INFORMATICS EUROPE

INFORMATICS RESEARCH EVALUATION

Floriana Esposito
Carlo Ghezzi
Manuel Hermenegildo
Helene Kirchner
Luke Ong

Informatics Research Evaluation

An Informatics Europe Report

Prepared by the *Research Evaluation Working Group* of Informatics Europe:

- **Floriana Esposito**, Università degli studi di Bari Aldo Moro, Italy
- **Carlo Ghezzi**, Politecnico di Milano, Italy
- **Manuel Hermenegildo**, IMDEA Software Institute and Universidad Politécnica de Madrid, Spain
- **Helene Kirchner**, Inria, France
- **Luke Ong**, University of Oxford, United Kingdom

Informatics Research Evaluation

March 2018

Published by:

Informatics Europe
Binzmühlestrasse 14/54

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

Other Informatics Europe Reports

- *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).
- *Informatics Education in Europe: Institutions, Degrees, Students, Positions, Salaries. Key Data 2011-2016* (2017, Cristina Pereira, Svetlana Tikhonenko).
- *Informatics in the Future: Proceedings of the 11th European Computer Science Summit (ECSS 2015), Vienna, October 2015* (2017, eds. Hannes Werthner and Frank van Harmelen, Springer Open).
- *Informatics Education in Europe: Institutions, Degrees, Students, Positions, Salaries. Key Data 2010-2015* (2016, Cristina Pereira).
- *Informatics Education in Europe: Institutions, Degrees, Students, Positions, Salaries. Key Data 2009-2014* (2015, Cristina Pereira).
- *Informatics Education in Europe: Institutions, Degrees, Students, Positions, Salaries. Key Data 2008-2013* (2014, Cristina Pereira, Bertrand Meyer, Enrico Nardelli, Hannes Werthner).

All these reports and others can be downloaded at:

www.informatics-europe.org

Executive Summary

Evaluation can be highly effective in improving research quality and productivity. To achieve the intended effects, research evaluation should follow established principles, benchmarked against appropriate criteria, and sensitive to disciplinary differences.

This report confirms the findings of the 2008 report on Research Evaluation for Computer Science, while incorporating recent developments.

1. Informatics is an original discipline combining mathematics, science, and engineering. Researcher evaluation must adapt to 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. To assess impact, artifacts such as software can be as important as publications. The evaluation of such artifacts, which is now performed by many conferences (often in the form of software competitions), should be encouraged and accepted as a standard component of research assessment. Another important indicator of impact are advances that lead to commercial exploitation or adoption by industry or standard 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. The order in which a publication in Informatics lists authors is generally not significant and differs across sub-fields. In the absence of specific indications, it should not serve as a factor in the evaluation of researchers.

8. In assessing publications and citations, the use of public archives should be favored. When using ranking and benchmarking services provided by for-profit companies, the respect of 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 innate in research. Researchers know that their work is subject to evaluation from the very beginning of their life as researchers. Research results submitted for publication are subject to a peer review process that scrutinises their scientific quality. Researchers are constantly evaluated during their lifetime: when hired, for promotion, for funding of research proposals, for assignment to specific roles in committees, for recognition through awards,… Researchers also often act as evaluators of their peers.
By and large, these innate evaluation processes are internal to the research community, and aim to guarantee its internal fairness and integrity through self-regulation. Increasingly, evaluation is mandated and regulated by exogenous processes and entities, and scaled from individuals to entire institutions. In many cases, governments have developed national evaluation standards and launched evaluation processes. Often, the end result of these are rankings. The main motivation of these efforts is to guarantee that tax payers' money is spent in research wisely 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 artifacts), an individual researcher, a research group, or an organizational unit (e.g., a department). It may even generalize to entire organizations (e.g., universities) or even countries. In any case, evaluation is performed 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 formulated upfront**, and the way it is conducted aligned with these goals. The **evaluation should 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 recognize and support research that can lead to advances in knowledge and impact on society. At the same time, **the effect of following wrong criteria or practices in research evaluation can have seriously negative long-term effects**. In particular, it may greatly damage the potential of future generations of researchers.

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 often be generalized to departments, since their research performance is largely determined by their individuals.

This report confirms the findings of the 2008 Informatics Europe report on the subject[3] and at the same time incorporates a number of new observations concerning the growing emphasis on collaborative, transparent, reproducible and accessible research.

These recommendations are also consistent with a recent statement of three national Academies (Académie des Sciences, Leopoldina, and Royal Society) on good practice in the evaluation of researchers and research programmes, that also provides recommendations on evaluator selection, overload, and training.

[1] Some aspects of department evaluation are addressed in the Informatics Europe publication "Protocol for research assessment in Informatics, Computer Science and IT Departments and Research Institutes." (2013, Informatics Europe, ed. Manfred Nagl).

[2] We use the term "Informatics" (more frequently used in Europe) with the same semantics as "Computer Science" (more frequently used in the US).

[3] Research Evaluation for Computer Science, Informatics Europe Report, 2008. Eds. Bertrand Meyer, Christine Choppy, Jan van Leeuwen and Jørgen Staunstrup.

2. Informatics and its specificity

Characteristics of Informatics

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

It is an original discipline with roots in mathematics, science, and engineering. In addition, Informatics is **pervasive** and is affecting the way research is traditionally done in many areas, bringing new approaches and perspectives. It results in new **interdisciplinary** research fields in which researchers from traditionally distinct areas collaborate. Interdisciplinary research is notoriously hard to assess and may result in undervaluing individual contributions.

The two basic and universal research paradigms (theory and experimentation) are equally prominent in Informatics, and are often both present in varying proportions in most research efforts. Very often, the outcome of Informatics research is the creation of new **artifacts,** which solve new problems or perform better than previous solutions.

Informatics research covers an extremely wide range and conceptually **diverse** set of aspects: from development of new computing devices to mathematical theory of algorithmic complexity, from human factors to automatic learning from big data, 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 development of digital technologies, informatics has a **high societal and economic impact.** Mastering our digital world and its evolution requires continuous progress of research and training in informatics. This is also true in other disciplines due to the increasing role of informatics in them.

Informatics research, as for 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 different subfields. These differences must be taken into account and the temptation should be resisted to adopt simplistic universal criteria.

The Informatics publication culture and its evolution

The publication culture within Informatics differs from 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 undergo a highly selective peer-review process, that makes them very competitive and often leads to lower acceptance rates than the best journals. They are more timely (shorter time to publish research results) and offer greater opportunity to get timely feedback from peers. They 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, without the space limitations imposed by conferences, which can make it hard to provide a thorough account of a body of research.

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, where a dedicated journal is used to publish proceedings of several different, related conferences. These tight combinations of conferences and journals **offer a promising and growing avenue** that combines the advantages of timely publication of conferences with the impact tracking of journals.[4]

- **Open Archives** (like HAL, ArXiv, etc.) provide opportunities to publish first versions and protect intellectual property of new results, 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, the 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, publication 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. In the absence of specific indications, the order should not serve as a factor in individual researchers' evaluation.

[4] Conferences vs. Journals in CS, what to do? Dagstuhl 12452: Publication Culture in Computing Research -- Position Papers, 2012, Dagstuhl.

3. How to evaluate the impact of research?

The ultimate goal of evaluation is to assess research impact. Impact, however, is an elusive concept and many possible dimensions contribute to it. First, we might distinguish between internal vs. external impact. The **external impact** of research measures its effect on society at large. For example, the invention of a new secure protocol may lead to the development of more secure networks on which society relies. Likewise, a new automated development environment can improve industry's productivity. Assessing external impact is often quite hard, and even unfeasible in a relatively short time frame. The impact of one's research can be indirect, and it may take years before it can be traced back to it. **Internal impact** refers to impact of one's research on other researchers. This means that other researchers can build on top of one's research. This kind of impact is what evaluation often tries to capture.

Another possible characterization of impact concerns the outcome of research. The outcome can have an impact since 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 algorithmic computational complexity. New knowledge can also be generated by research that performs empirical studies on the artificial world created by Informatics. Typical examples are discoveries of biological properties by applying learning algorithms to large data sets, or the automatic extraction of interaction patterns among software designers from open-source repositories.

Advances in practice refer instead to research that aims at developing new tools or methods to solve a given problem. For example, consider a technique to improve scalability of a technique to perform program verification.

Elements for evaluating impact

Bibliometrics

The most common approach to evaluating (internal) impact is through citations. Counting citations for a given paper is viewed as a metric to assess that paper's impact within the research community. This indicator is often used as an objective assessment of the quality of a piece of research described by a paper. The concept has been generalized to a whole set of bibliometric indicators that are used to evaluate researchers (e.g., through the h-index) and publication venues (through impact factor). Bibliometrics has mainly become popular under the pressure of external evaluations, but is also increasingly used (often tacitly) by internal evaluations that officially adopt a peer review scheme.

The main results of the already cited 2008 Report concerning bibliometrics are still valid and are worth restating here:

- **Numerical impact measurements, such as citation counts, have their place but must never be used as the sole source of evaluation.** Any use of these techniques must be subjected to the filter of human interpretation, in particular to avoid the many possible sources of errors. It has also been observed that they induce unethical behaviours and gaming. Bibliometrics **must be complemented** by peer review, and by attempts to measure impact of contribution other than publication.

- **Publication counts, weighted or not, must not be used as indicators of research value. They measure a form of productivity, but neither impact nor research quality.**

Numerical indicators should never be used for comparisons across disciplines, and in particular for **comparison of researchers who work in different fields**. This is also true for **different sub-areas within Informatics**, which can have different publication and citation cultures. In addition, one should never look at absolute values, but rather at orders of magnitude, because even using various sources does not allow a high degree of accuracy.

Different alternatives are available for assessing publications and citations. **Public initiatives, such as ArXiv, Dblp, HAL, or Zenodo provide very useful accounts of the publications for individual researchers, laboratories, or institutions, because of their very low data noise, and their use should be favored.** In addition to providing metadata about publications, ArXiv, HAL, and Zenodo also provide access to the full texts for many of the stored publications. Dblp provides very useful accounts of the publications for individual researchers. **CORE has established itself as a reference on conference impact rankings** similar to that provided by the ISI Web of Science in other sciences whose publication culture is journal-based. **Some for profit companies provide services that are also quite useful for ranking publications.** Google Scholar has emerged as a reference source of bibliometrics and ranking information and includes not just journals but also conferences, which, as stated before, play a prominent role in Informatics. Microsoft Academic is a similar initiative. Both have the drawback of a higher level of noise in the data. Also, they keep control of the data and use black-box ranking algorithms, in contradiction with the open access criteria. **Clarivate Analytics' "Web of Science" is still inadequate for most of Informatics** because it is mostly based on journal publications,[5] while Scopus, powered by Elsevier, provides benchmarking tools but is also journal-biased and clearly does not satisfy the open access criteria.

The emergence of the new tools mentioned above, and specially their **combined use**,[6] has greatly improved the availability and quality of bibliometric data for Informatics, thus largely overcoming the challenges derived from the conference-centred publication culture. However, this is only true when performing evaluations within the field. **Outside the field, or in larger-scale evaluations that involve several fields** (e.g., in rankings across departments, universities, countries, etc.) **the traditional tools based solely or largely on journal citations** (such as ISI Web of Science or Scopus) **are still used routinely, with results in a very inaccurate portrayal of the contributions of Informatics**. **The recent trend towards the tight coupling of conferences and journals,** with the papers of conferences appearing in a journal, or conferences incorporating journal-first papers into their program, **can be very useful in this context.**

Artifacts

Artifacts are another possible proxy for research impact. More generally, the impact of a piece of research is high if what is produced by it can be used by others to further advance science. This may mean that the research tool produced is strong enough to be used or extended by others, or that the experiment performed can be replicated in other contexts to derive new insights.

To assess impact, artifacts such as software can be as important as publications. The classical ways of evaluating the impact of artifacts have been to use download counts, numbers of users, etc. These are however considered rough metrics which are indeed insufficient. However, there is also interesting recent progress in this area. In the past decade **an increasing number of conferences and journals[7] have established artifact evaluation committees**, as part of or in addition to their selection committees, which

[5] Jacques Wainer, Cleo Billa, Siome Goldenstein, "Invisible Work in Standard Bibliometric Evaluation of Computer Science," Communications of the ACM, Vol. 54 No. 5, Pages 141-146. http://10.1145/1941487.1941517

[6] E.g., using Google Scholar for citations, filtered by the clean publication lists of Dblp, and CORE for ranking the publication venues given by Dblp, while accessing the actual papers via ArXiv, HAL, or Zenodo.

[7] https://www.acm.org/publications/policies/artifact-review-badging

give a peer review-based measure of quality.

In addition, **in many areas of Informatics, software competitions are regularly organized to assess progress in tools** (e.g., SAT solvers or learning tools). Such competitions help to evaluate impact, but more importantly should increase the quality of the tools. Hackathons are also an interesting evolution in this context.

Quality criteria for software and artifacts have to be stated and shared by evaluation instances or committees. **Quality assessment should take into account novelty, applicability, provability, availability, and reproducibility.**

Since it is often difficult to credit only one researcher in a team for a software development, it is important in the context of evaluation to clarify the precise contribution of each team member. This allows crediting artifacts and software to individuals in the same way as publications, and using them for promotion in selection committees.

In the same spirit as Open Archives, initiatives aimed at facilitating the wide access and archival of software (such as the Inria and Unesco-sponsored Software Heritage, github, etc.) should be supported and used.

Awards

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 recent development is the awarding, also by major conferences or societies, of "most influential paper award" or "test-of-time awards", for the paper 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.

Innovation and commercialisation

Technological innovation is vitally important to 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, or adoption by industry or standards bodies, are a highly valued form of impact, and an excellent indicator of the applicability of research. Nevertheless, technological development by itself cannot be a substitute for scientific research.

Open Science

Another important recent development in research evaluation and impact assessment is Open Science. Open Science is about extending the principles of openness to the whole research cycle, fostering sharing and collaboration as early as possible; it advocates practices such as open access publishing, open data, and open peer review. **In evaluating impact, Open Science criteria such as transparency, and accessibility of results, data, and algorithms should be applied, and the value of collaboration acknowledged**. As pointed out in the EC report on Open Science,[8] this relies also on stakeholder and citizen engagement, and goes hand in hand with research integrity, legal and ethical awareness of researchers:

"For the practice of Open Science to become mainstream, it must be embedded in the evaluation of researchers at all stages of their career. This will require universities to change their approach in career assessment for recruitment and promotion. It will require funding agencies to reform the methods they use for awarding grants to researchers. It will require senior researchers to reform how they assess researchers when employing on funded research

[8] Evaluation of Research Careers fully acknowledging Open Science Practices — Rewards, Incentives and/or recognition for researchers practicing Open Science. European Commission - Research and Innovation doi:10.2777/75255.

projects. This is about changing the way research is done, who is involved in the process and how it is valued; evolving from a closed competitive system to one that is more open and collaborative. Overall, a cultural change is needed in organisations and in the research community for the promotion of and engagement in Open Science."

Informatics should acknowledge Open Science practices in its research evaluation. The Open Science Career Assessment Matrix (OS-CAM)[8] represents a possible, practical move towards a more comprehensive approach to evaluating researchers through the lens of Open Science.

It is interesting to note that without the development of computing and communication resources, Open Science research would not be possible. **Informatics thus also has a prominent role to play in the adoption and development of the Open Science approach**.

In conclusion, evaluation of research impact cannot be reduced to a single dimension or a single metric. A multi-criteria approach is recommended.

4. Towards more quality and impact

The use of quantitative criteria in research assessment seems an inexorable trend. Increasingly, assessments resort to publicly available bibliometric indicators provided by a variety of sources, without questioning their soundness and trustability. Although we acknowledge the usefulness of quantitative data and bibliometric indicators, we stress the following:

- **The goal of research assessment is primarily to assess quality and impact over quantity.**[9] This is especially crucial in the case of research evaluation in hiring and promotion cases. 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 least publishable increment practices. We observed that some existing evaluation campaigns established in European countries adopt 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 artifacts, trying to identify impact factors like their novelty, their supporting artifacts, and their impact on others' work.

- **Quantitative data and bibliometric indicators must be interpreted in the specific context of the research being evaluated. They should never constitute the sole ranking criterion.** Different research areas and even subfields behave very differently. Even within a homogeneous set, they 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 insight 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.

[9] This is eloquently stressed by the CRA Best Practice Memo of February 2015 "Incentivizing Quality and Impact: Evaluating Scholarship in Hiring, Tenure, and Promotion," by B. Friedman and F.B. Schneider.

INFORMATICS EUROPE

www.informatics-europe.org