Informatics Research Evaluation, Revised Report (Draft 24-10-18)

An Informatics Europe Report

Prepared by the *Research Evaluation Recommendations Panel* of Informatics Europe and the European National Informatics Associations. Editors **Manuel Carro**, Universidad Politécnica de Madrid and IMDEA Software Institute, and **Pekka Orponen**, Aalto University.

Informatics Research Evaluation, Revised Report

October 2024

Published by:

Informatics Europe
Binzmühlestrasse 14/54

8050 Zurich, Switzerland
www.informatics-europe.org
administration@informatics-europe.org

Other Informatics Europe Reports

- *Survey about Diversity and Inclusion Initiatives (2024, Elisabetta Di Nitto, Antinisca Di Marco and the Informatics Europe Diversity & Inclusion Working Group)*
- *Proceedings of the 1st Early Career Researchers Workshop at ECSS 2021 (2021, Elisabetta Di Nitto and Standa Živný)*
- *Bridging the Digital Talent Gap: Towards Successful Industry-University Partnerships (2020, Enrico Nardelli, Cristina Pereira and rapporteurs in Artificial Intelligence, Cyber Security and Software Engineering, with the support of the European Commission, DG CONNECT)*
- *Informatics Education in Europe: Institutions, Degrees, Students, Positions, Salaries — Key Data 2013-2018 (2019, Svetlana Tikhonenko, Cristina Pereira)*
- *Ethical/Social Impact of Informatics as a Study Subject in Informatics University Degree Programs. (2019, Paola Mello, Enrico Nardelli, with contribution from the Working Group Members)*
- *The Wide Role of Informatics at Universities. (2019, Elisabetta Di Nitto, Susan Eisenbach, Inmaculada García Fernández, Eduard Gröller)*
- *Industry Funding for Academic Research in Informatics in Europe. Pilot Study. (2018, Data Collection and Reporting Working Group of Informatics Europe)*
- *Informatics Education in Europe: Institutions, Degrees, Students, Positions, Salaries — Key Data 2012-2017 (2018, Svetlana Tikhonenko, Cristina Pereira)*
- *Informatics Research Evaluation (2018, Research Evaluation Working Group of Informatics Europe)*
- *Informatics for All: The Strategy (2018, Michael E. Caspersen, Judith Gal-Ezer, Andrew McGettrick, Enrico Nardelli. Joint report with ACM Europe)*
- *When Computers Decide: Recommendations on Machine-Learned Automated Decision Making (2018, James Larus, Chris Hankin, Siri Granum Carson, Markus Christen, Silvia Crafa, Oliver Grau, Claude Kirchner, Bran Knowles, Andrew McGettrick, Damian Andrew Tamburri, Hannes Werthner Joint Report with ACM Europe)*
- *Informatics Education in Europe: Are We All In The Same Boat? (2017, The Committee on European Computing Education. Joint report with ACM Europe)*

All reports can be obtained from Informatics Europe at:

www.informatics-europe.org

Executive Summary

Evaluation is an indispensable instrument for improving research quality and impact. To achieve the intended effects, research evaluation should follow established and widely accepted principles, be benchmarked against appropriate criteria, and be sensitive to disciplinary differences.

This report builds on and updates the outcomes of the 2008 and 2018 *Informatics Europe reports on Research Evaluation for Computer Science/Informatics*, while aligning its recommendations with other recent documents on research evaluation, most notably the *CoARA Agreement on Reforming Research Assessment* (CoARA 2022). The report also contains updated analyses and recommendations in four topical areas of concern in Informatics: the responsible use of bibliometrics and credit assignment in contributions, assessing artefacts, Open Science, and interdisciplinary research, together with a discussion on the role of AI in research evaluation.

Our key messages are the following:

1. Informatics is an original discipline that combines aspects of mathematics, science, and engineering. Researcher evaluation must recognise and respect its specificity.
2. A distinctive feature of publication in Informatics is the importance of highly selective conferences. Journals have complementary advantages but do not necessarily carry more prestige. Publication models that couple conferences and journals, where the papers of a conference are published directly in a journal, are a growing trend that may bridge the current gap between these two forms of publishing.
3. Open archives and overlay journals are recent innovations in the Informatics publication culture that offer improved tracking in evaluation.
4. The impact of artefacts such as software, open datasets, and other research products such as trained machine learning models can be as great as publications. The evaluation of such objects, which is now conducted by many conferences, should be encouraged and accepted as an established component of research assessment. Another important indicator of impact is advances that lead to commercial exploitation or adoption by industry or standardisation bodies.
5. Open Science and its research evaluation practices are highly relevant to Informatics. Informatics has played a key enabling role in the Open Science revolution and should remain at its forefront.
6. Numerical measurements (such as citation and publication counts) must never be used as the sole evaluation instrument. They must be filtered through human interpretation, specifically to avoid errors, and complemented by peer review and assessment of outputs other than publications. In particular, numerical measurements must not be used to compare researchers across scientific disciplines, including across subfields of Informatics.
7. In Informatics, the order of authors often holds little significance and varies across subfields. Without clear guidelines, it should not be a factor in researcher evaluation. Instead, authors should be encouraged to clearly state the scope and role of their individual contributions to multi-author works.
8. In assessing institutions, researchers, publications and citations, the use of open research information provided by Open Science infrastructures should be favoured and supported. When using ranking and benchmarking services provided by for-profit companies, respect for open access criteria is mandatory. Journal-based or journal-biased ranking services are inadequate for most of Informatics and must not be used.
9. Any evaluation, especially quantitative, must be based on clear, published criteria. Furthermore, assessment criteria must themselves undergo assessment and revision.

1. Research Evaluation

Evaluation is concomitant with research. The work of researchers is subject to evaluation from the very beginning of their life as such. Research results submitted for publication are subject to a peer review process that scrutinises their scientific quality to determine their worth. The researchers themselves are constantly evaluated during their lifetime: when hired, for promotion, for funding of research proposals, when being considered for specific roles in committees, to be given recognition with awards, etc. Researchers also often act as evaluators of their peers.

By and large, these evaluation processes are internal to the research community and aim to guarantee its internal fairness and integrity through self-regulation. Increasingly, evaluation is also mandated and regulated by exogenous entities and scaled from individuals to entire institutions. In many cases, governments have developed national evaluation standards and processes. Often, their end results are rankings. The main motivation for these efforts is to guarantee that taxpayers' money is spent on research efficiently and leads to societal benefits.

Research evaluation can thus be performed for different goals. It can target a specific piece of research (documented by a single or several artefacts), an individual researcher, a research group, or an organisational unit (e.g., a department). It may even generalise to entire organisations (e.g., universities) or even countries. In any case, the goal of evaluation is to assess some explicit or implicit notion of value, or quality, of research. To achieve the positive objectives of research evaluation, **the specific goals of any such evaluation effort must be clearly stated**, and the way it is conducted must be aligned with these goals. The **evaluation must follow established principles and practical criteria, known and shared by evaluators and researchers**, and **take into account any specificities of the scientific field and area involved.**

Evaluation can have a tremendously positive effect in improving research quality and productivity. It is vital to recognise and support research that can lead to advances in knowledge and impact on society. At the same time, **the effect of following ill-conceived criteria or practices in research evaluation, or misusing metrics and indicators, can have seriously negative long-term effects**. In particular, it may greatly damage the potential of future generations of researchers. Furthermore, **very frequent evaluation can have a detrimental effect on fundamental research**, as the assessment may then accentuate low-risk, short-term developments at the expense of potentially high-gain, long-term work.

This report focuses mainly on the main **principles and criteria that should be followed when individual researchers[1] are evaluated** for their research activity **in the field of Informatics**,[2] addressing the specificities of this area. This subsumes evaluation of a specific piece of research and can to some extent be generalised to departments since their research performance is largely determined by their individuals.

This report reasserts the recommendations in previous Informatics Europe reports on the subject (Informatics Europe 2008, 2018), to which it incorporates a number of observations concerning the topical areas of assessing artefacts, the responsible use of bibliometrics, Open Science and interdisciplinarity research, and the role of AI in research evaluation.

An important development since the publication of the 2018 report has been the convergence of several reports and initiatives on revising research evaluation practices into the *CoARA Agreement on Reforming Research Assessment* (CoARA 2022). The recommendations in this report have been reviewed and updated to align with the CoARA recommendations.

[1] Some aspects of department evaluation are addressed in the publication (Informatics Europe 2013).

[2] Interchangeably "Computer Science" or "Computing".

2. Informatics and Its Specificity

2.1 Characteristics of Informatics

Informatics is a relatively **young science** which is **rapidly evolving** in close connection with technology.

Beyond the two basic and universal research pillars of theory and experimentation, which in Informatics research are often both present in varying proportions, a third important component in Informatics is the creation of **artefacts**, which provide new designs, tools or otherwise improved support for information processing tasks.

Informatics research covers an extremely wide and **methodologically diverse** set of areas: from the development of new computing devices to the mathematical theory of algorithms and complexity, from human factors to big data, machine learning and artificial intelligence, from studies of programmers' productivity to secure encryption methods. Interdisciplinarity, as in the case of bioinformatics, medical informatics, geo-informatics, or cognitive science, brings in even more diversity.

With the continuing digitalisation of society and the emergence of potentially disruptive computational technologies such artificial intelligence and quantum computing, Informatics has an **extremely high societal and economic impact.** Measuring up to these responsibilities requires not only continuous progress in Informatics research, but also putting more emphasis on Informatics education and the influence of Informatics on society, including ethical concerns.

Informatics research, as any other science, must be evaluated according to criteria that take into account its specificity. Universal criteria do not exist to evaluate research quality. This is also true for the different subfields of Informatics. These differences must be taken into account, and the temptation should be resisted to adopt simplistic, one-size-fits-all criteria.

2.2 The Informatics publication culture and its evolution

The publication culture within Informatics differs from most other sciences in the **prominent role played by conference publications**. Many subfields of Informatics have leading conferences with status, visibility, and impact comparable to or higher than their respective leading journals. Conference papers at these meetings undergo a highly selective peer-review process that makes them very competitive and often leads to lower acceptance rates than in the best journals. Conferences in Informatics provide a faster turnaround time than journals to publish research results, get feedback from peers, and build upon it, and also often have higher standards of novelty. These factors are crucial in a rapidly evolving field like Informatics, and as a result, in Informatics, **journals do not necessarily carry more prestige** than conferences. Journal publications are, of course, also important, especially for gathering previous research into an established body free of the space limitations imposed by conferences.

Bridging the dichotomy between conferences and journals, new alternatives are now in place that are changing the publication culture:

- **Coupled conferences and journals**, implemented in different ways: VLDB-style with continuous submission to the journal and presentation of the accepted papers at the conference; ICLP-style with the proceedings of the conference (i.e., all full papers accepted after two rounds of refereeing) being published as a special issue of a standard journal, in this case, TPLP; and its variant ACM PACMPL-style,[3] where a dedicated journal is used to publish proceedings of several

[3] VLDB: "International Journal on Very Large Databases"; ICLP: "International Conference on Logic Programming"; TPLP: "Journal on Theory and Practice of Logic Programming"; PACMPL: "Proceedings of the ACM on Programming Languages."

different, related conferences. These hybrid combinations of conferences and journals are a **promising and growing trend** that combines the advantages of timely publication of conferences with the impact tracking of journals (Dagstuhl 2012).

- **Open Archives** (arXiv, HAL, Zenodo etc.) provide opportunities to publish first versions of papers and protect the intellectual property of new results, at the same time giving online access to all proofs and materials (including data and software) sustaining these results. After possible feedback from peers, improved versions can be submitted to **overlay journals** according to their publication constraints. In this model, reviewers have access to the complete history of the results and can better evaluate their quality and impact.

Books remain specific and important to provide a comprehensive view of topics and contribute to education. Here again, prepublication in an open archive can help in sharing drafts, getting feedback from peers before the official publication, managing versions, etc.

Another specificity of the Informatics publishing culture is that unlike in other sciences such as Physics or Medicine, Informatics does not have a generally adopted convention regarding the semantics of the order in which a publication lists its authors. Thus, in the absence of specific indications, the order should not serve as a factor in individual researchers' evaluation. This issue is discussed more extensively in Section 4 of this report.

3. Research Evaluation for Increased Quality and Impact

The fundamental goal of research evaluation is to assess the quality and impact of research, for the eventual improvement of both. **Quality** is an elusive intrinsic characteristic for which a commonly accepted assessment method, even if imperfect, is peer review by a panel of informed experts. **Impact** is an observable external characteristic that takes many forms and can to some extent be measured by numerical indicators, but even then only with human expert interpretation. Quality is mostly a good predictor of impact, and impact is mostly a good indicator of quality, but the two are not coextensive. The CoARA agreement recommends to "focus research assessment criteria on quality [and] recognise the contributions that advance knowledge and the (potential) impact of research results" (CoARA 2022, p. 3). Assessing impact is often quite hard and, by its own nature, even infeasible in a short timeframe, since the impact of a body of research can be indirect, and it may take years before the eventual impact can be recognised.

3.1 Assessing quality of research

As recognised by the CoARA agreement, "research assessment should rely primarily on qualitative assessment for which peer review is central, supported by responsibly used quantitative indicators where appropriate" and "it is important that peer review processes are designed to meet the fundamental principles of rigour and transparency" (CoARA 2022, p. 5).

With the increasing availability of publicly available bibliometric data, research assessments have been resorting more and more on bibliometric indicators provided by a variety of sources, often even without questioning the sources' soundness and trustability. While we acknowledge the usefulness of quantitative data and bibliometric indicators when used responsibly (see Section 4), we stress the following:

- **The goal of research assessment is primarily to assess quality and impact over quantity** (cf. Friedman & Schneider 2015). Any policy that tends to favour quantity over quality has potentially disruptive effects and would mislead young researchers with very negative long-term effects. Such policies can lead to just focusing on publishing the least publishable increments, and in this respect, some European countries have established evaluation practices with indicators that raise serious concerns. To stress the importance of quality and impact, it is recommended that researcher evaluations focus on a relatively small number of high-quality publications and artefacts, trying to identify also their impact factors such as novelty, supporting artefacts, and impact on other researchers' work.

- **Quantitative data and bibliometric indicators must be interpreted in the specific context of the research being evaluated. They should never constitute the sole evaluation criterion.** Different research areas and even subfields of Informatics have very different characteristics. Even within a homogeneous set, bibliometric indicators only provide a very coarse assessment. For example, although a very high number of citations may indicate a potentially impactful piece of work, it may also indicate a widely referenced survey instead of an original and novel contribution (but which, of course, also has its value). **Human appraisal with respect to established criteria is needed to interpret data and discern quality and impact.** Numbers can only help; they are not a substitute.

- In addition to being established, known, and shared by evaluators and researchers, **assessment criteria must themselves undergo assessment and revision** in order to follow the evolution of science.

Other indicators for quality

Major conferences and scholarly societies in Informatics often grant "best paper awards" to the papers perceived to have the highest value of those accepted at a conference. A further development is the awarding of "most influential paper award" or "test-of-time awards", for papers perceived to have had the most influence in the area in the last *N* years. These distinguishing, peer-reviewed awards should be taken into account in evaluations.

3.2 Assessing impact of research

The impact of research can be assessed along many different dimensions. First, one can distinguish between external and community impact. The **external impact** of research measures its effect on society at large. The new GenAI techniques, for example, will potentially have an enormous impact on society. Likewise, the invention of a new secure protocol may lead to the development of more secure networks on which society relies, and a new automated development environment can improve the industry's productivity. **Community impact** refers to the impact of one's research on other researchers. This means that other researchers can build on top of one's research. Evaluation often tries to capture this kind of impact by e.g. citation indices.

An orthogonal dimension of impact concerns the outcomes of research. An outcome can have an impact because it advances **knowledge** in a given area, or because it advances **practice**. New *knowledge* can be created by theoretical research, similar to mathematics. A typical example is research on algorithms and computational complexity. New knowledge can also be generated by research that pursues empirical studies using tools from Informatics, sometimes also addressing Informatics itself. Typical examples are discovering properties of biological systems or physical materials using machine learning techniques on large datasets, or the automatic extraction of interaction patterns among software designers from open-source repositories. Advances in *practice*, on the other hand, refer to research that aims at developing new tools or methods to solve a given problem. An example of this would be a technique that improves the scalability of a technique to perform program verification.

Indicators for external impact

Technological innovation is vitally important for the general health of Informatics because it underlines the relevance of our discipline to society as a key driver of economic growth. Indeed advances that lead to commercial exploitation, spin-off activities, or adoption by industry (software licences granted to industry) or standards bodies, are highly valued forms of impact and excellent indicators of the applicability of research.

Indicators for community impact

In many areas of Informatics, software competitions are regularly organised to assess progress in tools (e.g., SAT solvers or learning tools). Such competitions are typically highly appreciated and contribute to increasing the quality of the tools, and success in them is a could be taken as an indicator of community impact.

4. Responsible use of indicators and credit assignment in contributions

The most common method of evaluating research impact is through citations. Counting citations is often viewed as a way to gauge a paper's influence within the research community. This approach has been extended into various bibliometric indicators, such as the h-index for researchers and impact factor for publication venues. While bibliometrics has gained popularity under the pressure of external evaluation, and publishing companies readily added them to their offer to the community, it is increasingly used, often tacitly, in internal evaluations that claim to rely on peer review. However, bibliometrics must always be complemented by qualitative peer review focusing on the research content and approaches measuring contributions beyond publications.

Two key principles are essential:

- Publication counts, whether weighted or not, must not be used to evaluate research value. While they may measure productivity, they are not indicators of research impact or quality.
- Numerical impact measurements, such as citation counts, must not be used in isolation. They must be interpreted by humans, taking into account the potential for errors and unethical practices like gaming the system. Additionally, these metrics should not be used to compare researchers across different fields, nor within subfields of Informatics, due to varying publication and citation cultures.

This is in line with the CoARA recommendations to "Base research assessment primarily on qualitative evaluation for which peer review is central, supported by responsible use of quantitative indicators."

It is also crucial to assess the individual contributions in multi-author publications. In some disciplines, the order of authorship reflects the level of contribution, but this is rarely the case in Informatics, even though Informatics researchers also face the necessity to produce publications with 'primary authorship' roles, particularly for publication-based PhDs. In these contexts, even the percentages of authors' contributions are sometimes assessed and documented. To address this, we recommend that researchers clearly specify each author's contribution in their publications and other scientific artefacts to facilitate fair evaluation. The CRediT taxonomy (credit.niso.org) provides a structured way to document contributions across the Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project Administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. In line with CRediT guidelines, we suggest that all contributions be listed, whether from authors or individuals acknowledged elsewhere. Individual contributors can be assigned multiple roles, and a single role can be assigned to multiple contributors. Where multiple individuals serve in the same role, the degree of contribution can optionally be specified as 'lead,' 'equal,' or 'supporting.' (https://credit.niso.org/implementing-credit/, 2024-07-29). Transparency in this regard is crucial for both evaluators and the researchers themselves.

When harvesting bibliometric data, public databases like ArXiv, DBLP, HAL, and Zenodo are valuable because they offer low-noise data on publications and often provide full-text access. DBLP makes available very useful accounts of the Informatics publications of individual researchers, while ICORE (https://www.core.edu.au/icore-portal) and CSRankings (https://csrankings.org) provide resources for ranking conferences similar to those furnished by Clarivate JCR and Scopus SJR for journals. The Clarivate bibliometric tools however are inadequate for much of Informatics because they mostly address journal publications, while Scopus provides also some coverage of conferences but is still journal-biased. Google Scholar is more inclusive of conferences than these, and thus more aligned with the Informatics publication culture, but suffers from high data noise.

While quantitative bibliometric indicators can provide insights into global trends at institutional levels, they should play a limited role in assessing individual researchers or projects. Furthermore, prominent indicators such as h-index are prone to manipulation, such as by using citation rings or paid authorship positions.

Evaluating individual contributions requires contextualising them within the standards and best practices of the specific field, which can only be effectively done by peers with domain expertise. Although peer review is inherently subjective, collective assessments by expert panels provide robustness to the process. Indicators can play a supportive role in this context. The value of quantitative indicators is highly questionable unless they are contextualised and explained appropriately.

5. Assessing artefacts

In this section, we explore the evaluation of research using outputs that go beyond traditional publications, such as software and other outcomes. While our primary focus will be on software, due to resource limitations during the preparation of this report, many of the points discussed are likely applicable to other research outputs. These may include datasets, collections of theorems to test automated theorem provers, formal specifications, proofs developed with interactive provers, complex build environments, RTL definitions of processors, and similar results.

5.1 Software artefacts

"Software artefacts" submitted alongside papers (Winter 2022) are now routinely evaluated in CS conferences. These artefacts serve as evidence to support the claims made in the papers by enabling the reproduction of the results therein. They should include detailed installation instructions or be provided as a container or virtual machine image for ease of use. These artefacts are immutable snapshots, ideally stored in repositories that offer long-term preservation (e.g., Zenodo, FigShare, Software Heritage). Quality "badges" may be awarded based on factors such as ease of installation, functionality, and how effectively the artefacts reproduce the results presented in the paper. Many CS conferences and societies provide detailed guidelines for this evaluation process (ACM, 2020; Rous, 2017).

It is important to note that reproducibility is not only relevant for conferences, but is also a fundamental pillar of Open Science, as defined by UNESCO (UNESCO 2021 a, UNESCO 2021 b). Reproducibility issues in sub-areas such as AI and machine learning are impacting other scientific disciplines that rely on their outcomes (Ball 2023; Baker 2016). Of course, no subfield of computer science is free from these negative effects, and therefore the growing adoption of artefact evaluation is laudable.

However, the academic value of these artefacts is tied to that of the associated paper. Their evaluation primarily confirms that the artefact aligns with the expectations set by the paper. As a result, their usefulness as independent assessment of research outcomes is not clear. Additionally, these artefacts are typically not expected to evolve over time by incorporating additional functionality or improving performance.

Efforts to ensure that artefact evaluation is as fair and rigorous as possible are ongoing. For instance, the International Association for Cryptologic Research (IACR) has established a task force to standardize guidelines and practices across the conferences it sponsors, which until now have largely been determined by conference chairs and steering committees. Additionally, the USENIX Security Conference will introduce a new topic in 2025, "Meta-Science in Security and Privacy," which will include reports and studies on the replicability—or failure to replicate—previous research results (USENIX 2025).

On the other hand, some software packages / systems are designed with the expectation of continuous evolution and improvement, incorporating additional functionalities over time. These packages often result from long-term collaborative efforts involving multiple researchers and developers. Some examples (among many others) include systems like Coq and CVC, which are extensively used in their respective fields.

These systems are often not associated with a single paper and, as such, cannot be classified as usual "artefacts." The techniques and ideas they implement may be described across multiple papers, but some key components of these systems may not be documented in any publication. Yet, they can represent significant research contributions and demonstrate an innovative use of technology. This raises the question of how to properly recognize and reward the effort involved in developing and maintaining such systems.

5.2 Why evaluate software systems

A strong argument for evaluating software systems as independent, standalone research contributions is that their authors deserve recognition for the significant resources and effort invested in their development and maintenance. These systems play a crucial role in advancing science. This reasoning can also be applied to other types of research outputs, such as those (partially) listed before. Therefore, we understand that to assess impact, research software and datasets are as important as publications.

Large, robust software systems that implement complex techniques are continuously used by researchers other than their original authors. Systems like Coq and CVC, previously mentioned, are examples of influential software packages. In the domain of datasets, "Thousands of Problems for Theorem Provers" (TPTP) serves as a key resource. TPTP is regularly used by theorem prover developers to evaluate the efficiency of their systems and make fair comparisons. Although the "deep knowledge" gained from collecting and categorising such examples may seem modest, the contribution to advancing the field of theorem proving must be acknowledged.

By releasing robust, well-engineered, and useful systems (or curated datasets), authors provide essential building blocks upon which more advanced research can be conducted. This creates a global benefit for both academia and society that hopefully pays off for the resources invested in the design, implementation, and maintenance of these systems.

On the other hand, well-engineered software systems and datasets are also an integral part of the Open Science ecosystem, which includes "open research data, open software, source code, and open hardware." These resources lower the barrier for technologically inclined, but not-yet-proficient-enough individuals, such as a first-year computer science student who may lack the skills to install complex software with dependencies and intricate configurations, to explore and better understand the outcomes of research funded by society.

5.3 Caveats

Developing a software system according to sound software engineering practices requires significantly more effort than making an artefact available.[4] However, researchers often lack the funding to hire skilled software engineers, programmers, and, when needed, data scientists or other specialised professionals. As a result, they must allocate their own time and resources, diverting focus from their primary research. Furthermore, maintaining such systems demands substantial resources.

In the current state of affairs, there is little incentive to develop well-rounded, high-quality software systems. This creates a vicious cycle: because software systems are not inherently valued, minimal effort is invested in demonstrating the feasibility of certain techniques, resulting in software of only average quality. The subpar quality, in turn, diminishes interest in using the software, as it may be difficult to use, leading to projects being abandoned for lack of impact. From a broader perspective, this represents a misuse of resources that ultimately harms both the academic community and society at large. Rather than rewarding authors who maintain and improve software systems, the current research evaluation system penalizes them, as the time and resources dedicated to these systems often come at the expense of traditional paper-based research, which is commonly recognized.

This may be partly because there are specific characteristics – particularly from a software engineering perspective – that need to be assessed in software but are not relevant to papers. An incomplete list of key aspects to consider in software evaluation includes:

- **Longevity and evolution:** Software systems are long-lived and undergo complex evolution. Properly accounting for this and distinguishing between genuine research contributions and routine engineering updates can be challenging.
- **Authorship:** Assigning ownership to specific functionalities or implementations of algorithms can be difficult. However, this challenge is not unique to software: it also occurs with multi-author papers. A suggestion for handling authorship in papers is provided elsewhere in this report, and a practical approach for software systems has been implemented by INRIA (discussed later).
- **Measuring impact:** Metrics such as the number of users or downloads, often used as proxies for a software's relevance or usefulness, can be difficult to estimate accurately or may be artificially inflated. Furthermore, these metrics may not always reflect the true impact in the research community or quality of the software.
- **Temporal context:** Evaluating a software system or its characteristics long after its initial development can be problematic, as the motivations and significance of the software at the time of its creation may be hard to fully appreciate.
- **Industry collaborations:** In cases of industry collaboration, software may be subject to usage restrictions, limitations on functionality, or restricted access to source code, making it difficult to conduct a comprehensive evaluation of its merits.

[4] This is not meant to minimise the challenges of preparing artefacts, but rather to highlight the clear reality that as systems increase in size and complexity, so do the difficulties of development and maintenance.

5.4 Recommendations

There is currently no universally accepted policy for evaluating independent software systems. Even in the case of papers, where established quality indicators exist (the relevance of conferences or journals, community impact through e.g. citations, and expert evaluations), different institutions and countries apply varying approaches. Given the challenges outlined above, we are unable to provide concrete recommendations for evaluating software systems at this time. Instead, we offer some general recommendations and considerations that should be taken into account to make evaluations as fair as possible. Additionally, a comprehensive framework would need to address not only software but also datasets, open hardware definitions, and other related outputs.

- **Avoid "bean counting":** Evaluations should not rely solely on quantitative metrics such as lines of code, number of projects, or downloads. While these metrics can provide useful insights, they should not be the primary or dominant source of information.
- **No penalty for lack of software releases:** Researchers with strong traditional paper-based research records should not be penalised for not releasing software. While this is a reasonable principle, its implementation in environments with limited positions and promotions remains a challenge.
- **Reward only research-driven software advancements:** Only software developments that generate new knowledge or research should be rewarded. Improvements demonstrating engineering skill but lacking a research component should not count toward research evaluation. For instance, maintaining a large system, although resource-intensive, should not be considered part of research evaluation.
- **Long-term software projects deserve recognition:** Software systems that have been in use and actively maintained over decades should be considered particularly significant achievements.
- **Consider the researcher’s role:** The role of a researcher in a software project should be factored into the evaluation, similar to the way author roles are classified in research papers, as discussed elsewhere in this report.
- **Short contributions:** Short contributions to large projects should generally not count toward research evaluation, unless they involve adding a key component or capability that significantly enhances the project. An exception to this exception is when the contribution involves merely adapting existing code, which is primarily an engineering task.
- **Cumulative value of successful artefacts:** A continuous record of successfully evaluated artefacts, even if they are not part of a large system, can be seen as having a cumulative value greater than the sum of individual papers or artefacts. This demonstrates a sustained commitment to validating the practicality of research ideas.
- **Value of contributions to components:** Contributions to components of larger systems (e.g., libraries, parts of compilers) should be valued independently, as they provide essential functionality for other researchers or to perform advanced training (at the level of PhD, for example). For example, contributions to systems like LLVM or Coq may be more valuable than creating a standalone toy compiler.
- **Weigh software releases appropriately:** Not all public releases of software should be weighted equally. Evaluation should focus on the changes introduced between releases. Alternatively, the latest version could be the focus of the evaluation, rather than treating each release with the same importance.
- **Citing software:** Software used in research must be properly cited. While it is common to cite

the paper where the software is presented, if a specific version of the software has been used in some research work)especially if it includes functionality not covered by any paper), the contributors to that version should be appropriately credited. This requires the use of permanent, immutable identifiers associated with the software version and with its authors.

- **Public software for public funding:** Software developed with public funding should be openly accessible to ensure transparency and accountability, especially when it is subject to evaluation.
- **Specific software evaluation criteria:** There is a need for distinct criteria to evaluate software. According to an internal study by the software subcommittee of the French National Committee for Open Science, the FAIR principles used for datasets are not easily adaptable to software, and specific metrics must be devised for effective evaluation.

The protocol established by INRIA (Canteaut 2021) for evaluating software systems may be one of the better defined ones. It involves a self-evaluation of software contributions, which is then reviewed by a committee. The self-assessment form and evaluation guidelines could serve as a valuable starting point for replication or adoption in other institutions and agencies. Along these lines, the French Ministry of Research and Education annually gives awards to outstanding software developed for education or research purposes or created in a scientific context; expanding such awards to a European level could be beneficial. Additionally, and in a related category, Italian universities are evaluated based on societal impact, a criterion that could also encompass well-maintained academic software systems.

6. Open Science

In the past ten years, Open Science has become one of the hot topics introduced in the context of the scholarly domain at large. Open Science concerns a plethora of different dimensions which aims at incentivising quality in science and recognising the diversity of research outputs, activities and missions. However, Open Science is not granted per se, but it is the result of a community effort that pushes on developing an appropriate policy environment to host it in the scholarly ecosystem. Part of this environment concerns also “research and researcher evaluation and assessment practices” (UNESCO 2021) .

While discussing and debating about how research is conducted, this “construct” (UNESCO 2021) also concerns how research should be evaluated. UNESCO pushes a lot on the importance of rewarding, in assessment processes, good open science practices – a necessary incentive for enabling the implementation and operationalisation of Open Science. In addition, it fosters a revision of the current research assessment practices to align them with the principles of Open Science. Indeed, the overall direction suggested is to build assessment processes on existing and well-known practices and guidelines – such as the San Francisco Declaration on Research Assessment (DORA 2012), the Leiden Manifesto (Hick et al. 2015), and the Coalition for Advancing Research Assessment (CoARA 2022). The main action points of these guidelines are:

- avoid using metrics developed for measuring a kind of entity (e.g. Journal Impact Factor for journals) as a proxy measure for assessing other kinds of entities (e.g. researchers);
- explicit the criteria used in the assessment, and prefer peer-review evaluation, supported by quantitative indicators when appropriate;
- recognise the diversity of research contributions in assessment exercises (datasets, databases, software, and other artefacts) in addition to classic print-like publications (journal article, book, book chapters, etc.);
- apply openness and transparency when providing data and methods used in the assessment exercise;

- enable anyone to verify both the data and the analysis;
- fight against metrics manipulations and incorrect uses;
- consider the variability of different types of research output and subject areas when comparing an entity against another entity (e.g. researchers);
- focus on the qualitative judgement of research outputs when assessing researchers.

Several of the aforementioned points would not be implementable without the availability of appropriate Open Science infrastructures such as open bibliographic and citation databases and metadata repositories, institutional current research information systems, open bibliometrics and scientometrics systems for assessing and analysing scientific domains. Indeed, UNESCO insists that investing in Open Science infrastructures and services is key and that the scholarly community should retain control and ownership over these infrastructures, which are a crucial means for providing the data used for devising metrics and indicators that may be used to support peer-review assessment: research information.

Research information refers to all the metadata related to the conduct and communication of research, such as bibliographic metadata of research outcomes (articles, software, datasets, methodologies, etc.), information on funding and grants, and information on organisations and research contributors. Several initiatives, such as the Barcelona Declaration (2024), and Open Science infrastructures push for making research information openly available and freely reusable for the scholarly community since various activities are characterised by using such information, including research assessment. Among these infrastructures, OpenCitations (https://opencitations.net), OpenAIRE (https://openaire.eu), Software Heritage (https://www.softwareheritage.org), and DBLP (https://dblp.org) also provide information about software and other Informatics-related artefacts in addition to classic publications in journals and conference proceedings. Instead, other infrastructures, such as PREreview (https://prereview.org), help in addressing other assessment-related tasks in a very transparent way, e.g. by enabling applying open peer review evaluation practices. These open reviewing processes permit the disclosure of the identity of the reviewers, the publication of reviews and, thus, the recognition of the effort that scholars put in reviewing and the possibility for a broader community to provide comments and participate in the assessment process (UNESCO 2021).

Supporting these kinds of Open Science and community-guided infrastructures is key to maintain an environment enabling transparent assessment workflows. Indeed, there is an urgent need to make research information to advance responsible research assessment and operationalise Open Science and promote unbiased, transparent and high-quality decision-making. The Open Science Career Assessment Matrix (OS-CAM) (European Commission 2017) represents a possible, practical move towards a more comprehensive approach to evaluating researchers through the lens of Open Science. In addition, other templates for Open Science assessment that involve Informatics as a domain of application have been studied in the context of specific research projects, such as GraspOS (https://graspos.eu/).

Informatics should acknowledge Open Science practices in its research evaluation. Informatics thus also has a prominent role to play in the adoption and development of the Open Science approaches and infrastructures, and its support is key to keep them sustainable in the long term.

7. Interdisciplinary Research

Informatics has for a long time been an important support provider for research in other fields, in the form of software tools for modelling, optimisation and visualisation, data management and analysis, etc. In an interdisciplinary collaboration this kind of work is often considered mundane by the companion area partners, and not really computing research by the Informatics community. There are also more ambitious collaborations, where the Informatics contribution is an integral element of the research agenda, and pursuing it requires both competence in the companion area and an ability to adapt existing and develop new methods in Informatics for the needs of the specific collaboration.[5] Even in these cases, however, credit is commonly assigned along disciplinary lines, so that recognition goes primarily to the companion area, and also on the Informatics side the computational methods contribution is considered as “an application”.

An added challenge is that in assessments that are based mechanically on publication venues and/or indicators, Informatics contributions to other areas than core Informatics are easily overlooked, because the relevant venues are either not indexed at all in Informatics databases, or are considered as “application area outlets”, no matter how prestigious in the companion area.

The challenge of overcoming disciplinary barriers in the assessment of integrative interdisciplinary research is a widely recognised problem (e.g. McLeish & Strang 2016). Some key guidelines for responsible evaluation of Informatics research in this context are:

- Recognise the value of the integrated research in its own terms, not as an “application” of Informatics.
- Assess (i) the depth of the integration and (ii) the novelty and significance of the Informatics contribution to the totality of the work.
- Be wary of numerical indicators and “top venue” lists oriented towards assessing Informatics disciplinary work.

8. The Role of AI in Research Evaluations

The role of AI influences research evaluation in two ways: first, how to handle AI being used in research, and second, how AI can be used to perform or assist in carrying out research evaluations. While the first point is heavily discussed and prominently addressed in this document, much less discussion exists regarding the second point. We believe it should also be addressed at this time.

Whenever a task is tedious, the question arises of how it can be automated or at least made easier by delegating parts to an automated process. A good example in the context of research evaluations is the h-index, which is widely used but also heavily criticised. Its advantage lies in the ease with which it can be obtained. It is much harder to judge the content of publications than simply to count them and their citation numbers. Until now, there was no easy way to delegate the refereeing of papers to an automated process. This has changed with the advent of Large Language Models (LLMs). It is certainly tempting to use such models to evaluate the output of researchers or entire departments (or even universities). At the time of writing this position paper, the generally available LLMs are hesitant to write entire evaluations. For example, Gemini primarily describes the process of writing an evaluation but does not provide one. ChatGPT can write an evaluation but is very cautious in comparing different computer science departments. Moreover, ChatGPT bases its judgments on citation numbers and very vague identification of key strengths.

Nevertheless, we can safely assume that LLMs can, in principle, be used for research evaluations and that they will be used for this purpose in the future, if not already today. The question for us is to understand the potential dangers and benefits of this development.

As with any other use of AI, some might call for the restriction or outright ban of using AI methods in evaluations that form the basis of important decisions, such as funding large projects or tenure

[5] Notable examples of the latter kind of collaboration are e.g. quantum computing, and the work on computational protein design that was awarded the 2024 Nobel Prize in Chemistry.

decisions. It is short-sighted to ban AI altogether: it might actually help in the decision-making process when used with discretion, and a ban cannot be enforced anyway. Rather than a ban, we propose policies for the responsible use of such technologies, if they are used at all:

- The use of generative AI in research evaluations should be communicated openly, including detailed logs of the communications and the way the information was used in the decision-making.
- Decisions should still be made by human experts, and the use of AI should be restricted to the lower levels of the decision process.
- Efforts should be made to verify critical data obtained from a computer.
- The use of AI must not be used to reduce the number of human experts in decision panels and their responsibility on the outcome of the panel.

The use of AI is a moving target, and today we are just beginning to harness its power. While AI has been instrumental in many areas for some time, its application in decision-making is still in its early stages. Research evaluations are among these areas, and we believe that further discussions in the near future are important as the field continues to evolve. Therefore, our conclusions at this time can only be considered preliminary.

Acknowledgements

We would like to acknowledge the input of a number of researchers who were generous enough to devote some of their time to answer our questions: Erwan Bousse, Carlos Esteban Budde, Karine Even-Mendoza, Hadar Frenkel, Arnd Hartmanns, Tobias Kappé, Raphaël Monat, Benoit Montagu, Marius Muench, Roberto Natella, Michael Rawson, Robert Ricci, Markku-Juhani Saarinen, Jesús Sánchez Cuadrado, Alexander Serebrenik, Caleb Stanford, Quentin Stiévenart.

References

(ACM 2020) Artifact Review and Badging Version 1.1. ACM 2020, https://www.acm.org/publications/policies/artifact-review-and-badging-current

(Baker 2016) M. Baker, 1,500 scientists lift the lid on reproducibility. Nature 533, 452-454 (2016). https://doi.org/10.1038/533452a

(Ball 2023) P. Ball, Is AI leading to a reproducibility crisis in science? Nature 624, 22-25 (2023), https://doi.org/10.1038/d41586-023-03817-6

(Canteaut 2021) A. Canteaut, et al. "Software Evaluation (Research Report)" Inria (2021). https://inria.hal.science/hal-03110728/document

(CoARA 2022) Coalition for Advancing Research Assessment. (2022). Agreement on Reforming Research Assessment. [Policy]. https://coara.eu/agreement/the-agreement-full-text/

(Dagstuhl 2012) Publication Culture in Computing Research -- Position Papers. Dagstuhl Perspective Workshop 12452, 2012. Eds. K. Mehlhorn, M. Y. Vardi, M. Herbstritt.
https://doi.org/0.4230/DagRep.2.11.20

(DORA 2013) San Francisco Declaration on Research Assessment. 2013. https://sfdora.org/read/

(DORI 2024) Barcelona Declaration on Open Research Information. 2024. https://barcelona-declaration.org/

(European Commission, 2017) European Commission. Directorate General for Research and Innovation. (2017). Evaluation of research careers fully acknowledging Open Science practices: Rewards, incentives and/or recognition for researchers practising Open Science. [Report]. Publications Office. https://doi.org/10.2777/75255

(Friedman & Schneider 2015) B. Friedman and F.B. Schneider, Incentivizing Quality and Impact: Evaluating Scholarship in Hiring, Tenure, and Promotion. CRA Best Practice Memo, February 2015. https://www.cra.org/cra/wp-content/uploads/sites/10/2016/02/BP_Memo.pdf

(Hicks et al. 2015) Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). Bibliometrics: The Leiden Manifesto for research metrics. Nature, 520, 429–431. https://doi.org/10.1038/520429a

(Informatics Europe 2008) Research Evaluation for Computer Science, Informatics Europe Report, 2008. Eds. B. Meyer, C. Choppy, J. van Leeuwen and J. Staunstrup. https://www.informatics-europe.org/services/publications/reports.html

(Informatics Europe 2013) Protocol for Research Assessment in Informatics, Computer Science and IT Departments and Research Institutes. Informatics Europe Report, 2013. Ed. Manfred Nagl. https://www.informatics-europe.org/services/publications/reports.html

(Informatics Europe 2018) Informatics Research Evaluation, Informatics Europe Report, 2018. Eds. Floriana Esposito, Carlo Ghezzi, Manuel Hermenegildo, Helene Kirchner and Luke Ong. https://www.informatics-europe.org/services/publications/reports.html

(McLeish & Strang 2016) T. McLeish and V. Strang, Evaluating interdisciplinary research: the elephant in the peer-reviewers' room. Nature Palgrave Communications 2, 16055 (2016). https://doi.org/10.1057/palcomms.2016.55

(Rous 2017) B. Rous. The ACM Task Force on Data, Software, and Reproducibility in Publication, 2017. https://www.acm.org/publications/task-force-on-data-software-and-reproducibility

(UNESCO 2021a), UNESCO Recommendation on Open Science, 2021. https://www.unesco.org/en/open-science

(UNESCO 2021) UNESCO. (2021). UNESCO Recommendation on Open Science (Programme and Meeting Document SC-PCB-SPP/2021/OS/UROS; p. 36). https://unesdoc.unesco.org/ark:/48223/pf0000379949

(USENIX 2025) USENIX Security '25 Call for Papers. https://www.usenix.org/conference/usenixsecurity25/call-for-papers

(Winter et al. 2022) Winter et al. "A retrospective study of one decade of artifact evaluations." Proc. ESEC/FSE 2022. https://doi.org/10.1145/3540250.35491