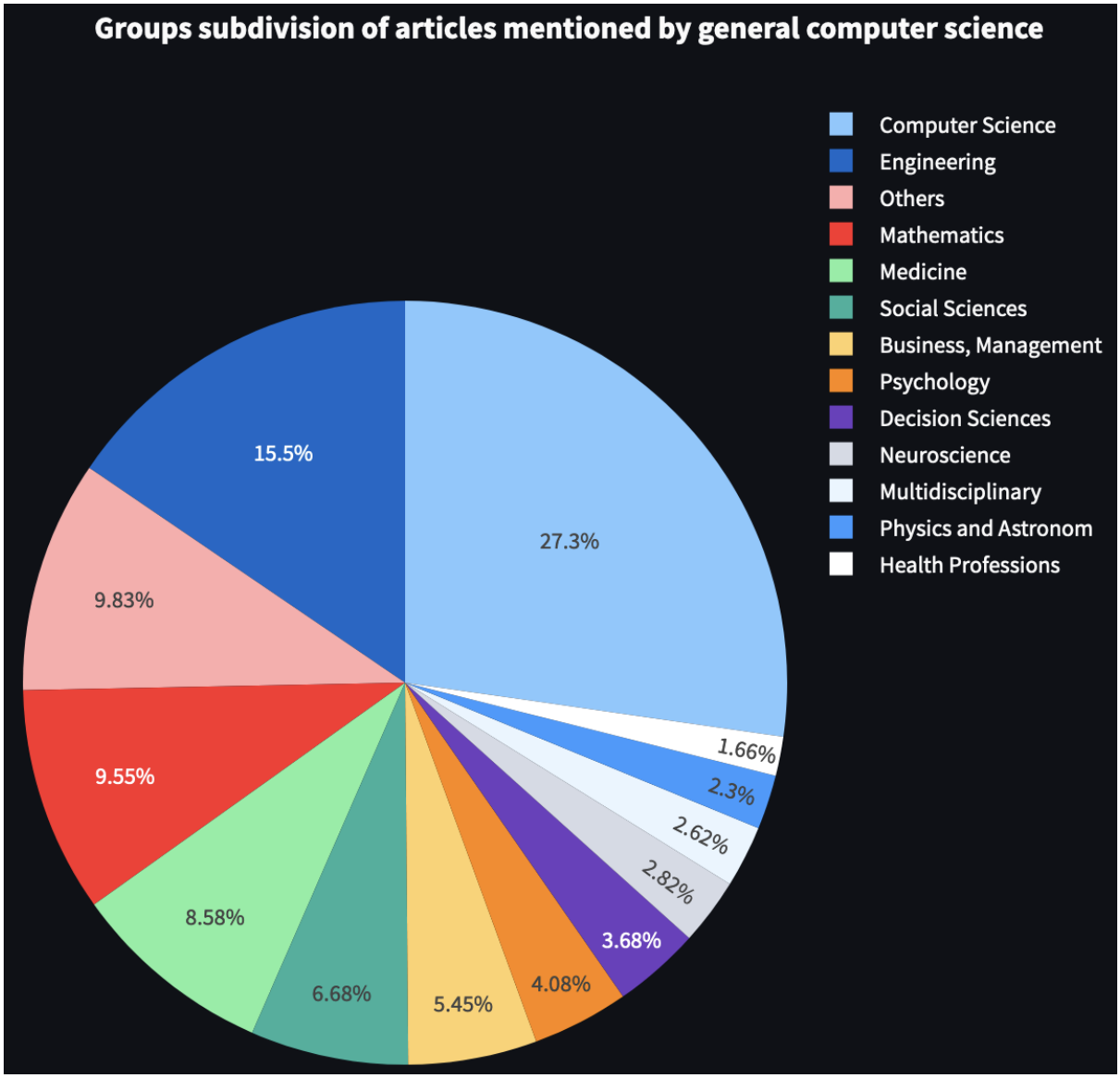
One of the measures to analyse and consider would be that related with the ratio of the reuse of Computer Science research by other disciplines and how much CS takes from other disciplines. For instance, using DBLP as a proxy to understand if a particular publication belongs to the Computer Science research, it is possible to understand how many citations in a collection, e.g. OpenCitations, involve Computer Science publications by comparing the DOIs of citing and cited entities with those available in DBLP. For instance, as introduced in https://querty.hypotheses.org/22, using the OpenCitations' COCI September 2021 dump and DBLP October dump, we found that more than 80 million citations in COCI involved at least one of the 4,637,865 entities in DBLP (considering only journal articles, conference proceedings papers, books and book chapters). As shown in the figure below, only 39% of these citations are between citing and cited entities both included in DBLP, while the rest of them either come from or go to publications not listed in DBLP – that, potentially, could not be Computer Science publications.



Indeed, from https://alerosae.github.io/OpenCitationsCharts/ which again uses COCI data and ASJC subject categories, it is clear that Computer Science (as well as other disciplines) has several interactions (shown via citations) to other fields. Indeed, citations represent connections between articles and, in turn, between subject areas. In the diagram below, it is possible to see how journal articles in different disciplines interact with other disciplines.



In addition, by looking at a customised diagram for Computer Science obtained for the same data from https://alerosae-opencitationscharts-app-web-oc-fe7rco.streamlit.app/, it is clear that Computer Science research interact a lot of other disciplines. Measuring such level of multidisciplinarity would be a reasonable indicator to use to understand how much CS contributes to the research of other fields.



additional diagrams to include from: https://alerosae-opencitationscharts-app-web-oc-fe7rco.streamlit.app/

As stated before, the comparison between various research organisations will be key to the identification of general trends of a more qualitatively based research assessment, in particular in relation to European organisations such as Informatics Europe or ERCIM, but also national organisation such as the SIF (Société Informatique de france). This will also help define priorities as to which infrastructural lacunae we need to fill in the context of the project.

3. Ambitions in terms of developing new ways of evaluating/monitoring open science

Given the strong open science culture in the computer science domain together with the variety of objects that researchers in the fields have to deal with in their everyday research practices, we have identified various ambitions that should be pursued to further a better open science based research assessment in computer science. These ambitions relate to three main aspects:

- **criteria** for research assessment: setting up a simple yet comprehensive overview of relevant criteria for research assessment in computer science, which can possibly be based upon openly accessible information;
- **monitoring**: we need to set in place the relevant monitoring mechanisms where we obviously lack data;
- **tooling**: we indicate where, to our knowledge, a specific development emphasis should be put in order to better equip the computer science field with means to deploy and exploit research productions.

From the point of view of criteria, we need to gather a compendium of existing but also possible assessment criteria for individuals and organisations that could be relevant in computer science and see which could be the result of some information extraction activity from the existing open resources available to our field. This should comprise some elaboration from the work of [Canteaut et al. 2021] and see whether there could be some room for a future white paper on the assessment of research software.

On the monitoring side, the recent progress made within the French open science monitor (BSO - Baromètre de la Science Ouverte) in the domain of detecting and qualifying mentions to data sets and software in full text publications (see Bassinet et al. 2023) makes it possible to anticipate a wider dissemination of such technologies at the service of open science based research assessment. Since software is by definition an essential part of research productions in computer science, we should design a flexible way to design an all-in-a-box software dashboard component, based upon the Softcite library[11], that would allow CS institutions or departments to produce, in a comparable way, their own software dashboard (along the properties used/created/shared), but also contribute to increasing the annotated corpus base to improve the corresponding accuracy for our field. We suggest to carry out a proof of concept of such a generic methodology on the basis of two possible corpora of open full text publications: a) the publications of Inria available in HAL, b) the IFIP digital library content and c) the CEUR proceedings, thus covering two transversally relevant datasets for the CS domain.

In parallel to this generalisation of the OS dashboard for software within the CS field, it is essential that a strong communication campaign is launched to systematise the use of good citation practices to software, in particular to advocate the use of the intrinsic identifier SWHID for all software components.

Finally, we need to identify the condition of setting a better background for the documentation and identification of conferences that could serve as a basis for an authority database that would be usable in publication repositories just as what is currently available for journals. Although this is potentially a long-standing item for the community, it is essential that we keep experimenting on this and develop possible proof of concepts. The basis for this could be manyfold: defining an API to DBLP or extension thereof, connecting to CORE or GII-GRIN-SCIE (GGS) Conference Rating (https://scie.lcc.uma.es:8443/), and in doing so, identify how we can build up from existing community platforms such as Lipics[12] and make sure that we are not too tied to conference rankings.

References

Alliez P., Di Cosmo R., Guedj B., Girault A., Hacid M.-S., et al.. Attributing and Referencing (Research) Software: Best Practices and Outlook from Inria. Computing in Science and

[11] https://github.com/softcite/software-mentions

[12] The LIPIcs series, which publishes proceedings of computer science conferences, is part of Dagstuhl Publishing (https://www.dagstuhl.de/dagpub)

Engineering, 2019, pp.1-14. ⟨ 10.1109/MCSE.2019.2949413⟩ . https://hal.science/hal-02135891

Bassinet A., Bracco L., L'Hôte A., Jeangirard E., Lopez P., et al.. Large-scale Machine-Learning analysis of scientific PDF for monitoring the production and the openness of research data and software in France. 2023. https://hal.science/hal-04121339

Bretthauer, D. (2001). Open source software: A history. https://opencommons.uconn.edu/libr_pubs/7/

Canteaut A., M. Angel Fernández, L. Maranget, S. Perin, M. Ricchiuto, et al.. Évaluation des Logiciels. Inria. 2021. https://inria.hal.science/hal-03110723

Ginsparg, P. (2021). Lessons from arXiv's 30 years of information sharing. Nature Reviews Physics, 3(9), 602-603.

Kim, J. Evaluating author name disambiguation for digital libraries: a case of DBLP. Scientometrics 116, 1867–1886 (2018). https://arxiv.org/abs/1806.10540

Kusserow, A., & Groppe, S. (2014). Getting indexed by bibliographic databases in the area of computer science. Open Journal of Web Technologies (OJWT), 1(2), 10-27.

Ley M. (2009). DBLP - Some Lessons Learned. Proc. VLDB Endow. 2(2): 1493-1500. https://dblp.uni-trier.de/xml/docu/dblpxml.pdf

Lin, J., Yu, Y., Zhou, Y., Zhou, Z., & Shi, X. (2020). How many preprints have actually been printed and why: a case study of computer science preprints on arXiv. Scientometrics, 124(1), 555-574.

Vrettas, G., & Sanderson, M. (2014). Conferences vs. journals in computer science. J Assoc Info Sci Technol. https://doi.org/10.1002/asi.23349.

Zuo X, Chen Y, Ohno-Machado L, Xu H. How do we share data in COVID-19 research? A systematic review of COVID-19 datasets in PubMed Central Articles. Brief Bioinform. 2021 Mar 22;22(2):800-811. https://doi.org/10.1093/bib/bbaa331.