OePRESS: an Open ePrint and Rigorous Evaluation System for STEM

Wanpeng Tan

Department of Physics and Astronomy, University of Notre Dame

wtan@nd.edu

March 21, 2025

Abstract

In light of the failed peer review system in basic research, we propose a novel open science initiative: an Open ePrint and Rigorous Evaluation System for STEM (OePRESS). In particular, three different types of research activities — original research, indirect contributions (e.g., participation in the evaluation system), and funding/time requests — should be evaluated using quantitative and continually refined metrics. This new community-based rigorous evaluation system will provide the best incentive for all members by rewarding them with accurate recognition of their innovative achievements and accurate credit for their service and other contributions to the community. The system rewards the quality, not the quantity, of accomplishments. Community members earn credits for their research and other activities in OePRESS, and as they accumulate experience and credits they can advance in their role in the community. High-risk, high-reward research projects will have a better chance of being funded. Eventually, when funding agencies and hiring institutions rely on this rigorous evaluation system to make their decisions, we will see a self-sustaining community of researchers striving for perpetual innovation and development.

Contents

1 Introduction to OePRESS

For basic research in rigorous STEM fields, we need a quantitative peer-review system to rigorously evaluate each individual paper, proposal, and person involved in all phases of research activities. The motivations, detailed discussions, and the sketch of a viable solution to such a system are presented in our recent paper titled “A Robust Community-Based Credit System to Enhance Peer Review in Scientific Research”. We will present more quantitative details of the full scheme of such a system – an Open ePrint and Rigorous Evaluation System for STEM (OePRESS) that is ready to be implemented.

Considering that results from a basic research field typically have less tangible products and much delayed impact, such an evaluation system is critical for continued innovation and success in the field. What's more, the evaluation system should be as rigorous as the field itself. Such a system should involve all members in the research community. Each and every member should be naturally motivated to innovate, to rate, review, and comment on others' contributions, and to serve the community.

The best incentive for all members to participate in these activities in the community is to reward them with accurate recognition of their innovative achievements and accurate credit for their service and other contributions to the community. If funding agencies and hiring institutions eventually rely on this rigorous evaluation system for their decisions, then our basic research community will become a self-sustaining ecosystem striving for perpetual innovation and development.

There are three different types of evaluation to be considered. One is direct evaluation of the innovation itself. The focus here should be on quality instead of quantity. Each original research paper may receive an achievement score (equal to its average rating) that can be allocated to its authors according to the significance of their contributions. Each member of the community can submit up to a fixed number (e.g., five) of scholarly achievements for evaluation. If one achievement comes from a single original research paper, then the achievement level is the same as the author-allocated achievement score. If the submitted achievement is synthesized from a group of closely related papers, then it will be reviewed by the community for eligibility of one level above the highest rating of the individual papers. In the end, an achievement level system will be the best for evaluation of each individual's academic achievements.

The second is evaluation of a member's indirect contributions to innovation including review (e.g., reviews/comments on an article, and standalone review articles), education (e.g., textbooks), outreach (e.g., pop science articles), and other service activities (e.g., moderation and committee/board membership). Although these activities do not directly contribute to scientific innovation, they no doubt expedite the innovation process, get innovative ideas assimilated by the wider community more quickly, and educate future innovators more efficiently. A quantitative credit system to evaluate all these kinds of contributions would be the best way to incentivize members of the research community to participate in such auxiliary yet indispensable activities. This leads to the self-sustaining path for the entire research community.

The third evaluation type involves proposals submitted to funding agencies and user facilities. Two kinds of funding requests need to be evaluated. One is for proposals that request funding for various service/support activities such as coordinating exchange visitor programs, and organizing conferences, summer schools, and other training programs.

The new member rating system may improve upon the current practices by agencies by involving the entire community instead of a limited panel, but we probably won't anticipate a dramatic improvement. However, in evaluating proposals of original research projects that request either financial support from funding agencies or time allocation at user facilities, we may benefit much more significantly from such a system. In this case, two directions involving different measures should be considered. One direction is for support of main-stream research. Again, the new rating system will improve upon current practices, but the resulting effect will not be radically different.

However, the other direction for support of the so-called high-risk high-reward projects (perhaps taking up 5% of the total budget) requires an overhaul of the evaluation system. A minimum scientific standard for credibility should be required for a given research proposal in this category to be considered for support. Once the credibility bar has been passed, the criteria for funding priorities should be based on the potential impact and feasibility or testability of the proposal. To prevent competition from proposals in the main-stream category, a high-risk factor needs to be considered and polarized review reports would be a good telltale sign.

Before introducing the structures and implementation details for OePRESS, we should make a few clarifications. OePRESS is open to all registered members and can also be viewed by the general public. All ratings on original research and other documents are adjustable and could be phased according to historical turning points; Also, all evaluation formulas and/or associated parameters are subject to improvement over time; As such, the evaluation system is dynamic and self-calibrating against potential abuse and fraud.

The new system could start on an existing preprint or eprint archive platform by adding a new layer for rigorous evaluation, or start completely from scratch. It should have its own eprint service but also fully interface with all other existing eprint/archive services. Eventually, OePRESS and the community it serves should strive to become self-motivating, self-evaluating, self-serving, self-calibrating, self-regulating, and self-sustaining.

Many of the arguments, considerations, and measures that were discussed in the earlier paper "A Robust Community-Based Credit System to Enhance Peer Review in Scientific Research" are still valid. However, some specifics, especially formulas and associated parameters, should be superseded by the updates proposed in this paper.

2 Nomenclature and Structures of OePRESS

2.1 STEM branches and community membership

The hierarchical basic research branches of STEM are organized as follows. The top level is disciplines that could initially include mathematics (Math), physics (Phys), chemistry (Chem), life science (Life), medical science (Med), earth science (Earth), computer science (Comp), engineering and technology (Engr), and other uncategorized sciences (Uncat). Its short form is made of typically 3-5 characters with the first letter capitalized. A special top level may be designated for cross-disciplinary fields (Cross). Here only basic research parts of these branches are considered and the directions with tangible or commercializable products, in particular, in Engr, are excluded, as they should be evaluated with different criteria.

The next level is sub-disciplines or fields that are abbreviated as an all-caps acronym. The special "MISC" denotes all uncategorized fields within a discipline. The third level is a subfield or area in terms of topic-methodology in lower case. The full notation of a research branch can then be written in the format of "Discipline.FIELD.area". For example, Phys.NP.nucl-ex represents the discipline of physics, the field of nuclear and particle physics, and the area of experimental nuclear physics. There may always be a designated "other" subfield to include other uncategorized areas within a field. In certain cases, the last two levels (field and area) are optional. There is a special top level "All" with no sub-level branches, which is used to represent all disciplines.

Membership of the research community is applied with a unique open-ID such as ORCID. All members are rigorously evaluated for their achievements in original research and meanwhile earn credits from their indirect contributions to STEM and service to the community. Their achievement levels should be a critical factor in determining their academic positions and funding levels. On the other hand, their credits should largely determine their administration/service roles in the community.

Entry-level members, or commenters, typically still students, may post documents or eprints if they are in good standing and may earn credit by making comments and other contributions to the community. Once members gain enough experience and credit in a subfield, they will be promoted to the rank of reviewer in that subfield. A reviewer can be qualified in multiple subfields and can rate and review, in addition to commenting on, documents submitted in the subfields in which she/he is qualified. The next step in the role, by gaining more experience and credits, is to become a moderator, who can invite members to write reviews and coordinate other efforts in the community. Moderators with the highest credits can take on leadership roles and be elected and serve on the governing board of a discipline (possibly for a fixed term). They may also have other leadership roles in various

other committees. Some committees may also include a small number of reviewers (and even commenters) elected at large for diversity considerations.

2.2 Parent documents

An independent primary document or eprint submitted to OePRESS or other publication services is called a "parent document" with an ID format of "yyyymmdd.xxxx" (also used for DOI) where "xxxx" represents a four-character alphanumeric string in lower case [0-9a-z]. The date here should be the appearing or received date when it was first posted at OePRESS or other eprint/publication services and "xxxx" is determined by the sequential order when it is entered into the OePRESS database. An alternative format could replace "'xxxx" with "[type]-nnnnnn" where "nnnnnn" is the sequential number (up to six figures) within a given type for easier identification at the OePRESS site. The possible document types are listed in Table 1. For example, parent documents in original research can be abbreviated as o-docs; 20240805.000z denotes that it first appeared on Aug. 5, 2024 and it is the 35th overall record entered in the database for that given date, and its alternative ID 20240805.o-12 indicates that it is the 12th original research document recorded by OePRESS for that date. Considering the potential issues associated with explicit dates in IDs, a more pragmatic approach is to employ a 12-digit alphanumeric character string format (i.e., base-36 numbers up to about 4.7×10^{18}). Each parent document is assigned to one primary research branch and can be cross-linked to up to two other branches. Even better, a weighted percentage can be applied to each linked branch to explicitly weigh the connection.

All parent documents can be rated by members (possibly with a reviewer status or higher). Each rating score S should be a floating point number with one digit after the decimal point and range from -1.0 to 5.0. The average rating score of each parent document is typically written as $\bar{S}$ or $\bar{S}_p$. Such scores for f/p/s/t-docs are used for funding or time allocation decisions. The scores of the other types of parent documents determine two types of merits: either achievement level for original research or member credit for other contributions and service in the community.

For o-docs, the average score represents the achievement level of the document, denoted by $\bar{A} = \bar{S}$. The rating standards and six achievement classes are detailed in Table 2. Revolutionary (A_{+4}) means among the best in a discipline or even across disciplines, worthy of a top prize (e.g., the Nobel prize); Disruptive (A_{+3}) means among the best in a sub-discipline or field, worthy of a good prize in a field; Innovative (A_{+2}) means among the best in a subfield or area, worthy of publication in a top journal (e.g., PRL in physics); Incremental (A_{+1}) means normal incremental research; Positive (A_+) means all the above; Useless (A_0) means incorrect or (nearly) useless for any other research; Detrimental (A_-)

Table 1: Types of parent documents and their single-letter abbreviations are listed.

id-format	abbrev.	parent document type
o-nnnn	o-docs	original research documents
r-nnnn	r-docs	review, expository, and survey articles
a-nnnn	a-docs	synthesized achievements for evaluation
b-nnnn	b-docs	textbooks, lectures, and other non-original research books
c-nnnn	c-docs	conference proceedings or announcements (+ schedule/program) for conferences, meetings, workshops, and summer schools
d-nnnn	d-docs	detailed discussions on issues in the community
e-nnnn	e-docs	pop-science essays for a general audience
f-nnnn	f-docs	applications for funding facility-oriented programs
p-nnnn	p-docs	applications for funding project-oriented original research
q-nnnn	q-docs	applications for funding R&D equipment and technology
s-nnnn	s-docs	applications for funding service/support activities (meetings, etc.)
t-nnnn	t-docs	applications for time allocation at research facilities
u-nnnn	u-docs	open source scientific softwares and codes for end users
x-nnnn	u-docs	open experimental equipment for end users
v-nnnn	v-docs	voting for election and awards
w-nnnn	w-docs	white papers and other official reports from official committees
i-nnnn	i-docs	issues or bugs in OePRESS
n-nnnn	n-docs	new suggestions for improving OePRESS
m-nnnn	m-docs	moderation decisions and committee reports in OePRESS

means fraudulent, plagiaristic, or harmful to other research. A typical rating score in each category is the half-integer within its own range and can vary depending on the quality. Most of credible research papers should fall in the score range of 1.0 – 3.0. To protect potentially disruptive ideas that may not be fully understood by the community at the moment, an o-doc that looks like crackpot should not typically be rated below -0.1, even if one thinks that it is full of crackpot ideas. The range between -1 and -0.1 should be reserved for the class of truly detrimental documents.

Table 2: Achievement level classes are shown with their corresponding score ranges. A special category "Positive" (A_+) includes all solid works (Incremental and higher levels) with $1 \leq \bar{A} \leq 5$.

Detrimental	Useless	Incremental	Innovative	Disruptive	Revolutionary
A_-	A_0	A_{+1}	A_{+2}	A_{+3}	A_{+4}
$-1 \leq \bar{A} < 0$	$0 \leq \bar{A} < 1$	$1 \leq \bar{A} < 2$	$2 \leq \bar{A} < 3$	$3 \leq \bar{A} < 4$	$4 \leq \bar{A} < 5$

In an a-doc submitted by a member, the synthesized achievement refers to a group of orig-

inal research documents (o-docs) that should be closely related to the same achievement topic for evaluation. The rating scheme should make sure that a synthesized achievement level is no more than one full level above the best level of individual o-docs. Eventually, a robust formula may be needed to automatically calculate the synthesized level from the related o-docs as discussed later in Sect. 4.

For the remaining parent documents, an earned credit point (ECP) system is applied. In these cases, each rating score should be submitted according to the standards listed in Table 3. Excellent means full of deep insights or excellent contributions; Very good means some deep insights or very good contributions; Good means insightful or significant contributions; Fair means modest or useful contributions; Useless means unnecessary or futile contributions; Detrimental is reserved for abusive, fraudulent, and plagiaristic documents. Again, the midrange score of each category is typical and most of creditable contributions should be in the range of 1.0 – 3.0. Details of the ECP system will be discussed in the next section.

Table 3: Standards for rating reviews, comments, and non-original research documents are shown with their corresponding rating ranges.

Detrimental	Useless	Fair	Good	Very Good	Excellent
$-1 \leq S < 0$	$0 \leq S < 1$	$1 \leq S < 2$	$2 \leq S < 3$	$3 \leq S < 4$	$4 \leq S \leq 5$

2.3 Author hierarchical schemes

Authors of o-docs and other credit-earning documents are typically divided into four classes: 1st-authors, 2nd-authors, 3rd-authors, and general-authors, with their merits (achievement level or ECP) reduced accordingly in that order. ECP of each eligible contribution should be conserved when distributed among its authors. However, in order to encourage collaborations, the achievement level of original research is shared among its authors and therefore the allocation is not conserved. As a result, authors in the first three classes for an o-doc should be limited to a fixed number (e.g., five).

Another scheme by combining 2nd-class and 3rd-class into a variable class is to assume that the significance level from each of the 1st-authors is 100% and the significance level from each author in the variable class is normalized in percentage accordingly. There is no explicit limit in the number of 1st-authors or variable-authors, but the sum of their significance percentages should be limited to a fixed percent (e.g., 875%). It is up to the authors to decide if they choose the four-class scheme or the variable scheme to report their individual contributions to the article.

2.4 Derived contributions and rating principles

All the other documents including comments and reviews directly and indirectly (via other comments and reviews) on a parent document are called “derived documents” and can be rated according to Table 3. The derived document ID format has three more alphanumeric characters affixed to the parent document ID like “yyyymmdd.xxxx.xxx” that could also be used as part of DOI. The three-character ID “xxx” also represents the sequential order for all derived documents under the same parent document.

There are four types of derived documents: invited reviews [i], contributed reviews [r], creditable comments [c], and uncreditable comments [u], where the letter in square brackets can be used as a prefix for its type. A minimum number of characters or bytes for UTF-8 encoding (e.g., >1000) should be required to post an invited or contributed review. The uncreditable comment type cannot earn credit, and is typically either too short (e.g., less than 64 characters or bytes) or not intended for earning credit.

Another way to identify these derived documents at OePRESS is to prefix their numerical ID with one letter indicating their type such as “[i/r/c/u]-nnnn” for reviews and comments, where the ID “nnnn” (up to four digits) also denotes the numerical order within its own type. All these derived creditable documents and all ratings are called derived contributions and their ECP calculations are discussed in the next section.

The principle of democracy requires that each member’s rating should be weighted exactly the same when it is used to calculate the average score, although ratings themselves can receive different credit points and ECP should generally be weighted with various factors. One can rate the same item up to 3 times to allow for a change of mind, but only the last rating will be used to calculate the average score for that item, although all rating attempts will be credited.

Two ratings on the same item from the same member must be separated by at least a certain amount of time (e.g., one week) and be significantly different (e.g., >0.9) to prevent gaming. To prevent abuse, the number of ratings associated with no creditable contributions for a given rater should be limited, e.g., no more than five times the number of those associated with creditable contributions.

2.5 Overview of community structures

Therefore, all members in the community are evaluated based on two types of merits. Their innovation accomplishments are determined by the achievement scores that are allocated to them in their top-rated original research works and can be denoted by “Ln(m)” or “Ln/m” where n is the highest achievement class [0-4] (or possibly the highest achievement score) that the member has ever reached and m is the number of evaluated achievements

in that class. For a member who has submitted more negative (A_-) o-docs than positive (A_+) o-docs by a number of a, the member's achievement level should be labeled by "L$-a$".

Their member roles as commenters (C), reviewers (R), moderators (M), leaders (L) in the community are mostly determined by their accumulated ECP. Therefore, a combined denotation for a member's status could be like "Ln/m[Role]" where [Role] may be a single letter of C, R, M, or L. For members with negative credits, [Role] is replaced with N. Table 4 shows the hierarchical structures of member roles and achievement levels with roughly matched traditional positions in the community and suggested basic annual funding levels. Such basic funding support, as discussed later in Sect. 5, is considered an investment on free exploratory research, as opposed to funding for specific research projects proposed in p-docs.

Table 4: Achievement class levels (L0-L4) of researchers are aligned with suggested basic annual funding levels, and roughly matched with their positions and roles in the community.

Level	≈Position	Achievements	Funding	≈Role
L0	Student	N/A	N/A	commenters
L1	Postdoc	A_{+1}	\$3-5k/yr	reviewers
L2/1	Fellow/Lecturer	$1 \times A_{+2}$	\$10k/yr	
L2/2	Assist. Prof.	$2 \times A_{+2}$	\$15k/yr	moderators
L2/3	Assoc. Prof.	$3 \times A_{+2}$	\$20k/yr	leaders
L3	Prof.	A_{+3}	\$50-150k/yr	
L4	Chair Prof.	A_{+4}	\$0.3-1.5M/yr	

3 Earned Credit Point System

3.1 ECP for parent documents

The earned credit point (ECP) system is designed for evaluating most of service and support activities that are not directly related to original innovation research. It particularly encourages members to participate in the evaluation of OePRESS. Ideally a member's ECP will be the most critical factor for determining their service and leadership roles in the community. In one way, ECP can be earned directly from most parent documents as follows,

$$\text{ECP(parent)} = a(\sinh(b(\bar{S} - c)) + d) \tag{1}$$

where

- $\bar{S}$ is the average rating score that the parent document has received in the range between -1.0 and 5.0,
- a is the parent-doc-type dependent normalization factor that should ensure an upper limit on ECP≤ 2500,
- b is the skew factor biased toward highly-rated documents (the higher b the more biased, e.g., $b = 1.6$ favors $S > 3$),
- c gives the symmetry line of the ECP curve (documents are encouraged above this line, e.g., $c = 1.5$ means the middle of fair / incremental),
- d is the offset calculated by setting zero ECP for a certain average score $\bar{S}_0$ (e.g., $\bar{S}_0 = 1$ means that negative ECP is incurred when $S < 1$).

For review-type parent documents, we can adopt the following formula,

$$\text{ECP(r-docs)} = 18(\sinh(1.6(\bar{S} - 1.5)) + 0.888) \tag{2}$$

for $\bar{S}_0 = 1$ and an ECP range of roughly between -500 and 2500 as shown in Fig. 1. We may set $a = 0$ or no ECP for a-docs since their sole purpose is for achievement evaluation and also for f/p/s/t-docs since their success will be awarded by funding support or time allocation at user facilities. For o-docs, we may reduce it by a factor of 10, i.e., setting $a = 1.8$. For the remaining types, we may set a reduction factor somewhere in between 1 and 10 (e.g., the option of $a = 9$ in Fig. 1).

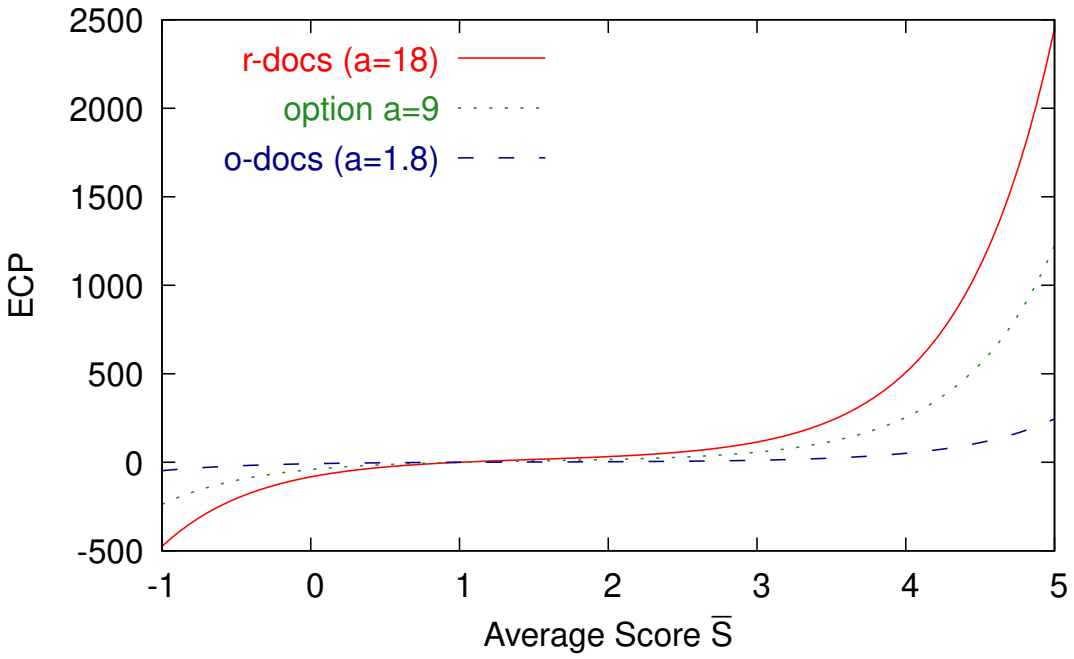

Figure 1: ECP as a function of the average rating score $\bar{S}$ for various parent documents.

3.2 ECP distributions among authors

ECP of a parent document should be distributed among all the authors according to the significance of their contributions,

$$\text{ECP(i-th author)} = \text{Total-ECP} \times \frac{p_i}{\sum_j p_j} \tag{3}$$

where p_j is the significance percentage of the j-th author normalized to that of each 1st-author (assumed to be 100%) in the variable-class author scheme. In the four-class author scheme, the points that each of 1st-authors receives can also be consistently calculated as,

$$\text{ECP(1st-author)} = \text{Total-ECP}/(n_1 + n_2/2 + n_3/4 + n_4/10) \tag{4}$$

where $n_{1,2,3,4}$ are the number of authors in each class and authors in a lower class receive half of the points of those in a class one level higher and general-authors receive 10% of the points of 1st-authors.

3.3 ECP for derived contributions

More commonly, ECP is earned through contributions, including ratings, comments, and (contributed / invited) reviews directly or indirectly (via other comments / reviews) on the parent document. Each of these derived contributions can receive ECP as follows,

$$\text{ECP(derived)} = f_o f_n f_a(\bar{S}_p) f_{eb}(N) f_t(T) f_d(\text{d-type}) f_p(\text{p-type}) \times \text{Raw-ECP} \tag{5}$$

where

- f_o is the openness factor (1 if posted with open-ID; 1/2 if posted semi-anonymously even if later with its associated ID revealed; 1/4 if posted completely anonymously and eventually connected to an account),
- f_n is the normalization factor between different disciplines and fields (possibly areas as well),
- f_a is the attention factor as a function of the average score of the parent document $\bar{S}_p$, as shown in Fig. 2 and Eq. 6 below,
- $f_{eb}(N)$ is the early bird factor for this N-th contribution within a given type of ratings/comments/reviews as shown in Fig. 2 and Eqs. 7-8 below,
- $f_t(T)$ is the time factor for the lapse of time T since the submission time of the parent document or the reset time of a new evaluation phase as shown in Fig. 2 and Eq. 9 below,
- $f_d(\text{d-type})$ is the discriminative rating factor that accounts for different types of derived contributions as shown in Fig. 2 and Eq. 10 below,

- f_p(p-type) is the factor that accounts for different types of parent documents; initially it can be set to 1 for o/p-docs, 1/2 for r/f/s/t-docs, and 1/4 for the rest,
- Raw-ECP is equal to $\bar{S} - 0.5$ for a creditable comment with an average rating $\bar{S}$; or its formula is similar to Eq. 2 with $a = 0.6$ for an invited review and $a = 0.4$ for a contributed review as shown in Fig. 3; or it is equal to $1-2|S-\bar{S}|/3$ for a rating with its submitted score S against the current average $\bar{S}$; see below for further discussions.

More details of the factors in Eq. 5 are elaborated below. The normalization factor f_n can initially be set to 1 for all fields and then it may vary as members in different fields may present a significant disparity in their rating behaviors. The alignment for normalization between different fields could fall on the average or, better yet, median ECP of members at reviewer or higher roles. Further considerations may be needed to account for the shape variations in ECP distributions among members in different fields.

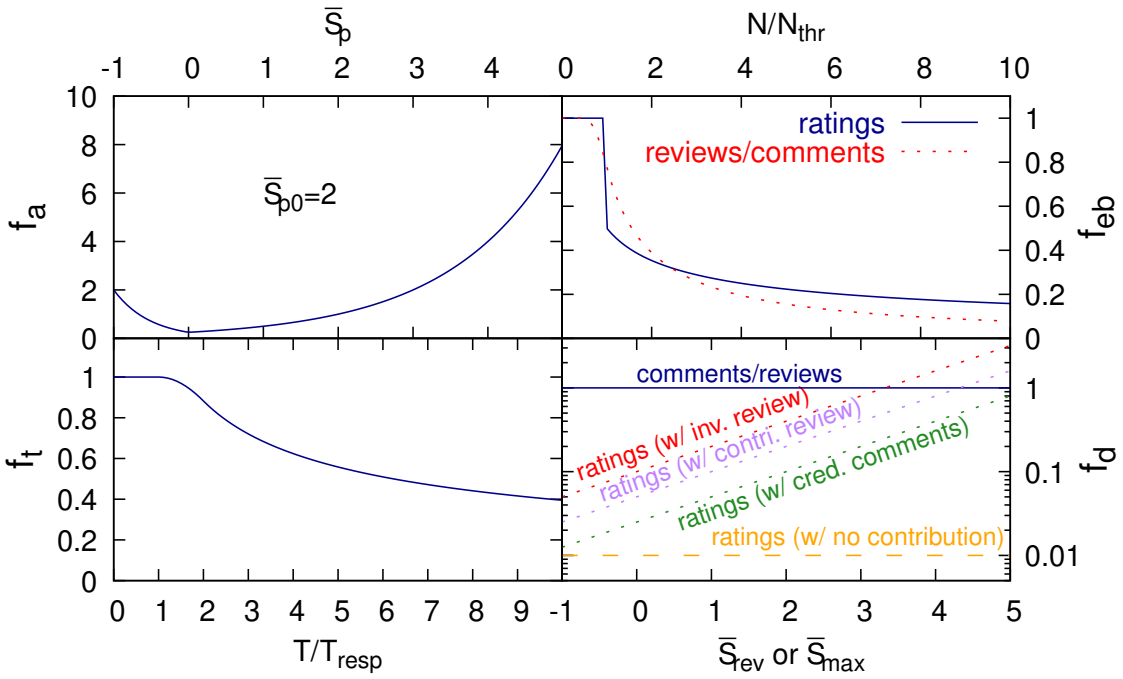

Figure 2: Various factors introduced in Eq. 5 are illustrated.

The attention factor f_a, as shown in Fig. 2, can be parameterized as,

$$f_a(\bar{S}_p) = \begin{cases} 2^{3|\bar{S}_p| - \bar{S_{p0}}}, & \text{for } \bar{S}_p < 0 \\ 2^{\bar{S}_p - \bar{S_{p0}}}, & \text{for } \bar{S}_p \geq 0 \end{cases} \tag{6}$$

where $\bar{S}_p$ is the average score of the parent document and the threshold score value $\bar{S_{p0}} = 2$ is used to scale the maximum possible ECP to be about 660 for an invited review.

The early bird factor f_{eb}, also shown in Fig. 2, can be parameterized for the N-th rating

as,

$$f_{eb}(N) = \begin{cases} 1, & \text{for } N \le N_{\text{thr}} \\ \frac{1}{2}\sqrt{N_{\text{thr}}/N}, & \text{for } N > N_{\text{thr}} \end{cases} \tag{7}$$

where the factor of 1/2 accounts for the practice that the average rating will not be revealed until number of the ratings has reached the threshold number $N_{\text{thr}} = 5$. For the N-th comment or review, it is parameterized as,

$$f_{eb}(N) = \begin{cases} 1, & \text{for } N \le N_{\text{thr}}/2 \\ \exp(-(N/N_{\text{thr}} - 0.5)^2), & \text{for } N_{\text{thr}}/2 < N \le N_{\text{thr}} \\ e^{-0.25} N_{\text{thr}}/N, & \text{for } N > N_{\text{thr}} \end{cases} \tag{8}$$

where the threshold number N_{thr} may be set to 6, 10, 50 for invited reviews, contributed reviews, and comments, respectively.

The time factor f_t, as shown in Fig. 2, can be parameterized as a function of the lapse of time T since the submission time of the parent document or the reset time of a new evaluation phase as,

$$f_t(T) = \begin{cases} 1, & \text{for } T \le T_{\text{resp}} \\ \exp(-(T/T_{\text{resp}} - 1)^2/8), & \text{for } T_{\text{resp}} < T \le 2T_{\text{resp}} \\ e^{-0.125}\sqrt{2T_{\text{resp}}/T} \approx 1.248\sqrt{T_{\text{resp}}/T}, & \text{for } T > 2T_{\text{resp}} \end{cases} \tag{9}$$

where T_{resp} is the possibly field-dependent response time (e.g., typically about two weeks for a 20-page document and additional one week per 10 more pages). This factor is similar to the early bird factor for a rating (without the 1/2 factor owing to publicity).

The discriminative rating factor for different types of derived contributions, as shown in Fig. 2, is presented here,

$$f_d(\text{d-type}) = \begin{cases} 1, & \text{for a creditable comment/review or negative Raw-ECP} \\ 0.4 f_r \times 2^{\bar{S}_{\text{rev}}-2}, & \text{for a rating (rater made an invited review)} \\ 0.2 f_r \times 2^{\bar{S}_{\text{rev}}-2}, & \text{for a rating (rater made a contributed review)} \\ 0.1 f_r \times 2^{\bar{S}_{\text{max}}-2}, & \text{for a rating (rater made creditable comments)} \\ 0.01 f_r, & \text{for a rating (rater made no creditable contributions)} \end{cases} \tag{10}$$

where f_r is 2 if the rating is on the parent document or 1 otherwise, $\bar{S}_{\text{max}}$ is the highest average score of the rater's comments under the same parent document, $\bar{S}_{\text{rev}}$ is the average score of the rater's review on the same parent document. If such a comment or review involves multiple authors, then its rating allocation to each author may be done in the same way as in the achievement level system discussed in Sect. 4.

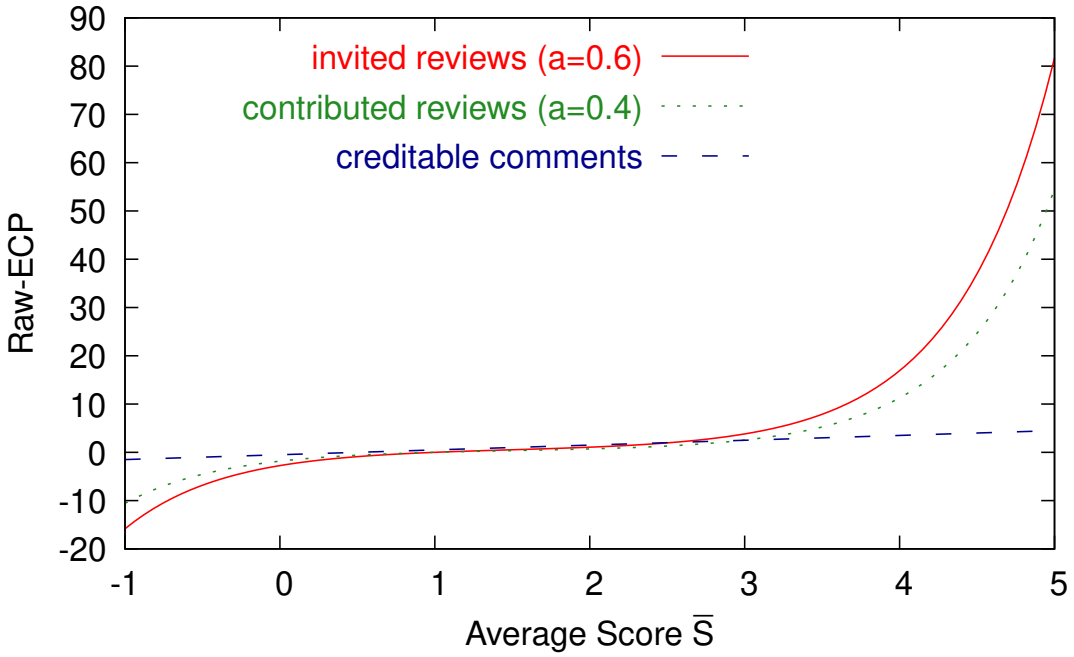

Figure 3: Raw-ECP as a function of the average rating score $\bar{S}$ for derived documents.

The principles behind the Raw-ECP formulas are as follows. Bad contributions must receive negative credits. A comment has to be better than $\bar{S} > 0.5$ to be positively credited and a simple linear credit curve is sufficient. The number of reviews should be limited and only good ones are encouraged so that the bias toward highly-rated ones is warranted. The parameter $a = 0.4$ means that a member should be confident that her/his contribution will receive very good ratings ($\bar{S} > 3$) to be better credited as a contributed review than a comment. For an invited review, it is a little more lenient and the break point occurs at about $\bar{S} = 2.5$ for $a = 0.6$. If a rating deviates from the average by more than 1.5, it will receive negative credit points to discourage malicious intentions and behaviors.

Such derived documents are typically written by a single author who takes the full ECP. If more authors are involved, the total ECP should be conserved and distributed to all the authors according to the above-discussed author schemes.

3.4 Invited reviews and ECP tagging

The moderator who invited a review may receive 10% of the credit points earned by the review. However, no such additional credits can be earned if a moderator invited herself/himself to write the review. No more invitation can be sent to the same reviewer if the reviewer has already received an invitation. So only one moderator will be credited for an invited review.

Before a contributed review has received the threshold number of ratings (i.e., its $\bar{S}$ is not yet public), a moderator can convert it to an invited review and also get 6% of the ECP if the number of invited reviews has not reached the maximum (e.g., 10). Similarly, the moderator can get 3% of the ECP if $\bar{S}$ is public but the conversion is still within the response time T_{resp}, and 1% otherwise.

A moderator can invite a maximum of five reviews. No more than 10 reviews can be invited in the first round. If five or more are posted within the response time T_{resp}, then no more invitations should be sent. Otherwise, a second round of invitations can be sent to make up the shortfall. If an invitee requests an extension and the inviting moderator grants it, the invitation will remain active for the agreed-upon extended period (up to another period of T_{resp}). If an invited review is posted after T_{resp} or the granted extension, it will be automatically treated as a contributed review. No more than 20 contributed reviews can be submitted. A member can only post one review (either invited or contributed) for the same parent document.

All parent documents and derived contributions are associated with a primary research branch and possibly up to two cross-linked branches. Therefore, all earned credits can be tagged with research branches and parent document types. If two branches are associated, then the primary branch takes 2/3 of the credit and the cross-linked one takes 1/3. If three branches are associated, then the primary branch takes 50% of the ECP and each cross-linked one takes 25%. Such tagging information on ECP will be helpful for certain decision processes, e.g., when a certain amount of credits in a research branch or from a certain type of activities is required for a position.

3.5 Anonymity and ECP

All members are associated with a single open ID such as ORCID and no duplicates are allowed. Their open contributions will receive full credit. Anonymity is not allowed for parent documents. However, members may contribute semi-anonymously and receive half of the credits for their derived contributions. In this case, a member uses her/his own account to post, so the contribution is linked to the account internally, but the public appearance shows that it is anonymous. The member has the option to reveal the association later, but no extra credits will be recovered. Complete anonymity is the last option for a comment, but not for a rating or review. This should be the last resort for a member who is afraid of retaliation or the like to have the chance to comment unofficially. In this case, the member must provide an email address or phone number to receive an identification code and the comment will be strictly moderated. The member has the option to use the code to link the comment back to her/his account and receive 1/4 of the credits.

4 Achievement Level System

4.1 Achievement level allocation from single o-doc

The achievement score $\bar{A}$ of an o-doc is simply the average of all its ratings: like other ratings for ECP and proposals, each member rating is counted with the same weight; but only the last rating is counted if a member has rated multiple times. In the four-class or variable-class author scheme, 1st-authors are all critical and most responsible for the results and hence should share the same $\bar{A}$ of the article. They all have equal ownership of the paper, and can change critical meta-data about the paper (e.g., notes about the article, other updates, and even retraction).

In general, the allocation of achievement level to the i-th author in an o-doc as shown in Fig. 4 is given by,

$$A_i = \mathsf{min}(\mathsf{max}(\bar{A} + \ln p_i,\ p_i\bar{A}),\ \mathsf{CAP} \equiv 5 + \ln p_i / \ln 2) \tag{11}$$

where p_i is the significance percentage of the i-th author normalized to that of each 1st-author (assumed to be 100%). For the 2nd class, $p_i = 50\%$; for the 3rd class, $p_i = 25\%$; for the general class, $p_i = 10\%$; for the variable class, $20\% \le p_i < 100\%$. However, the sum of $\sum_i p_i$ for all non-general authors must be less than 875% to be consistent with the four-class scheme. The last term in Eq. 11 is used to set the ceiling for authors not in the 1st-class according to their p_i values. In particular, the maximum value of A_i allocated to a 2nd-class author is 4, and it is 3 for a 3rd-class author. And a general-author's score A_i is capped at 1.678.

Eq. 11 is introduced mainly for the variable author-class scheme but it is also compatible and consistent with the four-class scheme. In particular, 2nd-authors are those who have made major contributions and thus should share an achievement level of the lesser of 4 and $\bar{A} - \ln 2 \approx \bar{A} - 0.693$ if $\bar{A} > 2 \ln 2 \approx 1.386$ or $\bar{A}/2$ otherwise (in the sense that they bear half the responsibility for bad research). 3rd-authors should have made significant contributions and thus share an achievement level of the lesser of 3 and $\bar{A} - 2\ln 2 \approx \bar{A} - 1.386$ if $\bar{A} > 8 \ln 2/3 \approx 1.848$ or $\bar{A}/4$ otherwise. General-authors share an achievement level of the lesser of 1.678 and $\bar{A} - \ln 10 \approx \bar{A} - 2.303$ if $\bar{A} > \ln 10/0.9 \approx 2.558$ or $\bar{A}/10$ otherwise. This non-conservation in merit allocation is intended to encourage collaborations. To avoid abuse, the number of authors in the first three classes should be limited (e.g., up to five for each class), or better yet, the sum of their significance percentages should be capped as discussed above, and the highest achievement level of non-1st-class authors should be limited properly as well.

The principle behind Eq. 11 is to treat the significance of high-level achievements exponentially and that of low-level ones ($\bar{A} \lesssim 1$–2) linearly. The corner introduced in such a

transition in Eq. 11 can be removed by imposing a smooth transition. One possible way is to consider that the two segments of a unit length starting from the corner are tangent lines to a circle and the arc between the two touching points provides the smooth connection for Eq. 11.

A more practical solution is to use quadratic Bézier curves to connect smoothly as shown in Fig. 4 (dotted lines). Specifically, the three Bézier points can be chosen from the corner and two points at $\bar{A}(\text{corner}) - 0.5$ and $\bar{A}(\text{corner}) + 0.5$, respectively, and therefore the connection equations are,

$$A_i(t, p_i) = -\frac{p_i \ln p_i}{1 - p_i} - 0.5p_i(1 - t)^2 + 0.5t^2 \quad (12)$$

$$t = \bar{A} + \frac{\ln p_i}{1 - p_i} + 0.5 \quad (13)$$

where t is the standard parameter in the range of $0 < t < 1$. Similar equations can also be obtained for the smooth connection to the upper limit of a non-1st-class author's achievement allocation,

$$A_i(t) = \text{CAP} - \Delta(1 - t)^2 \quad (14)$$

$$t = \frac{\bar{A} + \ln p_i - \text{CAP} + \Delta}{2\Delta} \quad (15)$$

$$\Delta = \min(\ln p_i - \ln p_i / \ln 2,\ 0.5), \quad (16)$$

where CAP is the upper limit defined in Eq. 11 and Δ defines the transition width to the upper limit.

4.2 Member achievement levels and synthesized achievements

Each member's achievement level is determined by a maximum number (e.g., five) of the member's top-rated achievements. As a possible simple approach, only achievements in the same top level class as shown in Table 2 may be considered. For example, if a member's five top achievements have author-allocated achievement scores of 2.7, 2.5, 2.1, 2.0 and 1.8, then the member is at the achievement level labeled by $L2/4$ or possibly by $L2.7/4$ to explicitly show the top score. Typically, an entry-level member (e.g., a student) has no original work with a rating above one (i.e., no A_+ o-docs), so the member is at the level $L0$. All possible achievement levels for members are listed here: $L-$, $L0$, $L1/1$ – $L1/5$, $L2/1$ – $L2/5$, $L3/1$ – $L3/5$, and $L4/1$ – $L4/5$. In particular, the level $L-$ may be assigned to a member if the member has more A_- o-docs than A_+ o-docs.

However, the above method may be too rough and a more quantitative and rigorous evaluation approach is needed. As a starting point, assuming that the the achievement allocation set $\{A_j,\ 1 \le j \le n\}$ of n achievements (n_+ of them positive) for a given researcher

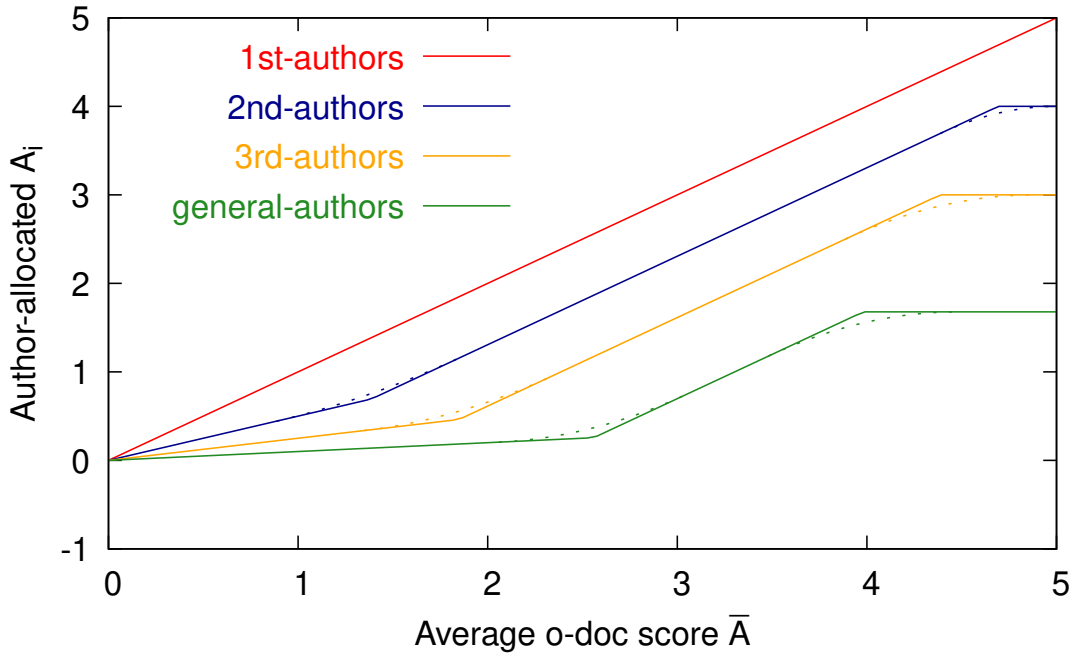

Figure 4: The allocation of achievement level to the i-th author in an o-doc as a function of the average rating score $\bar{A}$ for the original document. The dotted lines are smooth connection using quadratic Bézier curves within ± 0.5 in $\bar{A}$ from the corner.

is arranged in the descending order, a possible formula for the researcher's final achievement score can be proposed as follows,

$$\begin{aligned}
\alpha_+ &= \begin{cases} \sum_{j=1}^{n_+}(6^{A_j}-1), & \text{for } n_+ \le 5 \\ \sum_{j=1}^{5}(6^{A_j}-1) + \sum_{j=6}^{n_+}(6^{A_j}-6^{2.5})\theta(A_j-2.5), & \text{for } n_+ > 5 \end{cases} \\
\alpha_- &= -\sum_{j=1}^{n}(6^{-3A_j}-1)\theta(-A_j) \\
\alpha &= \alpha_+ + \alpha_-, \qquad (17)
\end{aligned}$$

where the total achievement score α is the sum of both positive (α_+) and negative (α_-) parts. The θ function is the unit step function for consideration of more than five positive and negative achievement cases. The base value of 6 is chosen to ensure that quality should trump quantity since five achievements of A will never score higher than one single achievement of $A+1$. The second term in α_+ takes into account cases when a researcher has more than five achievements with $A > 2.5$ where the level 2.5 can be regarded as the critical point of good quality. This emphasizes that quantity only matters when it is of good quality. The exponential factor 3 in the equation of α_- is chosen because fraudulent publications typically target an achievement level $L3$ instead of $L5$.

The scheme following Eq. 17 matches the above rough scheme very well if we equate

$L1/1$ with $\alpha \geq 6^{1.5} - 1 \approx 13.7$, $L2/1$ with $\alpha \geq 6^{2.5} - 1 \approx 87.2$, $L3/1$ with $\alpha \geq 6^{3.5} - 1 \approx 528$, $L4/1$ with $\alpha \geq 6^{4.5} - 1 \approx 3174$, etc. Accordingly, the sub-level m in Ln/m means m achievements equivalent of Ln, and hence we can equate Ln/m with $\alpha \geq m(6^{n+0.5} - 1)$ in general. Lastly, $L\_$ can be assigned if $\alpha < 0$.

An achievement can be counted and automatically evaluated from a single original research document (o-doc) using the achievement level allocated to the member. A synthesized achievement can also be submitted by a member for evaluation as an a-doc, referring to a group of closely related o-docs by the member. We may adopt a similar rating system, especially in the beginning of OePRESS, i.e., the rating score (S_a) range is also between -1.0 and 5.0 but its meaning is to what extent the synthesized achievement is really up to one full level above its best individual o-doc as elaborated in Table 5. The synthesized achievement level will be the allocated achievement level of the best-rated individual o-doc A_{max} + $\bar{S}_a/5$. This means that a synthesized achievement level can not be more than one full level higher than the best level of individual documents.

Table 5: Standards for rating a synthesized achievement submission with the synthesized level calculated as $A_{\text{max}} + \bar{S}_a/5$.

Abusive	No gain	additional $S_a/5$ to A_{max}	Definitely ($A_{\text{max}} + 1$)
$-1 \leq S_a < 0$	$S_a = 0$	$0 < S_a < 5$	$S_a = 5$

Upon the experience of the above manual rating system, we may develop a more robust and consistent way to automatically calculate the synthesized achievement level from a group of o-docs like the following,

$$A_{\text{syn}} = A_{\text{max}} + \frac{(\sum_j A_j^k)^{1/k}}{A_{\text{max}}} - 1 \tag{18}$$

where k may take the value of 6 to be a little conservative for at most 64 o-docs to be considered. The submitter of an a-doc may also include relevant r-docs for review. In such cases, the achievement score of each r-doc is calculated as,

$$A_i = \text{min}(A_{\text{max}}, S_i - 1). \tag{19}$$

where S_i is the average score of the r-doc that is allocated to the submitter. Note that A_i as calculated above can not be used separately and can only be used for evaluation of a synthesized achievement.

4.3 Further clarifications

The o-docs used in a synthesized achievement can not be used again either in another synthesized achievement or as a single o-doc achievement in order to prevent double

counting. If the number of submitted achievements (a-docs) is less than the maximum allowed (e.g., five) for a member, then the remaining individual o-docs with the highest allocated ratings will be automatically counted toward the member's achievement level. Therefore, only synthesized achievements need to be submitted.

In addition, no A_0 and A_- documents should be considered for synthesized achievements. Each of A_{+4} documents is so impressive on its own that it doesn't need to be synthesized for further evaluation. Therefore, an a-doc can only refer to o-docs of the caliber of A_{+1}, A_{+2}, and A_{+3}.

The achievement evaluation system is intentionally a bit more tolerant than the ECP system. In particular, $0 < S < 1$ is considered "Useless" at the moment, which may include unorthodox ideas that later turn out to be transformative, and therefore are not penalized in achievement evaluation. Contrarily, the same range with the same meaning is associated with a penalty of negative credits in the ECP system.

5 Funding and Hiring Processes

5.1 Funding programs

Funding support may be categorized into four major programs: service-oriented, facility-oriented, individual-oriented, and project-oriented. The first two programs fund service- and infrastructure-type activities as indirect contributions to innovation while the last two directly support original research. The service-oriented program evaluates proposals that request funding for various service or support activities including exchange visitor programs, and the organization of conferences, workshops, summer schools, and other training programs. The facility-oriented program may fund the establishment of large collaboration networks (especially cross-disciplinary), user facilities, and other collaborative centers. The funding agencies have been doing a decent job in the first two categories. The improvements in the new evaluation system may increase the openness of the process and involve the entire community rather than a limited panel. For a given proposal, a funding agency can easily invite referees from the existing pool of reviewers and moderators if they wish. They can also open it up as an f/s-doc to all members to solicit open comments and reviews from the entire community. Or do both.

The individual-oriented funding program that can be implemented under the evaluation system of OePRESS should reward individuals based on their achievement levels. Members at level $L1$ or higher who are affiliated with an academic or research institution can apply with minimal paperwork for basic research funding from national funding agencies such as NSF and NIH. Affiliation may be extended to other types of non-profit organi-

zations as long as such organizations are qualified to receive national funding and can properly handle taxes. This funding support is intended for unrestricted exploratory or discretionary research in basic STEM fields. It can be used to cover costs such as supplies, service fees, and travel for meetings and collaboration activities. Such basic funding support ensures that each eligible member in the research community should have some decent independence in conducting original research, much like start-up companies in the business world. The funded researchers are free to use the funds at their own discretion (e.g., to supplement support of an existing project of interest, or to pursue risky ideas).

Table 6: Funding levels for basic annual research support are suggested for individual researchers at different achievement levels.

	Number of Achievements or sub-Level at the Given Level				
Level	/1	/2	/3	/4	/5
L1	\$3k/yr	\$3.5k/yr	\$4k/yr	\$4.5k/yr	\$5k/yr
L2	\$10k/yr	\$15k/yr	\$20k/yr	\$25k/yr	\$30k/yr
L3	\$50k/yr	\$90k/yr	\$130k/yr	\$170k/yr	\$210k/yr
L4	\$300k/yr	\$600k/yr	\$900k/yr	\$1.2M/yr	\$1.5M/yr

As a simple rough approach, the basic funding levels are suggested in Table 6 for individual researchers at different achievement levels. The $L1$ funding level starts at \$3k/yr with increments of \$500 for each additional A_{+1} achievement and up to \$5k/yr. The $L2$ funding level starts at \$10k/yr with increments of \$5k for each additional A_{+2} achievement and up to \$30k/yr. The $L3$ funding level starts at \$50k/yr with increments of \$40k for each additional A_{+3} achievement and up to \$210k/yr. The $L4$ funding level starts at \$300k/yr with increments of \$300k for each additional A_{+4} achievement and up to \$1500k/yr. Obviously, this is just a suggestion for the funding levels in the United States and other countries can tailor it to fit their own fiscal budgets.

A more delicate funding scheme may take into account formulas similar to Eq. 17, that is,

$$\text{Annual Funding Amount } = f \times \alpha \tag{20}$$

where the adjustable factor $f = 100$ dollars will match the funding levels of the rough scheme proposed in Table 6 very well. The only extra condition to be considered would be the startup threshold, that is, a researcher's α score has to be high enough to pass the minimum funding level (e.g., \$3k) to be funded.

Such individual-oriented funding support is probably sufficient for theorists engaged in original research. However, for many experimental projects on original research, it is not adequate. The parent documents p-docs are intended for these more costly project-

oriented proposals. Such projects must be directly related to original research and must have very specific goals or aim to solve concrete problems.

There are roughly two types of projects that should be considered for funding support: main-stream research projects and high-risk high-reward projects. Again, funding agencies have been doing well in funding main-stream projects as discussed above on service- and facility-oriented programs. OePRESS can also improve on this in terms of openness and community involvement.

5.2 Funding high-risk high-reward projects

However, funding agencies have largely failed to support paradigm-shifting ideas, or in other words, high-risk high-reward projects. This funding area can really benefit from OePRESS. Ideally, national agencies should earmark a certain amount of funds (e.g., 5% of the total project-oriented budget for original research) for such projects. To select such worthy projects, we may need to implement a more reliable rating system for p-docs. First, each rater should give an overall rating score for a p-doc, as for other parent documents, which will be used to select main-stream projects for funding support.

In addition, three other scores should be provided on scientific soundness, potential impact if successful, and technical feasibility or testability. These scores should be used to identify high-risk high-reward projects. The score on scientific soundness can be used for setting a minimum scientific standard for credibility. Such a requirement for scientific standards could also be met (especially in the beginning of OePRESS) with 1-3 positive consultative/applicant-selected reviews and/or 1-3 highly-rated (A_{+2} or better) relevant o/a-docs. Exact numbers will depend on how much risk a funder is willing to take. In practice, we could also use the overall score to set the minimum scientific standard. It is important to note that this requirement should be a pass/fail standard that must be applied to exclude the competition of pseudoscience. The critical point is that we should not have a blanket exclusion of all fringe science ideas.

Once the credibility bar has been passed, the criteria for prioritizing funding for high-risk high-reward projects should be based on the potential impact and feasibility or testability of the proposal. In assessing the potential impact or high-reward factor, funders cannot rely, at least not entirely, on the positive reviews of experts selected by the applicant. Nor can funders rely on the opinions of randomly-selected expert reviewers, as they are likely to be biased in favor of the old paradigm. As one can imagine, the most unbiased reviews on the impact factor are likely to come from non-expert scientists in other fields. The best option would be scientists from immediately adjacent (sub)fields where the proposed new paradigm does not affect their methodologies much. These non-expert scientists may not be able to evaluate the technical details of the proposal, but they can probably tell how

impactful it will be if successful.

The other factor to consider is the testability and/or technical feasibility of such projects. A testable idea under a team with sufficient technical capabilities may have a good chance of success in the execution of the project and produce an immediate impact afterwards. In contrast to the evaluation of the impact factor discussed above, experts in the same subfield or similar subfields using the same technologies should provide the best assessment of this aspect.

The biggest pitfall that a high-risk funding program should avoid is inadvertently supporting mostly low-risk projects. If most randomly-selected experts give excellent ratings, there is near-consensus support from the community in the specialized field, and/or there are a huge number of citations in relevant publications, then such projects should not be considered for funding programs designed to support high-risk efforts.

To better quantify the high-risk factor, the variances of these rating score distributions should be considered. Disagreement among reviewers may also indicate the level of risk involved. If possible, the funder should invite non-specialists from adjacent subfields and/or even completely different fields to evaluate the proposal's potential impact.

A concrete procedure for funding decisions could be as follows. First, set a minimum scientific standard with the overall or soundness score and/or positive reviews/relevant achievements to select credible projects. Then, require that the variance or standard deviation of the overall or soundness score distribution be large enough to ensure that it is a high-risk project. Lastly, a combination of the impact/potential score from non-experts and the testability/feasibility score from experts will determine the funding priorities.

In short, funding programs that aim to support high risk, high reward proposals should set a low bar for scientific standards to filter out pseudoscience projects, and then focus on funding those highly-testable projects with great potential.

5.3 Hiring decisions

For hiring institutions, a candidate's achievement level should be heavily weighted for a research-oriented position such as university professors and research scientists while ECP and associated member roles should carry more weight for an administrative or service position in STEM fields.

Positions for the entry-level members should not be affected much by the new evaluation mechanism in OePRESS. Graduate students (typically at the $L0$ level) are paid through a teaching assistantship. They may also be funded by a fellowship, an individual's basic research funding, or funds from a facility- or project-oriented grant award. This is similar

to how PhD students are currently paid.

OePRESS will revamp the hiring structures for higher positions. Postdoctoral scholars (typically at the $L1$ level) should have a longer-term contract (e.g., five years or more) paid via either fellowship or teaching/research service, even though it may be a low-paying job. The current tenure system in universities should be replaced by an achievement-oriented promotion system.

Academic freedom should be protected, not just for tenured members in the current system, but for all members in the community including students. Hired members should be punished, disciplined, or even dismissed only for adequate cause including neglect of duty, and fraudulent, plagiaristic, or otherwise harmful research (A_- o-docs), but never for academic freedom.

Members at the $L2$ level may obtain long-term (or permanent) positions at universities (similar to lecturers and assistant/associate professors). The job titles may be better renamed as Lecturer I, II, III, IV, V depending on the number of their $L2$ achievements. The pay is mostly via teaching but the job duties may also include assigned research and service work. At national laboratories, the titles would be Scientist I, II, III, IV, V, corresponding to achievement levels $L2/1-5$, respectively, and the duties would be assigned research and development work.

The truly high-paying jobs will be available for members at $L3$ and $L4$ levels. $L3$ members can become professors. There should not be much more than ten professors even in a top academic department. $L4$ members may be appointed to a chaired professorship. There should not typically be more than a couple of them in a department. They should have freedom in what classes they teach and what other responsibilities they can take on. Their counterparts in the national laboratories would be senior scientists ($L3$) and chief scientists ($L4$).

Teaching and research positions in STEM at research institutions should be filled primarily on the basis of a candidate's achievement level. It is important that all hiring institutions strictly use the same title for individuals at the same achievement level. A lower-tier college may have to hire postdoc-level ($L1$) members to teach. In this case, their titles may be Instructor I, II, III, IV, V, corresponding to $L1/1-5$. College faculty on a pure teaching track (with a heavier teaching load) are also possible and at least part of their job performance may be evaluated from their contributions in b-docs and e-docs.

Hiring decisions on other administration/service positions may be based on both a candidate's achievement level and member role (or ECP). Such positions include not only deans and chairs at universities, administration positions like directors at research laboratories, and other similar positions in government, but also program officers at funding

agencies and editors at STEM publishing entities. Most service-type positions may require a minimum achievement level, but the hiring decision should be based primarily on the candidate's member role or ECP.

Most of the academic evaluation will be done through OePRESS on a consistent and equal basis. No more bias and politics will be introduced in contrast to the current limited and often flawed evaluation process at each individual institution. Universities will only have to do some additional teaching evaluation for initial appointments of teaching positions, as they have been doing via demonstration of a colloquium by the candidate.

6 Membership and Community Structures

6.1 Sources of finance and institutional membership

There are several possible sources for self-sustaining operation and development of OePRESS. The main one is probably from funding agencies and hiring institutions that rely heavily on the evaluation service of OePRESS. They may pay annual sponsoring membership fees to OePRESS at different sponsorship levels (diamond, gold, silver, etc.). In other words, the fees depend on how many members an institution has (e.g, weighted $\sum_n 2^n \times$number of Ln members, and/or $100 \times$moderator#$+10 \times$reviewer#$+$commenter#) or on the size of an agency's budget. Certain (small or low-budget) organizations can be exempt from fees and receive green memberships.

The benefits for being a sponsoring member institution may include posting calls for proposals and job announcements on OePRESS, inviting reviews on the evaluation of proposals, job candidates, and their achievements, and having access to other value-added services. For example, their affiliated members can create collaborative collections.

Donations, overlay journal services, and other value-added subscription services may be additional sources of revenue.

6.2 Member roles, privileges, and integration

All researchers can apply for a membership in OePRESS as long as they have a unique open ID (e.g., ORCID). No duplicate memberships are allowed. All members are evaluated from their submitted o/a-docs for their achievement levels as discussed above. All members can submit parent documents and/or make comments (and possibly review and rate depending on their roles) in the OePRESS system. All members can earn credits (ECP) through their contributions to various activities in OePRESS and play an ever-increasing role in the community based on their ECP. And all members can seek funding

support and apply for academic, research, or other related positions in STEM with the expectation of a decision based largely on their achievement level and ECP.

However, L_- members who have posted more A_- o-docs than A_+ ones or members with negative ECP will lose some of the privileges. Submissions from such a member including both parent documents and comments will automatically be placed under moderation. Submissions from members in good standing will not be moderated.

Any registered member (with a unique open ID) can be a commenter. The next level of role is reviewership, which has the privilege of reviewing and rating in addition to commenting. Criteria for reviewership include an achievement level of L1 or higher, an ECP of 100 or higher, and possibly at least one year in good standing. As the next step in the role, a moderator should have an achievement level of L2.5/1, L2/2 or higher, an ECP of 1000 or higher, and possibly at least three years of reviewership. They can participates in various moderation activities in addition to being a reviewer at the same time. Leadership, as the top role, requires at least an achievement level of L2.5/2, L2/3 or higher, an ECP of 5000 or higher, and possibly at least five years of moderatorship. They can be in charge of organizing and leading various committee activities in addition to being a moderator at the same time.

Members can submit parent documents in all research branches. However, if a parent document has multiple authors, only one of the 1st-authors can submit it and all its 1st-authors have to be in good standing for the submission to be accepted immediately.

Members can comment in all research branches. But reviewers can only review and rate in their qualified subfields or fields and moderators can only moderate in their qualified fields. Submissions should appear immediately (or possibly after a short grace period) except for posts from members in poor standing, but the average rating score of an item will not appear publicly until the number of ratings has reached the threshold number (e.g., five). The achievement score or ECP of a contribution will not be calculated and therefore will be assumed to be zero until the number of its ratings reaches the threshold number. A grace period like 5–10 minutes may be provided to allow members to withdraw or make changes to their submissions, in particular, to avoid accidental errors.

An eligible member can rate freely except when the number of the member's ratings associated with no creditable contributions has reached the limit (e.g., five times the ones associated with creditable contributions).

All members can create and manage their own collections of articles by grouping related parent documents according to a topic or their personal interest. Private collections are visible only to the owner. Public collections are visible to all and can be rated by members. Collaborative collections are shared with all members in the collaboration, but are invisible

to the public, and can not be rated (this may be a feature for members affiliated with member institutions).

The evaluation system and community structures should be integrated as a whole: members' role, position, and received funding support should all be tied to the quantitative evaluation results (such as achievement levels, ECP, and proposal ratings) by the entire community. Sponsorship and participation by member institutions and agencies will be critical to ensure such integration.

6.3 Community building and reorganization

If enough moderators and reviewers agree and join, a new sub-community at all three levels of disciplines, fields, and areas can be established, or can be a split-off of an existing (sub-)community. If the majority of two sub-communities agree, they can be merged into one. If a sub-community shrinks and its numbers of moderators and reviewers fall below certain thresholds, the sub-community should be merged back into its parent community. The corresponding STEM branches will also be renamed and reorganized accordingly.

The forming threshold for a sub-field (topic area) community can be at least 20 moderators and 200 reviewers. The terminating (merging-back) threshold should be halved, i.e., 10 moderators and 100 reviewers. In other words, a subfield should be terminated if it shrinks by half in size. The similar thresholds for fields should be about 10 times higher while the thresholds for disciplines should be multiplied by another factor of 10.

Appendix A: Further Discussions and Implementation Details

A.1 Disciplines, fields, and areas

The hierarchical structure of research branches in STEM is outlined in Sect. 2.1. Here we present a detailed example for the discipline of physics.

In the discipline Phys, we can initially define the following fields: AA (Astrophysics & Astronomy), AMOQ (Atomic, Molecular, Optical and Quantum physics), CM (Condensed Matter), NP (Nuclear & Particle physics), MISC (miscellaneous). The suggested subfields are as follows: AA.cosmology, AA.galaxies, AA.stellar, AA.solar, AA.planetary (planets, exoplanets, and astrobiology), AA.highenergy (high-energy astrophysical processes and phenomena), AA.im (instrumentation and methods); AMOQ.atomic, AMOQ.optics, AMOQ.molecular, AMOQ.quantum; CM.sc (superconductivity and superfluidity), CM.sces (strongly correlated electron systems), CM.soft (soft matter), CM.disordered (disordered

systems), CM.nano (mesoscale and nanoscale), CM.stat (statistical physics); NP.nucl-ex, NP.nucl-th, NP.hep-ex, NP.hep-th, NP.im (instrumentation and methods), NP.accel (accelerator physics); MISC.biophy (biophysics), MISC.mathphy (mathematical physics), MISC.chaos (nonlinear physics, chaos, self-organizing systems), MISC.fluid (fluid dynamics). In addition, they all have an "other" subfield for the uncategorized.

Under the cross-disciplinary top level, we can define the fields such as Cross.HPS (History and Philosohpy of Science), Cross.MS (material science), etc. A complete branch can include all three levels, first two levels, or only a top level. The special branch "All" represents all disciplines. Each parent document should be associated with one primary branch and can be cross-linked up to two other branches.

A.2 More parent document details and phased evaluation

A parent document may have multiple unique IDs: OePRESS DOI (yyyymmdd.xxxx) as the primary one, alternative OePRESS ID (yyyymmdd.[type]-nnnn), journal DOI, preprint DOI (e.g., from arXiv, osf.io, etc.), and other types of IDs. OePRESS may allow up to 16 revisions for each parent document posted locally. Each parent document entry may list three top rated public collections where it is included. Only o/a/f/p/s/t/u-docs (see Table 1) can be reviewed and other docs should allow comments only.

Once established, OePRESS may record metadata of previous publications and other documents from elsewhere into its database. For parent documents that exist elsewhere, no local storage is required and OePRESS can take DOI or other ID information to retrieve data from other services: arXiv.org, osf.io, bioRxiv.org, journals, etc.

In an i-doc (see Table 1) for each issue or bug raised in OePRESS, if approved, the submitter can earn credits, typically from one to ten credit points depending on its severity in well-known cases. For example, members earn credits for correcting errors in a parent document's meta information, while the ECP for the original authors will be reduced accordingly if they are responsible for the errors. In particular, if a document has multiple DOIs, identifying each unassociated new DOI will be worth one point and the authors' ECP will be reduced by one point. Finding duplicates at OePRESS will get the member 10 points and the submitter of the duplicate will receive -10 points. For i-docs in other cases, rating will be enabled and the rating score represents the severity level (typically for $S > 1$ to be a good issue/bug report). We may use the same formula for r-docs (Eq. 2), but with a smaller scaling factor $a = 1.0$. They will be closed when the issue is resolved.

New feature requests in n-docs can only be submitted and rated by reviewers and the above. A minimum voting rate (e.g., 30%) is required and the best rated request will be considered first for future implementation. After the minimum voting rate has been

reached, a notification will be sent to all relevant members who are eligible to rate, asking for more ratings. A final score will be tallied within a fixed period of time (e.g., two weeks). The proposer will earn credits according to the r-doc ECP formula (Eq. 2) with $a = 1.0$.

The issues raised in i-docs, if they can not be easily solved within the current system and have received a high average rating, can trigger the formation of a dedicated moderation committee. The issue submitter and up to three top-rated commenters are automatically members. In addition, three to ten moderators in research branches relevant to the issue can join within a fixed period of time (e.g., two to four weeks). If not enough moderators are interested, the issue will be closed. If more moderators want to join, those with higher achievement levels and ECPs ($5 + 5$) will be given priority. Within the announced time frame, the committee will provide a solution published as an m-doc. After a fixed comment period (e.g., three months), the solution or a revised version will be adopted if it receives high enough ratings.

Similar measures can be taken for n-docs. In particular, for suggestions related to changes in formulas and parameters in OePRESS. A highly-rated proposal in such changes will trigger the formation of a committee. The submitters and up to five top commenters are automatic members. In addition, five to fifteen moderators can join the committee. Eventually the committee will publish their report as an m-doc.

Phased Evaluation: Only o/a/f/p/u/n-docs can be reset for new evaluation phases. In a new phase, the average rating will be calculated using only new ratings, and all the N-numbers and the lapse of time T will be reset for new reviews, comments, and ratings. The N-number and T of an old contribution from an old phase will remain the same. However, the ECP of the old contribution will be recalculated using the new average rating in the new phase. A historical record of the average rating and/or ECP for all phases of the parent document could be kept in addition to an overall average rating and/or ECP. All relevant $\bar{S}$ and ECP will be re-calculated in a new phase. The criteria for a new phase may be that the shift in the average rating should be very significant (e.g., > 0.9) and the new period should be long enough (e.g., $> 5T_{\text{resp}}$). An AI technique would be very helpful in identifying such new phases. It could also require the approval of at least 3 moderators as a safety measure. The timestamp of the first landmark comment or rating may be used as the reset time.

A.3 OePRESS database tables

The main tables in the local database of OePRESS: members, parent-doc types, research branches, parent docs, comments, reviews, ratings. The following database tables can be partitioned according to their parent document types and STEM areas.

Table of Parent Documents: title, author/editor/organizer IDs (such as ORCID) and contribution schemes, OePRESS DOI (yyyymmdd.xxxx), doc-type, alternative-ID (nnnn), other DOIs and IDs from elsewhere and first and last update date/time for each source, link to the most current PDF, links to external reviews, links to external comments (from twitter, facebook, blogs, etc.), media links, [current phase#, reset timestamp, current N-numbers for review/comment/rating, average rating score $\bar{S}$, and ECP (for the current phase)], [N_0-numbers, $\bar{S}_0$, and ECP$_0$ (for overall)], first submission/received date/time, date/time (when this entry is added), modified data/time and revision numbers, [abstract, full text, other pdfs will be retrieved from other sources].

Table of Creditable Comments: timestamp, comment-ID (xxx), alternative-ID (nnnn), parent-doc-ID (yyyymmdd.xxxx), parent-ID, member-ID/author-scheme, anonymity level (open/semi-anonymous/anonymous), [k-th phase, N-th comment, lapse of time T (in the phase where this comment belongs)], [current phase#, number of ratings, number of comments, average score $\bar{S}$, and ECP (for the current phase)], [N_0-th, T_0, number of ratings, number of comments, $\bar{S}_0$, and ECP$_0$ (for overall)], comment text and/or PDF attachment.

Table of Contributed and Invited Reviews: it should be similar to the above table, but no anonymity is allowed and ID of the inviting moderator may be added.

Table of Uncreditable Comments: timestamp, comment-ID (xxx), alternative-ID (nnnn), parent-doc-ID (yyyymmdd.xxxx), parent-ID, member-ID, comment text.

Table of Ratings: timestamp, rating-ID, item-type, item-ID, member-ID, score S, [k-th phase, N-th rating, lapse of time T (where this rating belongs)], [current phase# and ECP (for the current phase)], rating# (up to 3 from the same member, 0 means invalid).

Table of $\bar{S}_{\mathbf{max}}$: parent-doc-ID (yyyymmdd.xxxx), member-ID, review-ID, $\bar{S}_{\text{rev}}$, comment-ID1, $\bar{S}_{\text{max1}}$, comment-ID2, $\bar{S}_{\text{max2}}$, comment-ID3, $\bar{S}_{\text{max3}}$ (review score and three top-rated comment scores).

Table of Members: member-ID, name, affiliation, start-time of membership, start-time of reviewership and qualified areas, start-time of moderatorship and qualified fields, start-time of leadership and qualified disciplines, achievement level Ln/m, list of top 10 o/a-docs and achievement scores (parent-doc-ID1, $A1$, ... parent-doc-ID10, $A10$), list of ECP total and partial values [ECP_total; ECP1, Branch1, Doc-type1; ECP2, Branch2, Doc-type2; ...], meta_key=ECP_History, meta_values={timestamp1; ECP1; timestamp2; ECP2; timestamp3; ECP3...}. The table can be updated weekly. Each time when a member's ECP is changed, the timestamp is checked and the historical record will be updated if the time has passed by more than a week.

Table of To-Be-Updated: parent-doc-ID (yyyymmdd.xxxx), new phase; parent-doc-ID, new rating; review/comment-ID, new rating; formula-ID, changed; .... Easy changes can

be updated on the fly, but complicated changes will be updated later in sequence according to this table. This will make doing updates more efficient. Each time when a new update request is received, it will be checked against the table and be only added if there is no duplicate. When an update request is executed and done, the corresponding line in the update table will be removed. When a formula is changed, all relevant ECPs and other calculations will be updated. When new ratings or a new phase for a parent document is listed, ECPs and/or achievement score of this document and derived contributions will be updated. If new ratings for a review or comment exist, relevant ECPs will be updated.

A.4 Web interface and coding

The main page of OePRESS should contain the list of all research branches in the format of Discipline.FIELD.area, possibly in a three-level menu scheme. Clicking on any of these branches at any of the three levels will lead to its individual portal page. A menu for all parent document types should also be present and linked to their portal pages. The main page should also include sections for About, FAQ, Donate, Contact, Registration and Login, bug reports (i-docs) / feature requests (n-docs) / moderation and committee reports (m-docs), instructions for registration and participation in OePRESS, and guidelines on how to submit, comment, review, and rate. A job list and a list of funding/proposal calls from affiliated member institutions and agencies may also be included on the main page and other relevant portal pages.

OePRESS should support comments and reviews in multiple formats: latex, markdown, pdf, doc, and other rich text formats (links, images, tables, etc). It should provide a real-time preview function to easily catch errors before submitting. A set of top-rated comments and reviews should be listed at the top by default. All comments and reviews can be viewed in two ways: a flat chronological order of posting with a back-link if it is a reply; a threaded layout ordered by the most recent update first.

The timestamps of a parent document should include the first submission/received date/time at OePRESS or elsewhere, publicly announced date/time if moderated, the last revision date/time, other major/minor revision date/time, most recent activity (comment / review / rating) date/time, and reset date/time of each evaluation phase.

The OePRESS DOI (yyyymmdd.xxxx or yyyymmdd.xxxx.xxx) will be the main ID. Other unique document IDs (e.g., ISBN, PUBMED/PMID, arXivID, ADS bibcode, bioRxiv, viXra.org, osf.io, and journal DOIs) associated with the same document will be directly linked to the main ID.

All ratings, achievement levels, and ECP can be updated instantly, daily or weekly, depending on the availability of processing power. The character/byte limits for uncreditable

comments and reviews may be suggestive and unenforced, i.e., as default settings for contribution types and members may reclassify their contributions within the grace period of submission. All submissions should be in English and contributions in other languages will be categorized as uncreditable comments unless a regional OePRESS system is implemented to serve a regional non-English community.

All comments and reviews should be published under the Creative Commons Attribution License CC-BY (4.0) or a similar license. The parent documents may be published under a different license.