



THE UNIVERSITY OF
MELBOURNE

Responsible Use of Research Metrics at the University of Melbourne

メルボルン大学における研究指標の責任ある使用

Professor Justin Zobel

Pro Vice-Chancellor (Graduate & International Research)

5 February 2021

ジャスティン・ゾーベル教授
副学長（大学院・国際研究担当）
2021年2月5日

※本文中の日本語メモは、発表者の了解のもとに付記した仮訳ですのでご了承ください。





Responsible use of research metrics (RURM) at the University of Melbourne

メルボルン大学におけるRURM（責任ある研究評価の使用）

Where we are:

- Formal adoption by University governance of a RURM principle:

The University of Melbourne is committed to the principle that assessment and recognition of the performance of individuals, teams, and academic units requires expert judgement and peer review, in accordance with the Leiden Manifesto and the San Francisco Declaration on Research Assessment. Under this principle, research metrics may be used to inform assessment, where appropriate, but not as the primary determinant of outcomes. The University will ensure that the use of research metrics is at best practice and respects the rights of individuals.

- There is general acceptance of the RURM principle amongst management and academics.
- We have an ongoing committee whose role is to oversee implementation and adherence in practice.

In this talk: How we got here; what comes next; reflections.

現状：

- RURMの原則を大学のガバナンスで正式に採用。

「メルボルン大学は、ライデン声明と研究評価に関するサンフランシスコ宣言に基づき、個人、チーム、学術単位の業績の評価と認識には専門家の判断とピアレビューが必要であるという原則にコミットする。この原則に基づき、適切な場合には、評価の情報提供のために研究指標を使用できるが、成果の主要な決定要因としては使用しない。大学は、研究指標の使用をベストプラクティスとし、個人の権利を尊重したものであることを保証する。」

- RURMの原則は、経営陣や研究者の間で一般的に受け入れられている。
- 本学には、実際に実施、遵守されているかを監督する役割を担う委員会が存在する。

本日の内容：どのようにしてここまで来たのか、今後の発展、振り返り



Background: internal context

背景：大学内の状況

- As is for the case for many institutions, we have had long-running conversations about
 - The representational diversity of our workforce: gender, culture, personal circumstances ...
 - The diversity in proficiency, that is, whether we were recruiting a sufficiently broad range of skills.
 - Recruiting people for characteristics such as their academic leadership, industry skills, diversity, etc., can ‘set them up to fail’ if their strengths aren’t modelled in performance metrics.
 - The increasing availability of rich publication data made it easy to create tools for scoring the performance of academics – enabling a reductionist capability that
 - Could be deployed without considering the impact on individuals and
 - Concealed the limitations and errors in the data sources.
 - Mid-level managers sometimes took decisions based on metrics (for individuals, labs/research groups, departments) that were not in the broader interests of the University
 - Some of these decisions demonstrated that metrics alone do not provide effective managerial guidance and are a poor way of communicating University strategy.
 - For a couple of decades we have used numerical teaching scores from student feedback for staff assessment – thus creating a culture of score-based measurement of performance.
- 多くの機関同様、長年議論をしてきた
 - 多様な労働者：性別、文化、個別の事情
 - 習熟度の多様性：十分に幅広いスキルの採用
 - 学術的リーダーシップ、産業界でのスキル、多様性などの特性を理由に人材を採用しても、彼らの強みがパフォーマンス指標でモデル化されていないならば、「失敗するように仕組まれてしまう」可能性
 - 豊富な出版データが利用できるようになったことで、研究者のパフォーマンスに点数付けをするためのツールを簡単に作成できるようになった。一方で、これは以下のように還元主義的な評価をもたらした：
 - 個人への影響を考慮せずに利用される
 - データソースの限界や誤りが隠蔽される
 - 中間管理職が、大学のより広い利益に反する（個人、研究室・研究グループ、部局の）指標に基づいて、意思決定を行うことがあった。
 - これらの決定の中には、評価指標だけでは効果的な管理・運営のガイダンスができず、加えて大学の戦略を伝えるには不適切な方法であることを示唆するものがあった。
 - この数十年、教育面では、学生のフィードバックから得られる数値的なスコアを教員の評価に使用してきたが、その結果、点数に基づいてパフォーマンスを測定する文化が生まれた。



Background: Australian context

背景：オーストラリアの状況

- Some competitor institutions in Australia developed and use highly metric-driven approaches to performance management (individual and academic unit).
- National assessment processes effectively mandate that all Australian institutions gather and report on highly detailed metrics, though this has no impact on funding.
 - Our major funding bodies do not make extensive use of metrics in funding decisions, however.
- Success in recruiting international fee-paying students is dependent on success in global metrics and many Australian universities have set rankings targets in their strategies.
 - However, globally, deliberate attempts to improve metric standing through performance frameworks have not been very successful.
- Strong performance in metrics is widely seen as a key institutional accomplishment.

That is, there are no specific national factors encouraging endorsement of DORA or Leiden, but there are factors that encourage the use of metrics.

- オーストラリアの競合機関の中には、パフォーマンス管理（個人および学内組織）のため高度に指標化されたアプローチを開発し、使用している
- 国の評価プロセスでは、全ての教育機関が高度に詳細な指標を収集し、報告することが事実上義務づけられているが、これは資金配分には影響を及ぼさない。
 - 主要な資金配分機関は、資金配分の決定に指標を広く活用していない。
- 留学生の採用に成功するかどうかは、世界的な評価基準において良い評価を受けられるかにかかっており、多くの国内大学では、大学戦略の中でランキングの目標を設定している。
 - しかし、世界的には、パフォーマンスを用いた指標上の評価の向上を意図的に試みても、あまり成功していない。
- 指標における優れたパフォーマンスは、大学としての重要な成果として広く認識されている。

つまり、DORAやライデンを支持するように促す特定の国内要因はないが、指標の使用を奨励する要因はある。



Towards signing of DORA

DORA締結に向けて

- There were deepening concerns about new metrics-driven performance assessments at Melbourne:
 - Some of which were being applied in unintended and inappropriate ways.
 - Some were implemented and designed by people with poor understanding of the concerns and impacts.
 - Which were leading to poor morale amongst academics who were assessed by numbers (such as citations) that they are unable to influence, or which seemed arbitrary and irrelevant.
- A senior academic with interest in DORA and Leiden took the opportunity of a general review of practice to initiate a working group (which I chaired) that was intended to examine University use of metrics.
 - This was the original aim, but the initial task was to develop a statement of principle against which RURM would be assessed.
 - Development of this RURM principle became the key activity.
- With the (unexpected!) support and encouragement of a wide range of senior and junior academics, the RURM principle was smoothly accepted by University governance boards, leading to signing of DORA.
- メルボルン大学では、新しい指標に基づいたパフォーマンス評価に対する懸念が深まっていた。
 - その中には、意図しない不適切な方法で適用されているものもあった。
 - 懸念や影響を十分に理解していない人々によって実施・設計されたものもあった。
 - これらは、自分では影響を与えることができないか、恣意的で無関係と思われるような数字（被引用など）で評価を受けた研究者の志気を低下させる原因となっていた。
- DORA とライデンに関心を持つシニアレベルの研究者が、一般的な指標の使用を調査し、メルボルン大学での指標の使用を検討することを目的としたワーキンググループを立ち上げた（私が議長を務めた）。
 - これが当初の目的だったが、最初の課題は、RURM を評価するための原則を作成することだった。
 - この RURM（責任ある研究評価の使用）の原則の開発が重要な活動となった。
- 予想外にも（！）、シニアから若手まで幅広い研究者の支持と励ましを得て、RURM の原則は大学のガバナンス委員会にスムーズに受け入れられ、DORA への署名につながった。



Current activity

現在の活動

- We signed DORA at the end of 2019 but COVID distractions have slowed our internal implementation.
- We are
 - Reviewing specific HR practices (our faculties have significant independence and freedom).
 - Developing materials for explaining the RURM principle to managers and academics.
 - Introducing examples of RURM-friendly assessment practices.
 - Providing a point of reference for concerns about uses of metrics.
- Our HR practices are already generally compliant with RURM but new stresses (COVID-related economic shock) are creating pressure to be more rigorous in assessment.
- 2019年末にDORAに署名したが、COVIDに妨害され、内部実装が遅くなってしまった。
- 現在の活動
 - 具体的な人事慣行の見直し（本学の学部は独立性と自由度が高い）
 - 経営陣や研究者にRURMの原則を説明するための資料を作成
 - RURMに準拠した評価の実践例の紹介
 - 評価指標の使用に関する懸念事項についての参考資料の提供
- 本学の人事慣行はすでに概ねRURMに準拠しているが、新たなストレス（COVIDに関連する経済ショック）により、評価をより厳格にしなければならないというプレッシャーが生じている。



Reflections

振り返り

- Truly innovative or disruptive work is often successful on longer time-frames than is captured in metrics, e.g., the work that led to many Nobel prizes, and may not be a clean fit for typical disciplinary journals.
- The same is true of work that is focused on practical impact.
- Some metric-driven practices are incompatible with respectful management of individuals and teams.
 - And they encourage simplistic decision-making that can harm the institution.
- Metric-driven assessment processes are poor for morale, generate conflict, and build a workforce of ‘similar’ individuals who are not developing other valuable skills, such as in academic management.
- Academics who are focused on metrics are not focused on mentoring colleagues, etc.
- Metric-driven processes have undesirable effects when the metrics change or are inconsistent:
 - A historical shift from ‘counting publications’ to ‘counting citations’ meant that some researchers moved from *strong* to *weak* and were left with a body of poorly regarded work.
 - Metrics that ignore disciplinary conventions can prevent people from collaborating; as can rules such as a requirement that a publication be reported in a particular discipline.
- 真に革新的で破壊的な研究は、多くのノーベル賞につながった研究のように、指標で捉えられるよりも長い時間軸で成功を収めることが多く、典型的な分野別学術誌には適さないかもしれない。
- 実社会へのインパクトに焦点を当てた研究についても同様である。
- 指標を重視したやり方の中には、個人やチームを尊重した管理・運営とは相容れないものもある。
 - また、それらは機関に害を及ぼす可能性のある単純化された意思決定を奨励する
- 指標を重視した評価プロセスは志気を下げ、対立を生み、他の価値あるスキル（例：大学の管理・運営能力）を身につけていない「似たような」労働力を生み出すことになる
- 指標を重視する学者は、同僚のメンタリングなどに力を入れていない。
- 指標主導のプロセスは、指標が変化したり、一貫性がない場合に望ましくない効果をもたらす。
 - 歴史的に「出版物を数える」から「被引用数を数える」へと移行したことで、一部の研究者は強い研究者から弱い研究者へと移行し、高い評価をされない研究と共に置き去りに。
 - 学問分野の慣習を無視した指標は、研究者を共同研究から妨げることがある：例えば、特定の学問分野で出版物が報告されていることを要件とするような規則がある。



Reflections ...

振り返り...

- Good organisational outcomes come from bringing a mix of skills to all activities
 - It is a mistake to measure all academics on the same simplistic scales.
 - High rankings are the result of supporting strong academics to pursue their passion for research.
 - Use of metrics as an institutional goal does not encourage strong research – it creates incentives to seek short-term rewards.
- Metrics are critical to good institutional decision making – they should not be discarded
 - But they need to be designed alongside guidelines that describe how they are to be used, and only ever used as a component of a decision-making process – not as a replacement for the process.
 - Reductionist use of metrics can displace more thoughtful or principled organisational messages about strategy, value, and purpose.

That is, good practice involves ensuring that metrics are used, but that they are used respectfully and within understood limitations; they need to be accompanied by well-crafted documentation and safeguards.

- 優れた組織的成果は、全ての活動に様々なスキルがもたされることから生まれる。
- 全ての研究者を同じ単純な尺度で測るのは誤り
- 上位にランクされるのは、研究への情熱を追求する強い研究者を支援した結果である。
- 指標を機関の目標として使用することは、強い研究を奨励するものではなく、短期的な報酬を求めるインセンティブを生み出すものである。
- 指標は優れた機関の意思決定に不可欠なものであり、不要とすべきではない。
- しかし、それらは、それらがどのように使用されるべきかを記述するガイドラインと一緒に設計される必要があり、（プロセスの代わりとしてではなく）あくまで意思決定プロセスの一要素として使用されるべき。
- 指標を還元主義的に使用するのではなく、戦略、価値、および目的に関するより思慮深いまたは原則にかなった組織のメッセージとする。

つまり、優れた実践とは、指標が使用されながらも、丁寧に、共通に理解された制約の範囲内で使用し、熟考された文書とセーフガードを伴う必要がある。



THE UNIVERSITY OF
MELBOURNE

Thank you

<http://www.unimelb.edu.au>





Appendix: Leiden Manifesto principles

付録：ライデン声明の原則

Quantitative evaluation should support qualitative, expert assessment.

Measure performance against the research missions of the institution, group or researcher.

Protect excellence in locally relevant research.

Keep data collection and analytical processes open, transparent, and simple.

Allow those evaluated to verify data and analysis.

Account for variation by field in publication and citation practices.

Base assessment of individual researchers on qualitative judgement of their portfolio.

Avoid misplaced concreteness and false precision.

Recognize the systemic effects of assessment and indicators.

Scrutinize indicators regularly and update them.

原則 1 定量的評価は、専門家による定性的評価の支援に用いるべきである。

原則 2 機関、グループ又は研究者の研究目的に照らして業績を測定せよ。

原則 3 優れた地域的研究を保護せよ。

原則 4 データ収集と分析のプロセスをオープン、透明、かつ単純に保て。

原則 5 被評価者がデータと分析過程を確認できるようにすべきである。

原則 6 分野により発表と引用の慣行は異なることに留意せよ。

原則 7 個々の研究者の評価は、そのポートフォリオの定性的判定に基づくべきである。

原則 8 不適切な具体性や誤った精緻性を避けよ。

原則 9 評価と指標のシステム全体への効果を認識せよ。

原則 10 指標を定期的に吟味し、改善せよ。

(日本語訳参照：小野寺 夏生、伊神 正貴、*研究計量に関するライデン声明について*、STI Horizon 2016 Vol.2 No.4、35-39頁 <https://www.nistep.go.jp/wp/wp-content/uploads/NISTEP-STIH2-4-00050.pdf>)



Appendix: Further internal context

付録：大学内の状況（より詳しく）

- We have been keen to encourage research of local or national significance – for example, on topics that are relevant to Indigenous Australians or which concern specialised agricultural practices.
 - These are important to our social mission but achievement might not be reflected in typical metrics.
- There were clear statements in our strategies that endorsed metric-driven decision making, e.g., ‘recruit highly cited researchers’
 - However, these are in conflict with other desirable organisational goals such as ‘recruit for leadership and organisational citizenship’.
- Some deployment of metrics was by individuals who were not experienced in academic management.
 - Some designers of metrics were not working to clear goals, but just wanted something to measure.
- メルボルン大学は、オーストラリア先住民に関連するテーマや、専門的な農法に関連するテーマなど、地域的、あるいは国家的に意義のある研究を奨励することに熱心に取り組んできた。
 - これらの研究は、本学の社会的使命にとって重要だが、その成果は一般的な指標には反映されていないかもしれない。
- 大学戦略には、「引用度の高い研究者を採用する」など、指標に基づいた意思決定を支持する明確な文言があった。
 - しかし、これらは、「リーダーシップと組織的市民性のための選考」などの他の望ましい組織目標と相反するものである。
- 大学での管理・運営の経験のない個人により、指標が使用されている例があった。
 - 指標の設計者の中には、明確な目標に向かって仕事をしているのではなく、ただ何かを測りたいと思っている人もいた。



Appendix: Further documented concerns with research metrics

付録：研究指標に関するさらなる懸念

- Metrics are typically simplified, summative, and standardised, and exclude individual circumstances.
- Some metrics can be gamed, or can create perverse incentives; journal impact factors and citation counts are two prominent examples.
- Metrics can provide the appearance of equity and transparency in assessment, particularly when large numbers of individuals are being considered, but in practice it can be impossible to design metrics to fully account for disciplinary differences.
- Metrics that relate to teams are sometimes devolved to individuals even when doing so is not representative of any individual's performance.
- There are differences across fields, not just in the levels of measured performance that are typically achieved by individuals on a particular metric, but in regard to which of the metrics are meaningful or relevant; such variation can apply even between fields within a single discipline.
- The shortcomings noted above are relevant to all fields of research, but the metrics that are deployed tend to be better matched to STEM than to HASS disciplines.
- Careless use of metrics, such as for comparative ranking of individuals, can violate privacy principles; anonymisation can be effectively impossible when the number of individuals is small.
- 指標は一般的に単純化、総括的、標準化されており、個々の事情を除外する。
- 指標の中には、ゲーム化されていたり、逆のインセンティブを生み出す可能性のあるものもある：ジャーナルのインパクトファクターと被引用数は、2つの顕著な例である
- チームに関する指標は、たとえチーム内のどの個人のパフォーマンスを代表するものではない場合でも、チーム内の個人に帰属してしまう
- このような違いは、特定の指標で個人が典型的に達成したパフォーマンスのレベルだけでなく、どの指標が意味のあるものであるか、あるいは関連があるかという点でも、分野間で差があり、1つのディシプリン内の分野間でさえも当てはまることがある。
- 上述の欠点はすべての研究分野に関連しているが、導入されている指標は、HASSの分野よりもSTEMの方に適している傾向がある。
- 個人の比較ランキングのようなメトリクスへの不注意な使用は、プライバシーの原則に違反する可能性がある。