EUTOPIA TRAIN
Transforming Research and Innovation

Open Science in research assessment

An overview of quantitative and qualitative approaches

Edited by Lennart Stoy (VUB), Elisa Maes (VUB), Tania Van Loon (VUB), Mojca Kotar (University of Ljubljana)

V1.0 – 28 September 2022

This project has received funding from the European Union's Horizon 2020 Framework programme under grant agreement No 101017419 EUTOPIA-TRAIN

Available on Zenodo via https://doi.org/10.5281/zenodo.7097264

1. Introduction

Research assessment reform is an urgent priority of European research policy.[1] While initially closely tied to the advancement of Open Science practices, this objective has grown into a larger reform movement which seeks to address many different shortcomings of dominant evaluation practices.

European University Alliances are directly concerned by this policy objective and are seen as possible testbeds for reform. EUTOPIA has also committed to implement a framework policy for research assessment and Open Science through the EUTOPIA-TRAIN project. We have explored the political momentum and its implications for European Universities, including EUTOPIA, in an earlier background note.

But how exactly implement a wide-ranging reform of research assessment within a university? What options are there to measure Open Science activities and outputs? How can a broader range of activities be reflected in assessment systems and methods? Which methods for qualitative assessment are being explored in this context? With this paper, we aim to present options that are being discussed and piloted at research performing organisations across Europe. We hope to stimulate the institutional transformation of our own alliance and other universities with the examples covered in this report.

Section 2 is a recap and overview of initiatives and declarations seeking to reform research assessment, including DORA, the Leiden Manifesto and others. Our summary focuses on the role of quantitative and qualitative assessment in these initiatives and declarations.

Quantitative indicators and metrics that could be used to measure Open Science activities and outputs are the subject of *Section 3*. The overview is based on existing source documents from various national and international frameworks. We also provide a summary of commonalities between these frameworks- as well as challenges when implementing quantitative approaches.

Section 4 is devoted to qualitative methods for assessment. Our overview is largely based on the emerging experience with narrative CVs, but other measures taken, for instance contextualising achievements and outputs, as well as contravening biases in qualitative assessment are addressed, too.

Assessment requires a comprehensive, up-to-date, and accessible evidence base. *Section 5* is therefore an excursion into this matter. Here, we focus on the underlying infrastructure and (open) information systems for assessment. We also summarise briefly the obstacles one encounters when assessing the FAIR-ness of research data. Lastly, *Section 6* summarises the main issues covered in this report.

2. A push for qualitative assessment

Over the last decade or so, research assessment has been subject of intense scrutiny, in particular its reliance on metrics and quantitative indicators. This attention has given rise to different declarations and recommendations how to implement responsible evaluation and assessment.

Probably the most renown is the 2012 San Francisco Declaration on Research Assessment (DORA). DORA's central recommendation is to move away from "journal-based metrics, such as Journal Impact Factors, as a surrogate measure of the quality of individual research articles". This reflects the

[1] See e.g. *Council conclusions on research assessment and implementation open science* (10 June 2022) https://data.consilium.europa.eu/doc/document/ST-10126-2022-INIT/en/pdf; ScienceBusiness (10 June 2022).*EU science ministers agree on research assessment reform.*https://sciencebusiness.net/news/eu-science-ministers-agree-research-assessment-reform.

recognition that the use of journal-based metrics is an inappropriate use of the intended purpose of the indicator. DORA also emphasizes the use of explicit criteria, considering that journal-based metrics might be used implicitly by reviewers to gauge the quality of articles, as well as the recognition of different outputs and impacts. DORA is now a community which collects case studies, a resource library, and a registry of assessment tools is developed in Project TARA.

The Leiden Manifesto is a declaration with the objective of fostering a *responsible use of metrics*. The authors of the manifesto, all experts in scientometrics, base this on the diagnosis of "that evaluation is now led by the data rather than by judgement. Metrics have proliferated: usually well intentioned, not always well informed, often ill applied". The Manifesto recommends that quantitative assessment should always support qualitative judgment and not supplant it. The authors also emphasise that performance or "excellence" occurs in different contexts which reviews and evaluations should consider.

The Metric Tide has been a pivotal report in raising awareness about the pitfalls of over relying on indicators with inherent flaws and introducing the concept of "responsible metrics".

The Hong Kong Principles for assessing researchers have the primary aim to foster "trustworthiness, rigor, and transparency" and are therefore geared towards reproducibility and research quality. They emphasise the role of Open Science practices such as stakeholder engagement, data sharing, and Open Access as potential indicators for responsible research practices along different stages of a research project. A summary of different dimension covered in the four declaration is shown in Table 1.

Table 1: Elements of international recommendations for responsible assessment[2]

	RECOMMENDATIONS	DORA	LEIDEN	METRIC TIDE	HONG KONG
METHOD	• **Journal-metrics as surrogate measure of quality**	✓			
	• Quantitative evaluation support qualitative assessment		✓	✓	
	• Qualitative judgment based on portfolios		✓		
	• Misplaced concreteness and false precision		✓		
CRITERIA	• **Explicit criteria used in evaluating**	✓			
	• Systemic effects of assessment and indicators		✓	✓	
	• Scrutiny and regular updating of indicators		✓	✓	
DATA	• **Open and transparent data and methods**	✓	✓	✓	
	• **Licence allowing unrestricted reuse**	✓			
	• Allowing those evaluated to verify data and analysis		✓	✓	
	• Best possible data in terms of accuracy and scope			✓	
VALUE DIVERSITY	• **All research outputs and broad range of impacts**	✓	✓	✓	✓
	• Missions of the institution, group or researcher		✓		
	• Excellence in locally relevant research		✓		
	• Variation by field in publication and citation practices		✓	✓	
	• Plurality of research and career paths			✓	✓
	• Responsible practices, complete reporting, open science				✓
	• Research activities and contributions				✓
	• Multilingualism and outputs in all languages				

University group LERU has also weighed in on the discussion on how to combine qualitative and quantitative assessment. Their recent report "A Pathway towards Multidimensional Academic Careers" argues that "assessment […] can and must be supported by objective data, but it cannot be reduced to simple metrics" and stipulates that "Quantitative measures can be useful in this process as a complement to a broad and sound qualitative approach". It also links assessment culture with the general research and professional culture in universities, warning that "a strong focus on number and reputation of publications can lead to a highly competitive, long-hours research culture, where bullying

[2] Taken from Bibliometrics and Research Policy, Nordic Workshop (2021): *NWB2021 - Janne Pölönen and Henriikka Mustajoki.pdf*. https://doi.org/10.6084/m9.figshare.16991992.v1; licensed under CC-BY 4.0

goes unnoticed and researcher wellbeing does not receive attention". Reforming research assessment according to this perspective is not just a question of finding more appropriate measures to gauge the quality and/or impact of research, but one of creating a better working environment within academia. In the words of LERU, "universities must invest in a positive research culture that stimulates a diversity of researcher profiles and recognizes a diversity of contributions".

The recently finalised *Agreement on Reforming Research Assessment*, developed by more than 350 institutions in Europe and beyond, in this light, takes strong inspiration from the various declarations and initiatives mentioned above. It aims, in a nutshell, to "Base research assessment primarily on qualitative evaluation for which peer review is central, supported by responsible use of quantitative indicators". Forming a coalition of organisations willing to contribute has the purpose to create critical mass which overcomes the collective action problem in a complex system.[3]

3. Quantitative indicators

Notwithstanding the debate about the responsible use of metrics in assessment, there are moves to measure and quantify Open Science activities and outputs. This can be for different purposes, such as monitoring or assessment, or different levels, such as individual, project, group or institution. In this section, we present quantitative indicators proposed by other studies, reports and frameworks concerning research assessment and Open Science.

The different frameworks presented in our previous report are the starting point for this exercise: OS-CAM, NOR-CAM, the Dutch Strategy Evaluation Protocol, and the Recommendation for the responsible evaluation of a researcher from Finland. The content from these resources is further enriched by two studies looking deeper into Open Science indicators:

- The report "Indicator frameworks for fostering open knowledge practices in science and scholarship" (hereinafter "*EC study*") was written by a group of experts for the European Commission and published in 2019. It aimed to develop comprehensive "**indicator frameworks**" and not individual metrics. It sought to accompany these frameworks by toolboxes which facilitate the application of a framework by a given community.
- The association of universities of science and technology, CESAER, released a "white paper on next generation metrics" in 2020 (hereinafter "*CESAER paper*"). This included a set of indicators for Open Science practice, along innovation and education metrics, in an attempt to look at indicators capturing different areas of activity within universities. These indicators are mainly intended to **measure the institutional performance** and not to evaluate individuals, e.g. for the purpose of career assessment

In the first step (section 3.1), we explore which principles and ideas the different frameworks and reports have in common. Secondly, we will present the main indicators proposed by them (section 3.2). Thirdly, we will summarize shortcomings and potential pitfalls of these indicators as highlighted in the source documents (section 3.3).

[3] See European Commission (20 July 2022). *Reforming research assessment: The Agreement is now final*. https://research-and-innovation.ec.europa.eu/news/all-research-and-innovation-news/reforming-research-assessment-agreement-now-final-2022-07-20_en

3.1 Common principles and approaches

3.1.1 Complementing qualitative judgement

The various frameworks share how they stress the importance of qualitative judgement (see Table 2 for a succinct comparison of the frameworks used nationally). All argue for a **supportive nature of quantitative indicators**. **NOR-CAM** explains the rationale to limit the usage of quantitative indicators with the argument that "bibliometric indicators are of limited value, because they are retrospective, take no account of the context and cannot replace decision-making responsibility". Acknowledging the complexity of academic work, **OS-CAM** reasons that "evaluating a researcher cannot be reduced to a number because their merits and achievements are a complex set of different variables, difficult to be summarised by a single figure." It argues that "A better approach is through multi-dimensional criteria evaluation, taking into consideration what is expected from a researcher and what is relevant for his/her career/recruitment". **CESAER** also recommends to use "a polychrome approach and to use indicators to complement qualitative expert assessment".

Table 2 Comparison of national assessment frameworks[4]

	THE NETHERLANDS	FINLAND	NORWAY
METHODS	• Emphasize quality of work over quantitative results	• Metrics support qualitative assessment • **Diverse group of evaluators** • Characteristics of research fields	• Balancing quantitative and qualitative goals
CRITERIA		• Transparent objectives & criteria • **Equity & non-discrimination**	• Transparency of criteria & merits • **Gender balance and diversity**
DATA		• Comprehensive, reliable, comparable • **Researcher self-evaluation**	• Systematic & structured documentation • **Applicant's own reflection**
VALUE DIVERSITY	• Enable diversification and vitalisation of career paths • Encourage all aspects of open science • **Acknowledge individual & team performance** • **Encourage high-quality academic leadership**	• Teacher and supervisory roles • Societal impact and interaction • Research & other communities • Open access effort • Research integrity • **Evaluation benefits the researcher**	• Recognise several competencies but not all by everybody • Assess all results, activities and competencies within Open Science principles • **NOR-CAM framework** (outputs, process, pedagogy, impact/innovation, leadership, other experience)

The **Finnish framework** adds that "Peer review should be the primary approach for evaluating individual researchers". **OS-CAM** explains that "good decisions require qualitative judgement, preferably by a panel of independent researchers who respecting the principles of openness, transparency and merit, assess the range of a researcher's achievements, whether this be for a new position, career advancement or for a funding grant. It is important for evaluators to consider profile and balance of the collective criteria".

3.1.2 Bigger picture: academic assessment

The frameworks are also conscious that evaluation of Open Science is not a practice occurring in isolation. **OS-CAM** adds that "To encourage and recognise Open Science activities, it is important to go beyond Open Science and frame this discussion in the broad context of the evaluation of researchers". The various **national frameworks**, also do not limit themselves to research and/or Open Science activities, but a chose broader approach encompassing many different dimensions of academic activity (Table 3). **CESAER** also highlights this in the recommendation to "Consider integrating principles of open science broadly within other initiatives such as the human resources strategy for researchers".

[4] Taken from Bibliometrics and Research Policy, Nordic Workshop (2021): *NWB2021 - Janne Pölönen and Henriikka Mustajoki.pdf*. https://doi.org/10.6084/m9.figshare.16991992.v1; licensed under CC-BY 4.0

Table 3 High-level dimensions of selected assessment frameworks[5]

Framework	High-level dimensions	
NOR-CAM	▪ Research output ▪ Research process ▪ Pedagogical competence	▪ Impact and innovation ▪ Leadership ▪ Other experience
OS-CAM	▪ Research output ▪ Research process ▪ Service and leadership	▪ Research impact ▪ Teaching and supervision ▪ Professional experience
Finland	▪ Research funding and research supervision and leadership experience ▪ Teaching merits and experience ▪ Awards and honours	▪ Assessment of scientific and academic merit ▪ Scientific and academic networking and community development
SEP	▪ Research products for peers ▪ Use of research products by peers ▪ Marks of recognition by peers ▪ Research products for societal target groups	▪ Use of research products by societal target groups ▪ Marks of recognition by societal target groups

3.1.3 Context of evaluation

The previous sections highlighted how recent assessment frameworks aim to prioritise qualitative judgement over quantitative methods. One of the reasons behind this development is that qualitative judgement is thought to better capture the **context of an evaluation**. According to **the EC study**, one should for example chose indicator frameworks instead of individual indicators. These should be sensitive to

> *"the purpose of an evaluation (e.g. monitoring, learning, resource allocation),*
> *the mission of the research under scrutiny,*
> *the level of evaluation (system, organization, researcher/group),*
> *disciplinary structures or cultures,*
> *stakeholders and beneficiaries of the research,*
> *and the research environment."*

The EC study also argues for such a comprehensive approach because it sees it as a challenge to develop single universal criteria, considering both diverging definitions of Open Science and the diversity of research practices. In the **Finnish report**, this is succinctly summarized as "The indicators used in assessment should be chosen to support the aims of the evaluation". In other words, one-size-fits-all indicators should be avoided and a discussion of which indicators are fit for which purpose should support the evaluation.

OS-CAM explains that "the weighting for each criterion should recognise the background of the researcher being evaluated". For instance, candidates from outside academia might have different backgrounds and strengths not recognised in assessment frameworks developed for academia. OS-CAM also emphasises that it offers an array of potential criteria and that it's not appropriate to expect a single person to fulfil all of them. **NOR-CAM** similarly states that "Requiring that all academic staff members

[5] Taken from Stoy, L., & Maes,E. (2022). *From impact factor to responsible evaluation. Overview of developments in research assessment and implications for EUTOPIA* (Version 1.2). Zenodo. https://doi.org/10.5281/zenodo.6323213; licensed under CC-BY 4.0

shall be assessed equally with regard to all three main objectives [of the framework] is therefore unrealistic and unnecessary." In short, the context of an evaluation should always be taken into consideration when applying specific criteria.

3.1.4 Access to underlying data

A recurring theme is the **accessibility of the underlying data**, which originates in part from the critique against the black-box nature of commonly used, commercially available bibliometric indicators. The **EC study** emphases that "all indicators should be based on FAIR data, the algorithms should be open source, and the frameworks should be governed by appropriate licensing and intellectual property regimes to prevent data monopolies or oligopolies. Non-transparent indicators should be excluded from the indicator frameworks used by funders, research performing organizations, and publishers". The **Finnish framework** explains further that "Being as open and transparent as possible in data collection, analytical processes and results is necessary. Those being evaluated should, as far as possible, be able to check both the data used and the results of the analysis". This recommendation reflects arguments of **DORA and the Leiden Manifesto** (see section 1).

3.2 Suggested quantitative indicators

Based on the principles outlined above, what type of quantitative indicators can or should be used to measure Open Science and basis for subsequent assessments or evaluations? We have collected the suggested indicators from the various source documents (Table 4). The table should be read in conjunction with the information and purpose of each of the sources:

- The indicators derived from the **CESAER paper** are intended to be used for evaluations at the institutional level, not for individuals or specific projects. Their ten potential indicators for science activities within a university include Open Science as well as other activities, although openness features prominently in the paper.
- The **EC study** collected 149 different indicators and judged them based on their appropriateness for being used during evaluations at different levels (institutional, project, individual) as part of the indicator frameworks. We are including the ones deemed as **suitable or promising for career assessment** (7 out of 149 indicators) in Table 4. This is based on the assessment of the original authors that they are "sufficiently developed and the data sources required to measure the indicator are of sufficient quality". This selection also means that several Open Science indicators are excluded from the summary, such as Open Access publications or research data.
- In the Norwegian assessment framework, **NOR-CAM**, open science would be measured under research outputs. The source for the underlying data are the institutional/national CRIS installations and other databases. While the report is not explicit about the openness requirements of these indicators, it explains that "Open access to research results and openness in the research process are key objectives that improve verifiability and quality, and increase the utilisation of knowledge. The degree of openness should therefore be indicated in descriptions and reflections in all of the categories". It adds that, "Emphasis is placed on open access to published works and other results, as well as whether the data adhere to the FAIR principles."
- The **Finnish framework** provides more detail on specific Open Science-related activities. It explicitly mentions "merits in producing and sharing research and information materials" as a possible criterion. Data for such activities are to be collected under a **national knowledge base**. No further detail about the exact design of the indicators is given but it is mentioned that "Researchers' activities to promote open access to research outputs is considered as part of the evaluation."

Table 4: Overview of Open Science indicators proposed by selected frameworks (Source: own compilation)

Framework		OS-CAM	NOR-CAM	SEP	Finland	CESAER	EU study
Dimension	Level	Flexible	individual	Groups	Individual	Institution	flexible
Publications (Open Access)		▪ Publishing in open access journals ▪ Self-archiving in open access repositories	▪ Published works	▪ (Open access) Journal articles and reviews (refereed/non-refereed) ▪ (Open access) Books, source publications and exhibition catalogues (refereed/ nonrefereed) ▪ (Open access) Book chapter (refereed/ non-refereed)	▪ Open access to scholarly publications	▪ Open access publications ▪ Top 10% most cited publications ▪ Citation impact ▪ Interdisciplinary publications ▪ International PhD students and postdocs ▪ Repository traffic	(included but not considered mature)
Research data		▪ Using the FAIR data principles ▪ Adopting quality standards in open data management and open datasets ▪ Making use of open data from other researchers ▪ Being aware of the ethical and legal issues relating to data sharing, confidentiality, attribution and environmental impact of open science activities	▪ Datasets	▪ Digital infrastructures and databases ▪ Data sets	▪ Open access to and quality and impact of research data	▪ Open and FAIR data sets	(included but not considered mature)
Citizen science and societal engagement		▪ Actively engaging society and research users in the research process ▪ Sharing provisional research results with stakeholders through open platforms (e.g. Arxiv, Figshare) ▪ Engaging in open innovation with partners beyond academia	▪ Citizen science		▪ Societal interaction	▪ Citizen science projects ▪ Publications with non-academic sector	▪ Nr. papers co-authored with civil society actors ▪ Nr. of refereed publications authored with non-academics
Training and education		▪ Training other researchers in open science principles and methods ▪ Developing curricula and programs in open science methods, including open science data management ▪ Raising awareness and understanding in open			▪ Open access to and the quality of information content, educational quality and	▪ Open science training ▪ Open Education Resources	▪ No. of students in sandwich [sic!]

Framework		OS-CAM	NOR-CAM	SEP	Finland	CESAER	EU study
Dimension	Level	Flexible	individual	Groups	Individual	Institution	flexible
		science in undergraduate and masters' programs			impact of research-based educational materials		
Software		▪ Using open source software and other open tools ▪ Developing new software and tools that are open to other users	▪ Software	▪ Software		▪ Open source software	▪ Nr of citations to software
Collaboration		▪ Widening participation in research through open collaborative projects ▪ Fully recognizing the contribution of others in research projects, including collaborators, co-authors, citizens, open data providers ▪ Engaging in team science through diverse cross-disciplinary teams					▪ Publication of co-author statements ▪ Openness on contributorship
Peer review		▪ Involving stakeholders in peer review processes ▪ Contributing to open peer review processes ▪ Examining or assessing open research					▪
Other		▪ Taking account of the risks involved in open science ▪ Securing funding for open science activities	▪ Methodologies ▪ Artistic results ▪ Research reports				▪ Simulation results

- In the **Dutch Strategy Evaluation Protocol**, evaluations of research groups are to be composed of a self-evaluation report, which is supplemented by quantitative indicators. According to the source document, research units can use quantitative indicators for the areas "research activity, progress and impact". These need to be justified and the "direct relationship between the arguments with regard to the aims and strategy on the one hand and the type of robust data underpinning the self-evaluation on the other" should be explained. The examples provided in Table 4 are considered "evidence" for research outputs.
- Finally, OS-CAM, being entirely dedicated to Open Science, lists a host of different evaluation criteria that can be used to embed Open Science in evaluations of across academic tasks (outputs, activities, service/leadership, impact, teaching/supervision, professional experience.). It is not the purpose to reproduce the entire matrix here, but we highlight several quantifiable criteria. These can be found under indicators for "research outputs", such as publications, datasets, software and funding, for the "research process", such indicators as pre-print sharing, cross-disciplinary work, contributorship practices etc.

For this exercise, we have only listed indicators and examples with a connection to openness. As highlighted in section 3.1, most frameworks emphasise the context as well as broad ranges of activities and outputs. This is reflected in their overarching suggested elements and corresponding indicators, concerning e.g., education, leadership roles, service to the scientific community, mentorship, editorial work etc. However, we have chosen to focus the overview in Table 4 on Open Science-related indicators.

The collection of indicators also highlights how the frameworks converge with regards to the Open Science practices they include. All cover Open Access publications and research data, be it with different foci. Our selection of 'mature' indicators from the EC study meant that these dimensions are not shown in the table, although they are considered in the source document. Other prominent practices include software sharing as well as citizen science activities. It showcases how widening the criteria used for assessment can incorporate typical Open Science practices.

3.3 Challenges

Considering the many voices arguing for a supportive nature of quantitative indicators it is unsurprising that the source documents also warn against the (mis)use of quantitative Open Science indicators . These go beyond the recommendation to use metrics only as supplementary to peer review.

Some acknowledge the difference of means to make outputs open and accessible. The Finnish framework recognizes this shortcoming and states that "Providing open access to research outputs often requires both time and other resources". It also acknowledges that there are differences in areas such as science communication, opportunities for commercialisation of research, ability to provide open access to output etc.

Others are cautions about consequences of measuring specific types of outputs and stress considering the effects of metrics before implementing them. The EC study provides explanations about their potential but also the risks and why some have not been deemed appropriate by the expert authors. They warn, for instance, that an indicator on communicating and sharing data could "privilege sharing data as an end-goal over generating quality data". Likewise, counting citations of data journals could "encourage symbolic citation behavior" or "over-emphasiz[e] citation analysis". An indicator for "Openness on contributorship" could "easily be gamed." Looking at Open Access publications risks that "Monitoring easily shifts into evaluative mode". It also warns that underlying data for some of these indicators is often proprietary. Finally, using one indicator will usually lead to the production of the type of output that it captures (also known as Maxwell effect). I

Other **indicators are lacking maturity** at the time of writing of the **EC study**. For example, counting publications with associated datasets suffers from the fact that "DataCite is still weakly developed, and data are sparse". Similarly, software citations are not a mature indicator yet, because "Archiving and referencing software is not yet a stable practice". In other words, these may be used once adoption and underlying data are more widespread.

A similar challenge exists for **OS-CAM** indicators. The coverage of OS-CAM criteria by CERIF (a common data standard for research information systems) was only fully achieved for some 14% of criteria by 2021. 52% of criteria are only partially covered.[6] In short, the technical infrastructure to capture quantitative indicators requires further development – an issue explored in more detail in section 5 of this report.

4. Qualitative methods

Many of the presented national frameworks and declarations, as well as the European coalition to reform research assessment put a strong focus on using qualitative approaches for assessment. Particularly interesting here is the role of academic CVs, how they are structured and what activities are covered by them (Woolston, 2022). Overarching recommendations in the form of 'Ten ways to improve academic CVs for fairer research assessment' have been shared by Strinzel et al. (2022) on behalf of the CV Harmonization Group (H-Group). These cover issues such as instructions for users of CVs, prioritizing achievements over prestige, focusing on recent achievements which are relevant, covering a broad range of activities and outputs, balancing incentives and a cautious use of metrics, using the academic age, using narratives instead of lists, as well employing open and interoperable data systems.

In this section, we will provide an overview of the idea behind and the use of narrative CVs at several European funders and research organisations (section 4.1). A related approach is to ask for qualitative assessments of achievements and outputs in proposals or CVs (section 4.2) and to ban specific types of proxy indicators from being used (section 4.3). We will also briefly address work that has been done to mitigate bias in qualitative assessment in section 4.4.

4.1 Narrative CVs

A major way to foster more qualitative assessment are so-called narrative CVs. These are CVs prepared in a way that a person, e.g. an applicant for a position or a grant, reflects on their achievements and/or career goals by answering a set of open-ended questions. This is an alternative to purely listing information about positions, publications etc. without providing contextual information.

The Resume for researchers, developed by the **Royal Society**, is a seminal type of narrative CV. The template contains the personal details, four modules, a personal statement, and additions. The four modules comprise of contributions to 1) the generations of knowledge, 2) to the development, 3) to the wider research community, and 4) to broader society. Each part is to be filled in a narrative style, although some parts of the personal details and additions can likely also be in a list format. British funder UKRI has adopted the Resume for researchers in its processes.

Narrative CVs have also been introduced by the **Luxemburgish funder FNR**. A template is available

[6] Mustajoki et al. (2021). *Making FAIReR assessments possible. Final report of EOSC Co-Creation projects: "European overview of career merit systems" and "Vision for research data in research careers"*. https://doi.org/10.5281/zenodo.4701375)

online, which is structured in three sections (*Personal Statement; Personal details – Individual narrative profile; Key outputs, contributions, and achievements*). All sections are to be filled in in a narrative way and go beyond Open Science practices. The template allows, for instance, to "mention significant life events, career breaks, secondments, volunteering, part-time work and other relevant events or experience", which might not fit the format of traditional CVs.

The **Swiss funder SNSF** also tested a "new CV format in biology and medicine". This template combined different quantitative and qualitative approaches. For instance, the template included h-indexes, the Relative Citation Ratio, and calculations of the 'academic age' of researchers. The narrative part was split into a 'project related narrative' and 'Contributions to science'. Notably, the CV template also allowed to cite five works, including "any other type of research output, such as preprints; patents, software etc". According to a preprint summarizing the results of the test, "the implementation of a new, well-structured CV format is not only feasible but also something that many stakeholders welcome" (Strinzel et al. 2022). Following the pilot, SNSF has announced to introduce a narrative CV template as default.

The **Dutch research funder NWO** uses a uniform narrative CV format, for example in its Talent Programmes Veni, Vidi, and Vici. The templates are available on the NWO website. Here, applicants are asked to write a narrative academic profile. Key outputs can be listed and must be motivated with a short statement about the importance and relevance of the item. With specific regards to Open Science, the form comes with the request to mark openly available outputs. In the academic profile, applicants are also invited to (optionally) explain their "Contributions to (FAIR) open data and open science".

The **University of Glasgow** tested the resume for researchers with early career researchers using mock review panels. According to the study report, the objectives of the narrative CV were appreciated but the additional workload deemed a challenge. Recommendations stemming from the study highlight that the narrative CV "does not exist in isolation from the rest of the system" and that, for example, attention needs to be paid to unconscious bias and that support to fill in the narrative template is needed. A short online course "Narrative CVs, tips for how to write one" has been developed as part of the project.

4.2 Achievements in context

Closely related to the development of narrative approaches is the contextualization of research outputs and achievements. This can be done as element within a narrative CVs. Research funders are also implementing such elements in their proposal templates independently from narrative CVs.

The **NWO** invites applicants to motivate why they selected a given output and how it is relevant and important. Practically, the proposal templates ask to give key outputs together with "a motivation for the selection of each of these output items: how does this output show your abilities/qualities as a researcher and/or how is it related to the Vici project.".[7]

A similar move has been made by the **European Union**, which asks applicants for Horizon Europe funding to list five achievements and to provide a brief summary of their relevance. This can be "publications, widely-used datasets, software, goods, services, or any other achievements relevant to". These are to be substantiated by "a short qualitative assessment of its impact and (where available) its digital object

[7] Proposal template for NWO Talent Programme Vici - Social Sciences and Humanities (SSH) 2022 https://www.nwo.nl/sites/nwo/files/media-files/Vici%202022%20Full%20proposal%20form_0.docx (last accessed 27/07/2022)

identifier (DOI) or other type of persistent identifier (PID)."[8]

This fulfils two objectives. On the one hand, the request to explain the relevance of a given output or achievement aims to shift the focus away from proxy indicators (e.g., the publishing venue) to the quality and relevance of a given output/achievements. On the other hand, the focus on output or achievements broadens the scope from a focus on publications to other relevant activities or outputs, such as research data, software, mentorship etc.

4.4 Excluding certain assessment approaches

The misuse of proxy indicators is a main driver for the research assessment reform movement. Consequently, organisation have moved to ban certain metrics from being used in evaluations.

In the aforementioned **NWO proposal template**, applicants are being requested not to "mention H-indexes, impact factors, or any type of metric that refers to the journal, publisher, or publication platform, rather than to the individual output item". The proposal form explains that "the scientific content of a paper is much more important than publication metrics or the identity of the journal in which it was published."[9] The objective is to steer away from using proxies for quality, such as the publication venue or publisher, and instead focus on the content of an achievement in itself.

Likewise, the **European Research Council** has signed DORA and adapted its evaluation processes. Here, the approach differs from NWO in that it only bans the use of the journal impact factor as a proxy indicator. The programme guide makes this explicit, allowing "field relevant bibliometric indicators" when providing an achievement track record but forbidding the journal impact factor.[10] In the general Horizon Europe programme, the advice to evaluators is provided in the video "How to evaluate Open Science in Horizon Europe proposals" from March 2022. The video explains that the "**significance of publications [is] to be evaluated on the basis of proposers' qualitative assessment** and not per Journal Impact Factor" (emphasis in original).

Finally, **Luxemburgish funder FNR** emphases in their peer review guidelines that reviewers should "focus on the content and quality of scientific outputs, rather than their number, the venue of publication, or the aggregate metrics of the journals and researchers.".[11] This guideline is explicitly justified with FNR becoming a signatory of DORA.

4.3 Mitigating bias in qualitative assessments

Introducing qualitative assessments does not come without challenges. While qualitative approaches seek to prevent a misuse of metrics, they open another risk of subjective interpretations of the quality or relevance of achievements and/or outputs as explained in narrative texts. Therefore, proponents of qualitative assessments put an emphasis on bias mitigation measures, including reviewer training and guidance, as well as analyzing review processes for potential problems.

[8] Horizon Europe Programme Standard Application Form (HE RIA, IA), Version 4.0, 21 January 2022 https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/temp-form/af/af_he-ria-ia_en.pdf (last accessed 27/07/2022)
[9] Proposal template for NWO Talent Programme Vici - Social Sciences and Humanities (SSH) 2022.
[10] ERC Work Programme 2022. European Commission Decision C(2021) 4860. https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/wp-call/2022/wp_horizon-erc-2022_en.pdf
[11] FNR Peer Review Guidelines – ATTRACT 2022 Call. https://storage.fnr.lu/index.php/s/JLp1VtEQXGKhT4N#pdfviewer (last accessed 28/07/2022)

As briefly outlined in our previous report, reliance on context-insensitive quantitative indicators can be problematic for equality, diversity and inclusion. In this light, acknowledging potential biases and mitigating them becomes even more important for qualitative assessment approaches to deliver on their promise of making assessment more fair.

DORA itself has released a report addressing concerns and solutions concerning biases in narrative CVs. It acknowledges the fact that using qualitative methods has caused confusion about "expected content and how it should be evaluated." In particular, the report recommends to build a better foundation and understanding how narrative CVs should be filled in. It also underlines the need to train reviewers, researchers, and other staff about evaluation methods and processes. Finally, it stresses the need to monitor the implementation of narrative CVs.

To give one example, the **Luxemburgish funder FNR**, already mentioned above, provides extensive guidelines for the peer review of proposals (see FNR download center). The research funder emphasises that its reviewers should "be sensitive to legitimate delays in active researchers' activity, and personal factors".[12] In addition, FNR practices the optional participation of observers for unconscious bias in evaluation panels. The observer may add "independent comments about the panel proceedings and the management of unconscious bias to the panel report", but is not involved in the scoring of a proposal in itself.[13]

In the **Netherlands, NWO** has launched an initiative on inclusive assessment which likewise seeks to "optimise its evaluation processes and broaden the often limited ideal image of what a good researcher or a good proposal is". The initiative offers guides and videos for written assessment and evaluation meetings.

UK funders such as the **Medical Research Council (MRC)** and the **Engineering and Physical Sciences Research Council (EPSRC),** both parts of UKRI, also provide guidance on unconscious bias and responsible use of metrics or, respectively, act to "better understand implicit reviewer bias and how to reduce this impact by using alternate approaches to ensure a fair research funding system".

In short, organisations are actively working to mitigate possible biases in more qualitative assessment approaches. Naturally, potential biases and their mitigation measures concern more than Open Science. This is evident in the link to equality, diversity and inclusion initiatives.

5. Information systems and infrastructure for assessment

Many of the studies and frameworks explored in this report place an emphasis on the openness and FAIRness of their underlying data. Data used for evaluations should itself be findable, accessible, interoperable, and re-useable (FAIR). This makes assessments more verifiable and reproducible and stands in contrast to proprietary, closed systems under the control of commercial entities, such as the assignment of Journal Impact Factors. While not eliminating biases, the openness of such as system would also mean that data and results can be scrutinized and biases be identified more easily.

An added benefit is also highlighted by Strinzel et al. (2022), who emphasize that automated systems

[12] FNR Peer Review Guidelines – ATTRACT 2022 Call. https://storage.fnr.lu/index.php/s/JLp1VtEQXGKhT4N#pdfviewer (last accessed 28/07/2022)

[13] Unconscious bias observers can also be part of university hiring committees. Departments at the University of York have used such a scheme since at least 2014 (https://www.york.ac.uk/chemistry/ed/resource-hub/unconscious-bias-scheme/).

"reduce the burden on researchers and institutions alike of checking and correcting the content of CVs.".

Openness of such a system is crucial, on the one hand, to allow global use of these tools and, on the other hand, to be independent from costly commercial suppliers of proprietary data. In the words of Strinzel et al. (2022), "such licenses and open technologies are vital to preserve a common pool of shared information, which can be used by any organization in the world with access to the internet without costly licensing fees or subscriptions".

Building such a system in form of a comprehensive, up-to-date, and accessible evidence base is not trivial. This section introduces briefly some ideas about such an infrastructure and (open) information system for assessment under the term 'open researcher profile'. We also summarise the obstacles one encounters when assessing the FAIRness of research data, a related technical issue.

5.1 Open researcher profiles

The thinking behind such open researcher profiles has been subject of two major studies in recent years. The first was commissioned by Knowledge Exchange, a group of European organisations working on e-infrastructures for research in 2021. Their study "Openness Profile: Modelling research evaluation for open scholarship" explores a system for open academic CVs from a more technical point of view.

A representation of the OP as a user-curated portfolio of contributions to open scholarship. The OP would include a narrative component to enable the research contributor to contextualise their work as well as contributions drawn from their ORCID record and from the web. Other contributions would be allowed that have no URL as descriptive text.

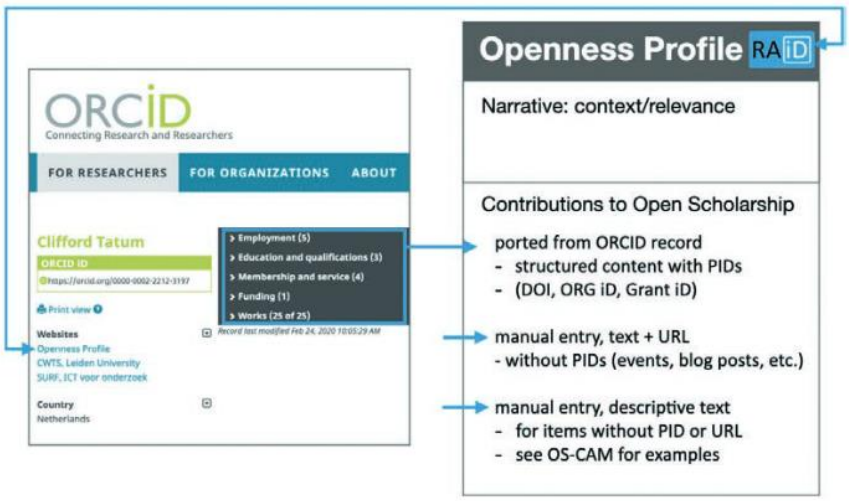

Figure 1 Openness profile schematic[14]

The report provides an analysis of what kind of information infrastructure could be used to capture and evaluate open scholarship from the perspective of research funding and research performing organisations. Their proposal is an **infrastructure based on fetching information from a researcher's ORCiD record** (Figure 1). Here, the records collected by ORCiD would be enriched by descriptive information provided manually by the user, resulting in the 'Openness Profile', which is a "user-curated portfolio of contributions to open scholarship". This would include both the outputs and the role of a contributor, for example using the contributor roles taxonomy (CRediT).

Another report "Making FAIRer assessment possible", funded through a European Open Science Cloud (EOSC) co-creation fund, also analysed assessment infrastructures, though with a stronger focus on FAIR data as a topic of assessment and the FAIRness of the underlying data in itself. The report provides three

[14] Taken from Jones, Phill, & Murphy, Fiona. (2021). Openness Profile: Modelling research evaluation for open scholarship. https://doi.org/10.5281/zenodo.4581490; licensed under CC-BY 4.0

major recommendations, namely 1) comprehensiveness, 2) interlinking/interoperability and 3) openness of the system. It also includes a three-step plan for assessment (see Figure 2 below). First, it is about **recognising the diversity of practices** and outputs related to FAIR date. Second, it stresses the need to "make it possible" by **building the corresponding infrastructure**. Third and finally, this needs to culminate in assessments that reward the practices from the first step by employing the infrastructure from the second step.

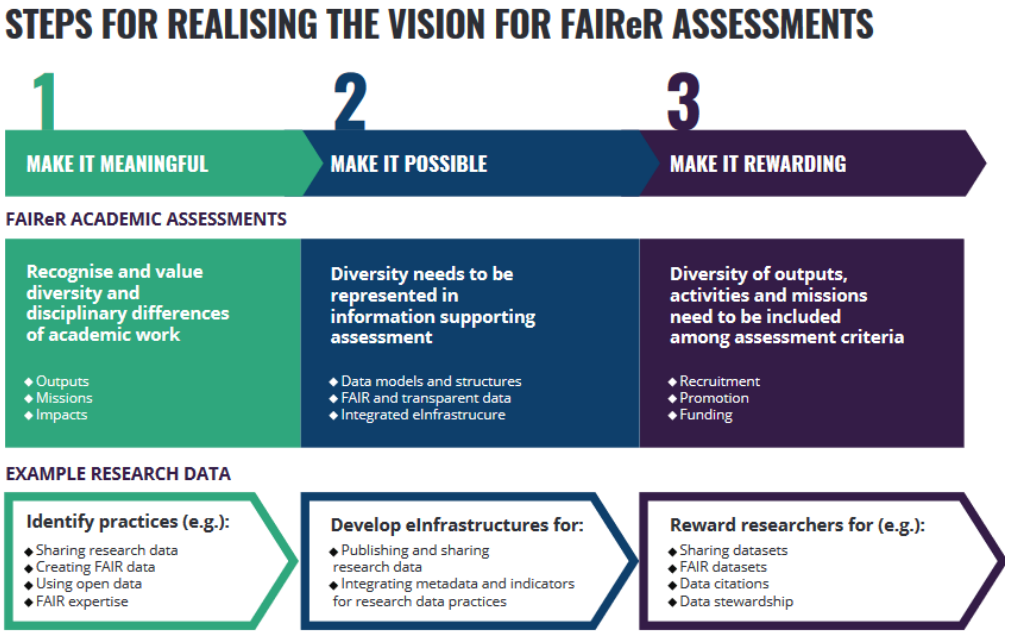

Figure 2 Steps to realise FAIRer assessments [15]

What these two studies have in common is their focus on the information infrastructure to support assessments. Especially when moving towards more 'complete' pictures of research, it is important to create systems which capture a broad range of activities and outputs, are easy to use, automated and interoperable across organisations. This requires investment in current research information systems.

The ownership of assessment data, as highlighted by the examples in section 3.1.4 is also an important concern for stakeholders. In several examples, notably Norway and Finland, the need for a corresponding national infrastructure for CVs, based on national and institutional CRIS was highlighted. Similar objectives are part of the European Open Science Cloud (EOSC) initiative.

5.2 Measuring FAIR-ness of data

One growing, major dimension of Open Science is FAIR and open research data. Researchers and research-performing organisations are encouraged to manage and share their research data accordingly. However, measuring the degree of FAIR-ness of data is not trivial, nor is the resulting question how to use it to assess an institution, a group, a researcher etc.

It is not trivial because the FAIR principles are **not intended as fixed standards** but as a "guide to data publishers and stewards to assist them in evaluating whether their particular implementation choices are rendering their digital research artefacts Findable, Accessible, Interoperable, and Reusable"

[15] Taken from Mustajoki et al. (2021). *Making FAIReR assessments possible. Final report of EOSC Co-Creation projects: "European overview of career merit systems'' and "Vision for research data in research careers"*. https://doi.org/10.5281/zenodo.4701375); licensed under CC-BY 4.0

(Wilkinson et al. 2016). In other words, the FAIR principles do not prescribe concrete steps for each of the four dimensions, which makes implementation and measurement a complex tasks.

Implementing different elements of the FAIR principles, such as meeting domain-relevant community standards or metadata being well described, is highly context dependent. Standards also change over time and therefore require constant maintenance. On the other extreme, the existence of permanent identifiers or license descriptions is rather easily verifiable. One can contrast this with the relatively straightforward check if a research article has been given an Open Access license and a persistent identifier.

There have been attempts to develop methods to evaluate research data and metadata according to the FAIR principles. The FAIR Data Maturity Model (Bahim et al. 2020) has classified different dimensions of the FAIR principles as 'essential', 'important', and 'useful', which helps assessing the most crucial qualities to consider a given data set as 'FAIR'.

Different FAIR assessment tools have been built based on such considerations. A comparison of five such tools by Peters-von Gehlen et al. 2022 finds that "neither one […] is fully fit-for-purpose to evaluate (discipline-specific) FAIRness".The authors recommend to develop hybrid approaches combining manual and automated assessments to account for the complexity it takes to consider a dataset FAIR.

This problem is exemplified by the example of the F-UJI Automated FAIR Data Assessment Tool. F-UJI produces a visual report with information how well a dataset meets the FAIR principles (Figure 3). The dataset receives an overall score and scores as well as a "FAIR level" for each FAIR dimension.

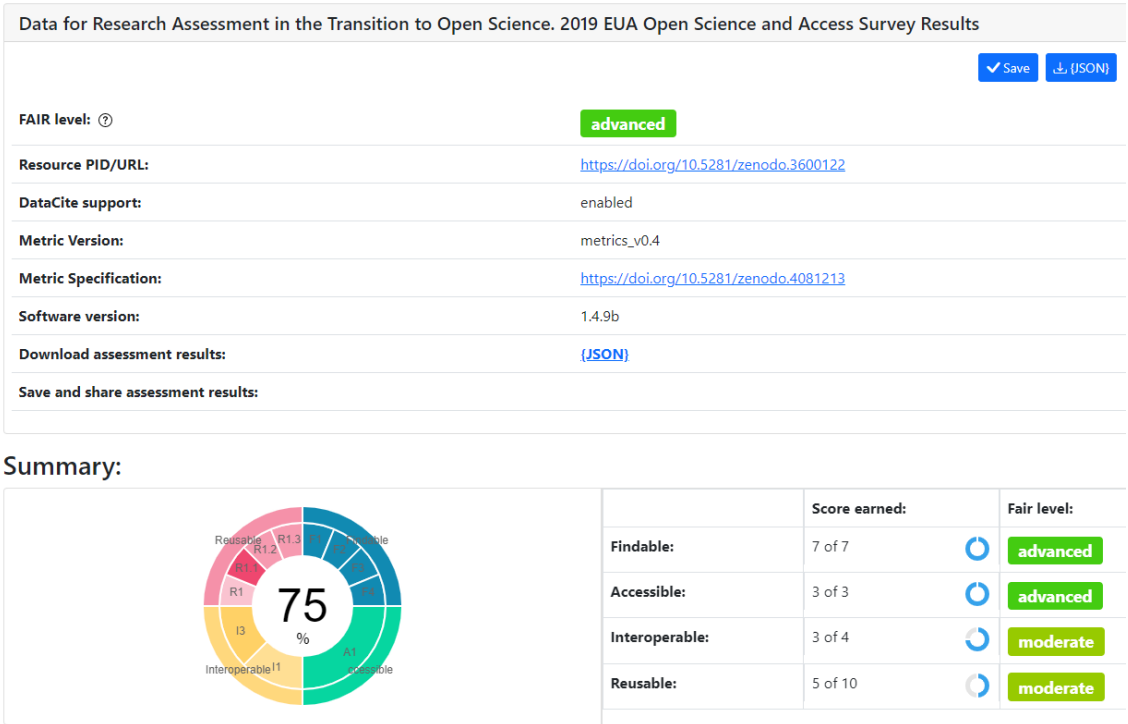

Figure 3 F-UJI report summary for a dataset

However, when using such a tool, there is a risk of re-introducing methods in assessment that were not developed for this purpose – a problem that most frameworks presented in this report sought to mitigate. Consider for example the question how to assess what makes a dataset 'FAIR enough'. Should overall scores be used or individual values for Findable, Accessible, Interoperable, or Reusable? In the given example, the score is lowest for Reusability. This could be mirroring the status of a community in which no widely adopted standards exist (yet).

These are questions which need to be deliberated and, ideally, be acknowledged in an evaluation. This is particularly important when the objective is to replace rather opaque indicators such as the journal Impact factor or h-index, which have turned into tools far beyond their original purpose. In short, while promoting FAIR data, one should be reflective about how it can be measured and implemented in assessment.

6. Summary

This report started with the objective to make different quantitative and qualitative approaches to include Open Science in assessments more visible and to provide practical ideas how to implement them. While it is for each institution or network to set their own methods and parameters for assessing research and researchers, there are several key takeaways from the various source documents used for this report.

- A clear trend is to **blend quantitative and qualitative approaches**. Almost every source highlights the supporting character of metrics – of any kind – to qualitative assessment based on peer review.
- There appear to be **few mature quantitative indicators** to be used to assess Open Science activities and outputs. If used, they mainly cover the Open Access status of publications or research data.
- The national frameworks guiding evaluations are **not necessarily using Open Science indicators as a rigid requirement but as encouragement and recommendations**. They acknowledge differences in opportunity for researchers in different fields and disciplines. This is possibly also limited by the respective national context of the framework, e.g. that enabling Open Access of articles via transformative agreements. In countries with a very comprehensive coverage to publish in Open Access, the rewarding aspect of Open Access to articles may even be negligible.
- There is emphasis, besides qualitative judgement, to **use metrics in context.** This also implies that the purpose of a (proxy) metric and the phenomenon that seeks to capture (e.g. quality, impact) should be reflected about during an assessment. In some cases, specific metrics are explicitly banned from being used in assessments by research funders (e.g. Journal Impact Factors).
- Regarding **qualitative methods**, funders and other research organisations are exploring **narrative CV formats**. These have the potential to capture wider contributions and achievements of researchers. However, they introduce new challenges concerning the workload, training needs, or unconscious biases.
- In the literature, there is also a recognised need for an **open and interoperable infrastructure** that supports assessments. This would ease access to information across countries and institutions to underlying data. Ideally, information would be standardised in order to minimise work required to produce CVs. Some countries, Norway and Finland, plan to build a national infrastructure for this purpose and the pan-European EOSC initiative has similar objectives.

However, there are remaining questions and challenges. First, using openness as the new yardstick is not a measure of quality, which is emphasised in the EC study referenced throughout this report. The overarching question is how to identify and define research quality in a way that is better than existing indicators. Openness, transparency, and reproducibility/replicability may be elements of that, but probably not the only ones.

Second, responsible use of metrics should remain a prerogative, even when measuring Open Science. Otherwise, the Maxwell effect–measuring a variable will inevitably lead to more of the measured item –may lead to unintended consequences under the guise of Open Science.