Research Paper

JIS

Journal of Information Science
1–10

DOI: 10.1177/01655515211040654
journals.sagepub.com/home/jis
SAGE

A typology of research discovery tools

Andreas Nishikawa-Pacher
TU Wien Bibliothek, Austria; Vienna School of International Studies, Austria; Department of Legal and Constitutional History, University of Vienna, Austria

Abstract
There has been a proliferation of new research discovery tools that aid scientists in finding relevant publications. To obtain a general overview of this development, this article generates a conceptual typology of all possible research discovery tools by drawing from the information-theoretical concepts of redundancy/variety. Bibliometric links between scholarly publications can thus exhibit 'redundancy' (i.e. expectable linkages between academic works) or 'variety' (i.e. original co-occurrence patterns). On the redundancy-reproducing end of the typology are machines that harness extant co-citations or keyword queries, such as academic search engines and paper recommender systems. The variety end of the spectrum harbours services that enable categorial browsing or that suggest publications randomly, such as journals' tables of contents or random paper bots. The typology has implications for understanding how the design of research discovery platforms may ultimately shape aggregate citational networks of science.

Keywords
Bibliometrics; citations; information retrieval; randomness; redundancy; research discovery; serendipity; systems theory; typology; variety

1. Introduction

To find scientific publications, one could use *Google Scholar* and search for specific keywords go to the university library and rummage in the bookshelves, access *Twitter* and follow researchers announcing their latest publications, browse through the tables of contents of scholarly journals, whether online or in print, or look for interesting titles in the reference lists of the papers already read. While highly heterogeneous, what these approaches have in common is their function: research discovery.

There has been a proliferation of such research discovery tools, especially in the digital realm. Nourished by technological advances and machine-readable metadata, dozens of pieces of software and online apps have sprung up, all of which demand user-input based on which they then provide links to research papers. Platforms like *CitationGecko, CoCites, ConnectedPapers, LENS, the Observatory of International Research (OOIR), SciLit, Semantic Scholar* and *wizdom.ai* are just a few examples. They now complement traditional sources of research inspiration, such as conference meetings, email alerts journal and book clubs. While each new tool may be helpful on its own, there is a downside to this dynamic: it leads to a rather confusing landscape of research discovery tools. Obtaining a systematic overview seems almost impossible.

Scholarly discussions on research discovery tools have not helped so far; instead, they offer an internally fragmented field in which each part focuses on a narrow area. Some look at specific software [1], others study 'serendipity' or unexpected ways of encountering useful research [2] and still others analyse co-citational links [3] to render them visible in network graphs [4]. While each cluster is useful in itself, none attain such a degree of abstraction that comprehensively subsumes all possible research discovery tools. Rather than a messy taxonomy of existing platforms based on mutually disconnected empirical impressions, the meta-scientific field could gain clarification by generating a conceptual typology [5] that would highlight the simple but universal commonalities behind all of these tools.

Corresponding author:
Andreas Nishikawa-Pacher, TU Wien, Library, Resselgasse 4, 1040 Vienna, Austria.
Email: andreas.pacher@tuwien.ac.at

This article takes on this conceptual task so as to find a common variable operating behind all possible research discovery tools. It does so by drawing from the information-theoretical distinction of *redundancy* and *variety*. Redundancy means that knowing one piece of information allows one to draw inferences about other information [6]; the revelation of the other datum would then seem expected, predicted and superfluous. Variety, however, means that the presence of one (known) piece of information does not disclose other (as yet unknown) information. The revelation of the other datum then seems surprising, unexpected and novel.

To briefly summarise the results of an application of the redundancy/variety-distinction onto the system of science, one obtains a conceptual typology based on a dimension ranging from redundancy-reproducing to variety-enhancing research discovery tools. The redundancy-reproducing tools can be (1a) citation-based or (1b) query-based, while the variety-enhancing ones may be (2a) category-based or (2b) randomness-based. On the redundancy side, therefore, are (1a) platforms that draw from extant bibliographic couplings (e.g. *CitationGecko, ConnectedPapers* and *CoCites*) and (1b) academic search engines or recommendation tools that use semantic proxies to find publications relevant to a user-input (e.g. *Google Scholar* and *Mendeley Recommendation*). They systemically reproduce bibliometric redundancy because they draw from pre-existing links between scientific works. On the other side are variety-enhancing tools that are (2a) category-based with a 'browsing'-feature (e.g. a simple *Curriculum Vitae* or CV, a set of journals' tables of contents like *JournalTOCs* or an *Altmetric*-powered ranking of trending papers, such as *OOIR*) and (2b) randomness-based sources of research discovery (e.g. random paper bots). The closer a tool reaches the variety end of the spectrum, the more its recommendations are unbiased by publication dates, measurements of impact, research fields, extant citational links and academic or semantic networks.

The following section delves deeper into the theoretical background behind informational redundancy and variety. Based on these concepts, the part thereafter develops the typology of research discovery tools and offers illustrative examples. The typology demands universality in the sense that every possible research discovery tool should find a place in this scheme. A final section then discusses, inter alia, how the dominant landscape of research discovery tools may ultimately shape aggregate citational patterns of science, which could be currently skewed in favour of redundancy-reproducing machines. After pointing out a few limitations of this article, it suggests further venues for the scientific-experimental exploration of research discovery.

2. The concepts of redundancy and variety

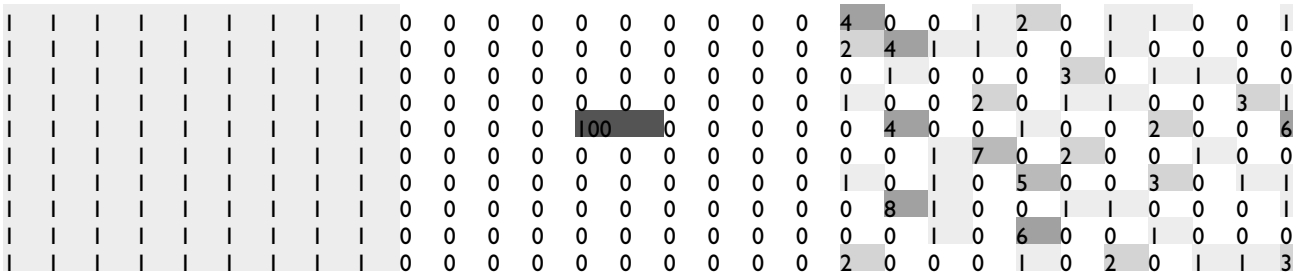
Every item of information provides a surprise, for it momentarily appears and then disappears, resolves uncertainty and renders any given context slightly different than before [7,8]. However, the surprisefulness inherent in a piece of information can vary by degree. For instance, there is less surprise in cases of informational *redundancy*, that is, when the knowledge of one piece of information allows inferences about other (as yet unknown) information [6]. A non-redundant system, however, harbours *variety*, or a rich plurality of distinct events so that an observer's knowledge of one piece of information does not disclose another one [8,9]. At their extreme, redundancy reproduces self-circularity in a permanent repetition of the same, while variety increases a system's openness towards more and more stimuli.

To apply this concept to the system of science, the unit of observation whose surprisefulness varies may be regarded as scientific publications and their citational interlinkages. One could, thus, imagine a scientific system in which every research finding must be cited once and only once. Such a condition would ensure the maximum *variety* of reference patterns within the system. One may also imagine the opposite, namely, a structure under which only a single publication must be cited in every work, while all other research outputs remain uncited. This pattern would amount to a repetitive reproduction of bibliometric *redundancy*.

Neither scenario is wished for, as they both push the scientific system into entropy. The first scenario (of maximum variety) does so by constantly bringing forth surprising citations without a predictable structure conducive to accumulative knowledge growth. The second one (of maximum redundancy) engenders entropy due to a permanent repetition of the same. A well-functioning scientific system, in contrast, would combine both variety and redundancy (cf. the right illustration in Table 1). A new publication conveys familiarity by referencing already-known works, but it simultaneously energises the system with new information requiring new citations – for example, through an innovative perspective, an unusual theory, an original question or novel data [10].

One may illustrate this abstract concept with a (semi-fictitious) example. Say that Merje Kuus [11], or *MK14*, is often co-cited alongside Fiona McConnell [12], or *FM17*, as both study similar aspects of diplomatic practices. Now, if a paper on diplomacy cites *MK14*, it will most probably cite *FM17*, too. Assume that another paper on diplomacy, one by Liudmila Mikalayeva [13], or *LM12*, has 0 citations so far despite its topical relevance to both *MK14* and *FM17* (at least until this very paper's publication). Redundancy is reproduced when, despite the same topical relevance, *MK14* and *FM17* accrue dozens of co-occurring citations while *LM12* is left with none. Perhaps there are dozens of other papers in

Table 1. Three variants of how 100 citations may be distributed among 100 (in the centre: 99) scientific publications.



left: variety (each paper accrues one citation); centre: redundancy (each paper is uncited, except for a single publication that accrues 100 citations); right: some papers are cited more than others in a seemingly random distribution.

Table 2. Examples illustrating the concepts of bibliometric redundancy, variety and a combination of both.

Concept	Informational value	Example: a set of citations to ...
Redundancy	Low surprise	{Garfield [21], Merton [22], and Rossiter [23]} (unsurprising because all three are often co-cited)
Variety	High surprise	{Garfield [21], Margulis [24], and al-Khwarizmi [25]} (surprising because these three have never been cited together)
Balance of both	Combination	{Garfield [21], Merton [22], and al-Khwarizmi [25]} (balanced because while Garfield and Merton are often cited together, which conveys familiarity [=redundancy], the reference to al-Khwarizmi adds an innovative perspective [=variety])

the same research field suffering similar fates like *LM12*. In total, there may be 25 papers on that topic, but only the most probable two (namely, *MK14* and *FM17*) are realised: the 'maximum information content' of the system remains in a non-realised state [14]. Variety, however, would be added when a paper citing *MK14* would cite *LM12* as well, rather than *FM17*.

Note that the distinction variety/redundancy does not refer merely to the dyadic relation between two papers, but rather to an aggregate network of publications within a given research domain (in the example above, the research domain studying certain diplomatic practices). It is a whole system's structural property rather than just a dyad between a 'citing' and a 'cited work'. As such, there may be manifold concrete triggers behind the system's overall pattern of variety or redundancy. Many authors may have failed to cite *LM12* because they did not do thorough literature research [15,16], or search queries may not have yielded *LM12* due to that paper's usage of terms that differ slightly from *MK14* and *FM17*. Also, it is possible that citational biases were at play as *LM12* was published in a journal of lower visibility not indexed in popular bibliographic databases [17–19], or it is possible that the authors behind *MK14* and *FM17* were well-connected personally within the academic community or well-known due to high institutional positions or prestigious conference presentations, while the author behind *LM12* was not. Alternatively, it could be a mixture of many reasons, a kind of intersectionality of grounds that led to this bibliometric discrimination against *LM12* or the bibliometric privileging of *MK14* and *FM17*. Whatever the actual historical reasons, for the analytical detection of informational variety/redundancy, the concrete, situational sources of these aggregate patterns do not matter.

In addition, while redundancy and variety serve as two opposing sides of a distinction, they are not opposites: the scientific system can increase both its redundancy and its variety simultaneously. A paper can cite both *MK14* and *FM17*, on one hand, and *LM12*, on the other hand. What the system requires is a mixture of both, conveying familiarity to enable the temperate growth of a given knowledge-field, but, at the same, rendering the familiar more interesting [20] by injecting a surprising novelty into it, a new perspective, a new theory, new data, with new scholarly references (see Table 2 for another illustration).

Neither redundancy nor variety bears an inherent value of 'good' or 'bad' [8]. Redundancy is not to be condemned per se, but rather exercises vital functions in communicative settings. If there were no redundancy, that is, if every paper was cited or recommended only once, one would exercise great labour in searching for a specific reference or publication,

Table 3. A typology of research discovery tools ranging from those that reproduce bibliometric redundancy (comprising citation-based and query-based tools) to those that generate bibliometric variety (comprising category-based and randomness-based tools).

Citation-based tools	Query-based tools	Category-based tools	Randomness-based tools
Redundancy		Variety	

rendering an informational loss more probable. Redundancy thus saves time and insures against research waste [21,26–28]. Great variety, however, is not per se to be longed for, but it rather generates an unpredictable system whose capacity may not suffice to carry the immense plurality of unique information within it. Extreme cases of variety would risk high amounts of wasted research [29]. Both one-sided redundancy and one-sided variety ultimately render a system entropic: maximum redundancy would mean entropy due to an almost complete loss of any energising surprises (again, think of the scenario where every publication only cites the same single paper over and over again and nothing else) and maximum variety would mean entropy due to an almost complete loss of predictable structures (again, think of the scenario where every single publication is cited only once and never again).

So far, this section has discussed the concept of variety/redundancy concerning scientific publications in general. One can also direct this lens onto *research discovery tools*, that is, to machines that draw from a set of scientific publications to recommend a selection of them to users. Here again, neither extreme would be wished for. Research discovery would be inadequate if it recommended publications only once and never thereafter (maximum variety), as well as if it always recommended the very same paper and nothing else (maximum redundancy). Like reference patterns, the landscape of research discovery tools should instead find a balance between variety and redundancy.

But what might a typological application of the information-theoretical concepts of redundancy/variety onto research discovery tools look like? The next section will attempt to flesh out such a typology based on that variable.

3. A typology

This section suggests that all research discovery tools can be located within a dimension ranging from bibliometric redundancy to bibliometric variety. At the *redundancy* end of the spectrum are tools that (at least indirectly) reproduce extant reference patterns or semantic linkages between scientific publications. They may be citation-based (drawing from citational proximity) or query-based (drawing from semantic proximity). In contrast, all other tools are at the *variety end* of the spectrum. These variety-generating ones may be category-based (enabling a 'creative browsing' feature) or randomness-based. Table 3 illustrates the typology.

The following will outline each of the types abstractly and offer impressionistic examples drawing from the existing pool of research discovery tools. These examples merely serve as a plausibility probe for the conceptual typology. Moreover, the illustrations here may be weakened because many of these platforms host not just a single function, but rather multiple features, some of which may be redundancy-reproducing, while others are variety-generating – these (and other additional) aspects will be discussed in the section thereafter.

3.1. Redundancy-generating tools

Research discovery tools reproduce redundancy when they strengthen existing inter-publication networks based on citational or semantic links. There are two such subtypes, namely, *citation-based* and *query-based* tools.

3.1.1. Citation-based tools. First, citation-based research discovery tools are most prone to generate bibliometric redundancy, for it is their very function to reproduce extant reference patterns. When *MK14* and *FM17* are heavily linked inter-citationally (to use that semi-fictitious example from above), then a citation-based research discovery tool receiving the user-input *MK14* will output *FM17*. At the same time, it will remain mute on *LM12* despite its topical relevance. Again, the tool does so because it merely reproduces pre-existing inter-citational links.

Such citation-based tools thus indulge in 'the gravity of highly cited papers' [30]. The user-input comes in the form of a token for a scientific publication (or a set of multiple publications), such as by having the user insert a publication's title or Digital Object Identifier (DOI). The tool then looks for the citations to or from that publication. It may also detect

second-order (or *n*th-order) references, that is, references to publications that cite or are cited by those publications which are citationally inter-linked to the user-input's publications. Examples for citation-based tools are *CitationGecko, ConnectedPapers* or *CoCites*. Many query-based search engines likewise contain default settings that reproduce reference patterns, such as *Google Scholar's* standard approach of listing results not only based on their semantic relevance to the user-input but also based on their citational impact.

3.1.2. Query-based tools. The second type of research discovery tool at the redundancy end of the spectrum is query-based. Many examples of this are, in effect, search engines that require keywords to search for papers containing similar strings. The most frequently used tool, *Google Scholar*, is primarily built around this function. With varying functionalities (e.g. advanced search options, Boolean operators and semantic web searches), other databases, likewise, work in this fashion: *Dimensions, LENS, Science Open, Scopus, Semantic Scholar, Web of Science* and so on, mainly (but not exclusively) revolve around presenting a search bar into which users type their queries. Beyond search engines, some recommender systems may likewise draw from semantic distance formulae to narrow down their paper recommendations to users (e.g. through real-time feedback query expansion techniques, cf., [31]). In addition, tag-based platforms that enable users to attribute categorical labels to scientific publications in a manner of a 'collaborative filtering' operate on the same principle. Such tools (like *CiteULike*) aggregate semantic commonalities across research outputs [32]. Whatever the precise approach, query-based tools are similar to citation-based platforms in that they are prone to reproduce bibliometric redundancy; for semantic similarities may serve as a proxy for extant inter-publication links.

User profile-based recommender systems likewise fall within this redundancy end of the spectrum; they may either follow the logic of co-citations or semantic proximities. Their main difference is that their user-input does not come in the form of actively inserted information of publications but rather in algorithmically generated user profiles. For instance, *Mendeley Suggest* [33] or *ResearchGate*'s recommender system look at the papers stored (or saved or 'liked') by a user to recommend proximate publications. As it is their very task to generate user-near suggestions, they cannot but reproduce bibliometric redundancy – be they based on citational links or on semantic nearness.

To be sure, such discovery platforms *can* recommend papers that have been hitherto unknown to specific users in a concrete setting, especially if they take publications from a 'long tail' [34] of possible outputs. The recommendations are then indeed 'surprises' to an individual user, which might lead one to believe that these tools enhance variety, not redundancy. However, from an aggregate viewpoint of informational flows, they nevertheless reproduce redundancy as they draw from a system-wide 'pre-existing knowledge basis' [35] in the forms of extant linkages to a given citation, keyword or user profile. It is thus that they reproduce aggregate bibliometric redundancy across the system, even if it may seem variety-enhancing to a single user.

3.2. Variety-generating tools

We have seen how citational proximity and semantic similarity reproduce redundancy across the scientific system. On the other side are tools that enhance variety; but how do these variety-generating research discovery tools operate? There can be two basic approaches – on one hand, *category-based* tools that enable users to browse through research outputs, and on the other hand, *randomness-based* tools that are almost completely detached from any citational biases, semantic links or common categories.

3.2.1. Category-based tools. First, a variety-enhancing research discovery tool may be based on the selection of certain categories. Such category-based platforms often facilitate 'creative browsing' through publication lists that happen to be bracketed together under a specific category [36,37]. There is a multitude of such instruments, but they are seldom analysed as fully fledged research discovery tools, even though they serve the same function of research discovery, that is, the conveying of links to scientific publications. Even a mere CV or social media profile of an individual scholar [38] can be subsumed under this category: they enable a profile-based discovery of research. To use a fictitious example, the author of *LM12* (which remained uncited) may have another publication, *LM15* (which is highly cited), so that when people browse through that person's CV (perhaps because they came across that person through *LM15*), they might encounter the unheard-of *LM12* and start citing it over time [39,40]. As *LM12* had remained uncited before, these novel citations will insert bibliometric variety into the scientific system, which, in turn, was rooted previously in someone browsing *LM*'s CV – that is, rooted in a variety-enhancing, category-based research discovery tool.

There are many other examples of category-based tools: institutional repositories operate on the same basis; they allow for institution-based research discovery. A more common way to explore publications uses a journal-level approach, that is, browsing through a specific outlet's table of contents, either in print or digitally (an experience that can be enhanced through RSS feeds, email alerts or APIs). Some platforms aggregate such tables of contents from a greater set of journals; examples include *Current Contents* and *JournalTOCs* and other so-called *Current Awareness Services* [41]. Once a specific research domain defines that set of journals, we have discipline-based platforms such as those narrowed down to Criminology (e.g. *Criminology Papers*) or Philosophy (e.g. *Philosophy Paperboy*). These tools then regularly send out titles (and abstracts) of the latest papers published in a specific research domain to their users.

While many category-based platforms simply look at the *latest* papers based on their publication dates, others may incorporate additional indicators, such as whether the publications of a specific research domain published in a certain date range (e.g. the past 30 days) are 'trending'. For instance, *OOIR* lists the latest papers from political science and related domains and ranks them according to their *Altmetric Attention Scores* [42] based on social media mentions and news attention. *OOIR* thus harnesses altmetrics for variety-enhancing purposes [43].

Discovery tools of this type list the latest or trending papers according to a pre-selected category. That chosen unit may be defined by a certain scholar, by a shared research institution, by a common journal or a set of journals or by a research domain. The more general the categorial unit, the less redundant the research discovery tool. The knowledge of the category does not reveal knowledge of specific publications anymore, as opposed to co-citation patterns. Most importantly, the listed outputs of these category-based research discovery tools are detached from extant reference networks. One may browse around and encounter *LM12* (which had remained uncited before) alongside highly cited articles, and this discovery of *LM12* structurally heightens the probability of it being cited after that. Such tools thereby do not reproduce bibliometric redundancy, but tendentially enhance citational variety instead.

However, even category-based platforms are not completely free from biases conducive to an indirect redundancy [44]. They at least implicitly demand that the user has specified a more or less detailed category, such as an individual scholar, a specific set of journals or a certain scientific discipline, to which the user may already have developed a systemic affinity. In addition, browsing-featured tools usually list the *latest* papers – but if bibliometric patterns bear a biased citation life cycle towards newer publications to the detriment of older ones [45,46], then even these platforms reproduce this tendency, and thereby serve to nourish at least a small degree of systemic redundancy. Tools that list *trending* papers perhaps likewise incorporate a systemic bias whose structure is reflected in citation patterns, such as that papers in highly cited (and better-networked) 'top journals' will more probably be 'trending' on social media and news platforms [47,48]. Thus, even category-based platforms may not only be conducive to variety, but also indirectly contribute to redundancy.

3.2.2. Randomness-based tools. In contrast, an even more variety-enhancing research discovery platform would list new *and* old, trending *and* non-trending, discipline-specific *and* discipline-deviating papers with equal probability. Such truly variety-enhancing tools would be based on a large sample of scientific publications and a high degree of randomness [30].

Such a tool may not even exist yet (if not life itself, or at least social media networks, are to be seen as a grand field for random research discovery, as some works on serendipity implicitly suggest; cf., [49–51]). A fictitious example would be a huge corpus of the scientific literature in all research domains covering various centuries that would regularly (e.g. hourly) spit out a random publication of this corpus through, say, a Twitter bot. It may link to a philosophical paper on epistemology from a German journal of the 1820s, and in the next instance, a medical publication from a low-ranked journal from the early 2000s, then a highly cited sociological work from the 1970s, before finally recommending this very article you are reading right now. The pattern of outputs should be highly unpredictable, with every new recommendation equally improbable as any other ones, finally enabling a truly 'blind variation' [52,53] that would constantly nourish the further evolution of science. Only such a machine would ensure freedom from extant biases, thus serving as a research discovery platform that is veritably variety-enhancing for the whole system of science. It would then be up to the system's evolution – or, as earlier centuries would say, up to individual geniuses and their sagacity – to fetch these random surprises and to de-randomise them into scientific knowledge [8,54].

While many 'discount serendipity because it is not viewed as a formal search strategy' [55], variety-enhancing tools offer systemic benefits. As they do not require pre-meditated queries or proxies thereof, they can be utilised even with a 'vague, fuzzy, or even unspoken information need, when users do not quite know what they're looking for' [30]. Skimming lists of papers based on some arbitrary categories or immersing oneself into total serendipity by harnessing the randomness of life might all enable unexpected and surpriseful 'information encounters' [56], and thus produce innovative, bibliometrically measurable ties.

Table 4 summarises the typology.

Table 4. Two types of research discovery tools (redundancy-reproducing and variety-generating), with two subtypes each.

Type	Subtype	Examples
Redundancy-reproducing	Citation-based	"Cited by"-lists, *CoCites, CitationGecko, ConnectedPapers*
	Query-based	Academic search engines (e.g., core function of *Google Scholar*), or recommender systems based on corpora of papers (e.g., *Mendeley Suggest*)
Variety-generating	Category-based	CVs, journals' tables of contents, institutional repositories, *OOIR*
	Randomness-based	Fictitious example of a 'random paper bot' drawing from a huge corpus of publications from multiple disciplines and centuries

CV: Curriculum Vitae; OOIR: Observatory of International Research.

4. Discussion

This article proposed a conceptual typology of all possible research discovery tools based on a dimension ranging from bibliometric redundancy (the reproduction of extant citational links between scientific publications) to variety (the generation of novel ties between research outputs).

The universality of the redundancy/variety-distinction is underlined by its ability to subsume highly heterogeneous exemplary tools, including a journal's table of contents, a researcher's CV, scholarly databases, academic search engines, institutional repositories or bots that regularly recommend random scientific works to its users. This typology thus departs from previous attempts to categorise scientific discovery services, as they had done so with a narrower focus by looking, for instance, only at query-based approaches [57]. An overarching typology seemed not to have existed yet, and it is this conceptual gap that this article seeks to fill.

Such a conceptual typology may aid our understanding of the broader role of research discovery. System-wide citation patterns begin with the way researchers find publications to skim, read, interpret, evaluate and cite [58]. The algorithms of research discovery tools and information retrieval systems, thus, gradually grow into aggregated reference networks within the scientific system. Their design may thereby generate greater or lesser system-wide entropy.

It is possible, for instance, that the most frequently used research discovery tools are shaped such that they tend to reproduce redundancy by narrowing down the element of systemic surprise. This rather speculative assumption of a current retrieval paradigm may be propped up by the observation that the 'ideal' type of a variety-enhancing tool does not even seem to exist yet: there is not yet the fictitious bot that spits out random scientific publications from a huge corpus from centuries of publications of which each work has an equal probability of being listed, regardless of the publication's age, discipline, authors, citational impact or publication outlet. In addition, one of the most frequently used machines for research discovery, *Google Scholar* [59], is by default a query-based and, therefore, a redundancy-reproducing one (it is even partly citation-based, as it sorts results based on algorithms that take into account the number of citations by default).

In effect, if researchers only used, say, *Google Scholar* and *Mendeley Suggest* to discover relevant publications, they would probably reproduce bibliometric redundancy – as is visible in the standard assumption of a Pareto distribution in bibliometrics, according to which 80% of all citations go to just 20% of all research outputs ([59–61]; but note that there is evidence of a gradual change in this pattern, cf. [62]). This 'can end up reinforcing habits rather than exposing students and researchers to new information, sharply limiting the researcher's view of the world' of scientific papers [30]. They would only discover publications whose content does not deviate too much from a semantic or co-citational closeness to the pre-specified user-input (e.g. a typed-in query of keywords or a list of publications one has already read). Finding only papers of topical proximity, such an approach would minimise the element of surprise, and thereby inhibit innovative cross-pollination from distant fields (see also studies on 'disruptive' publications, e.g. [63–65]). The scientific system would only contain a slow tempo of variety, with only rare energisation through new information and original reference patterns. While those instruments are not 'bad' as such, the assumption that the dominant landscape of research discovery is redundancy-reproducing might point to the need for consciously complementing them with other, variety-enhancing tools. The focus should be on *complementary* combinations, as the limitations of one search strategy can be overcome with other paradigms of information retrieval – there can be, in other words, 'no single web service that satisfies all demands' [57].

The typology developed above might contain two core limitations: first, actual research discovery tools often harbour multiple functions, some of which are redundancy-reproducing while others are variety-enhancing, thus rendering it

difficult to identify their precise place in the typology. A journal's website may show both the latest papers (which is variety-enhancing) and its most-cited ones (which is redundancy-reproducing). Many academic search engines sort their results by 'impact' measures such as citation counts (which is redundancy-reproducing) but allow users to sort by date (which is variety-enhancing). *Google Metrics* (which is built in to *Google Scholar*) enables a category-based browsing tool (which is variety-enhancing), whereby each publication list is ordered by a journal-level *h*-index, and thus, by citational impact (which is redundancy-reproducing). To make it more complicated and to repeat a statement from above, redundancy and variety are not veritable opposites. Both can be heightened at the same time – classification schemes and thesauri, such as the Medical Subject Headings (MeSH) in *PubMed* [66,67], possibly exemplify a combination of both semantic-based (thus redundancy-reproducing) and category-based (thus variety-enhancing) venues. Given the heterogeneous multi-functionality of many research discovery platforms, the typology developed in this article should perhaps claim to cover merely snippets or single functions of research discovery tools, rather than whole platforms in their entirety.

A second limitation is that this conceptual typology remains blind towards the actual, concrete, situational reasons behind a given scientific discipline's aggregate structure of redundancy or variety. In an analogy, a statistical description of the frequency of letters in the English language would not in itself explain the historical causes behind why 'e' occurs much more often than the letter 'q' (meaning that the appearance of 'e' reproduces informational redundancy, while that of 'q' generates variety) [6]. A high degree of redundancy may arise due to flawed literature searches, due to language barriers [68,69], variation in academic networking [70], structural gender-related biases [71,72] or editors violating the norms of publication ethics [73]. Whatever the specific reason, the typology remains blind to the concrete triggers.

It is this blind spot which further research could tackle in order to study practical aspects behind the conceptual typology. One could analyse to what extent the designs of research discovery tools really shape the broader network of citational interlinkages in given scientific areas. Experiments could be devised to expose researchers only to specific discovery tools and see how this treatment affects their citation patterns in subsequent publications. As regards technical implementations, future efforts could generate a truly variety-enhancing tool like that fictitious random paper bot outlined above, and see how it affects scientists' serendipitous drive for innovation. Whatever the next direction, much discovery seems to await the research on research discovery.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The authors acknowledge TU Wien Bibliothek for financial support for editing/proofreading.

ORCID iD

Andreas Nishikawa-Pacher https://orcid.org/0000-0001-5149-6294

References

[1] Martín-Martín A, Thelwall M, Orduna-Malea E et al. Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: a multidisciplinary comparison of coverage via citations. *Scientometrics* 2021; 126: 871–906.
[2] Foster AE and Ellis D. Serendipity and its study. *J Doc* 2014; 70: 1015–1038.
[3] Rousseau R and Zuccala A. A classification of author co-citations: definitions and search strategies. *J Am Soc Inf Sci Tec* 2004; 55: 513–529.
[4] van Eck NJ and Waltman L. Citation-based clustering of publications using CitNetExplorer and VOSviewer. *Scientometrics* 2017; 111: 1053–1070.
[5] Collier D, LaPorte J and Seawright J. Putting typologies to work: concept formation, measurement, and analytic rigor. *Polit Res Quart* 2012; 65: 217–232.
[6] Shannon CE. Prediction and entropy of printed English. *Bell Syst Tech J* 1951; 30: 50–64.
[7] Bateson G. *Steps to an ecology of mind*. Chicago, IL: University of Chicago Press, 1999.
[8] Luhmann N. *Die Wissenschaft der Gesellschaft*. Frankfurt am Main: Suhrkamp, 1992.
[9] Ashby WR. Requisite variety and its implications for the control of complex systems. In: Klir GJ (ed.) *Facets of systems science*. Boston, MA: Springer, 1991, pp. 405–417.

[10] Acemoglu D, Akcigit U and Kerr WR. Innovation network. *Proc Natl Acad Sci USA* 2016; 113: 11483–11488.
[11] Kuus M. *Geopolitics and expertise*. Chichester: Wiley, 2014.
[12] McConnell F. Liminal geopolitics: the subjectivity and spatiality of diplomacy at the margins. *T I Brit Geogr* 2017; 42: 139–152.
[13] Mikalayeva L. Reporting under international conventions: a genre analysis. *Hague J Dipl* 2012; 7: 287–312.
[14] Leydesdorff L, Johnson MW and Ivanova I. Toward a calculus of redundancy: signification, codification, and anticipation in cultural evolution. *J Assoc Inf Sci Tech* 2018; 69: 1181–1192.
[15] Bornmann L and Daniel H. What do citation counts measure? A review of studies on citing behavior. *J Doc* 2008; 64: 45–80.
[16] Milard B and Tanguy L. Citations in scientific texts: do social relations matter? *J Assoc Inf Sci Tech* 2018; 69: 1380–1395.
[17] Tijssen RJW, Mouton J, van Leeuwen TN et al. How relevant are local scholarly journals in global science? A case study of South Africa. *Res Evaluat* 2006; 15: 163–174.
[18] Vera-Baceta M-A, Thelwall M and Kousha K. Web of Science and Scopus language coverage. *Scientometrics* 2019; 121: 1803–1813.
[19] Acharya A, Verstak A, Suzuki H et al. Rise of the rest: the growing impact of non-elite journals. *arXiv* 2014: 14102217.
[20] Davis MS. That's interesting! Towards a phenomenology of sociology and a sociology of phenomenology. *Philos Soc Sci* 1971; 1: 309–344.
[21] Garfield E. Citation analysis as a tool in journal evaluation. *Science* 1972; 178: 471–479.
[22] Merton RK. The Matthew effect in science: the reward and communication systems of science are considered. *Science* 1968; 159: 56–63.
[23] Rossiter MW. The Matthew Matilda effect in science. *Soc Stud Sci* 1993; 23: 325–341.
[24] Margulis L. *Symbiosis in cell evolution*. San Francisco, CA: W. H. Freeman & Co Ltd, 1981.
[25] al-Khwarizmi M. *Al-Jabr*. Baghdad, 820. English translation in: Rosen, F. (ed.). *The algebra of Mohammed ben Musa*. London: Oriental Translation Fund, 1831.
[26] Leydesdorff L. Theories of citation? *Scientometrics* 1998; 43: 5–25.
[27] Min C, Ding Y, Li J et al. Innovation or imitation: the diffusion of citations. *J Assoc Inf Sci Tech* 2018; 69: 1271–1282.
[28] Riviera E. Scientific communities as autopoietic systems: the reproductive function of citations. *J Am Soc Inf Sci Tec* 2013; 64: 1442–1453.
[29] van Leeuwen TN and Moed HF. Characteristics of journal impact factors: the effects of uncitedness and citation distribution on the understanding of journal impact factors. *Scientometrics* 2005; 63: 357–371.
[30] Maloney A and Conrad L. *Expecting the unexpected: serendipity, discovery, and the scholarly research process*. Thousand Oaks, CA: SAGE, 2016.
[31] Khalid S, Wu S, Alam A et al. Real-time feedback query expansion technique for supporting scholarly search using citation network analysis. *J Inf Sci* 2021; 47: 3–15.
[32] Heck T, Peters I and Stock WG. Testing collaborative filtering against co-citation analysis and bibliographic coupling for academic author recommendation. In: *Proceedings of the 3rd ACM RecSys' 11 workshop on recommender systems and the social web*, Chicago, IL, 23 October 2011.
[33] Ransom S. Mendeley's vision for supporting researchers. *Mendeley Blog*, 15 May 2019, https://blog.mendeley.com/category/mendeley-suggest/ (accessed 6 November 2020).
[34] Liu Y, Xiong Q, Sun J et al. Topic-based hierarchical Bayesian linear regression models for niche items recommendation. *J Inf Sci* 2019; 45: 92–104.
[35] Vasconcelos AC, Martins JT, Ellis D et al. Absorptive capacity: a process and structure approach. *J Inf Sci* 2019; 45: 68–83.
[36] Bawden D, Foster A and Rafferty P. Encountering on the road to Serendip? Browsing in new information environments. In: Foster A and Rafferty P (eds) *Innovations in information retrieval*. London: Facet Publishing, 2011, pp. 1–22.
[37] O'Connor B. Fostering creativity: enhancing the browsing environment. *Int J Inf Manag* 1988; 8: 203–210.
[38] Delfanti A. The financial market of ideas: a theory of academic social media. *Soc Stud Sci* 2021; 51: 259-276.
[39] Fang H. A transition stage co-citation criterion for identifying the awakeners of sleeping beauty publications. *Scientometrics* 2019; 121: 307–322.
[40] van Raan AFJ. Sleeping beauties in science. *Scientometrics* 2004; 59: 467–472.
[41] Loesch MF. JournalTOCs. *Techn Serv Q* 2012; 29: 89–90.
[42] Priem J, Taraborelli D, Groth P et al. Altmetrics: a manifesto, 2010, http://altmetrics.org/manifesto/ (accessed 13 August 2021).
[43] Holbrook JB. Designing responsible research and innovation to encourage serendipity could enhance the broader societal impacts of research. *J Responsib Innov* 2019; 6: 84–90.
[44] Kaltenbrunner W and de Rijcke S. Filling in the gaps: the interpretation of curricula vitae in peer review. *Soc Stud Sci* 2019; 49: 863–883.
[45] Adams J. Early citation counts correlate with accumulated impact. *Scientometrics* 2005; 63: 567–581.
[46] Cano V and Lind N. Citation life cycles of ten citation classics. *Scientometrics* 2005; 22: 297–312.
[47] Ferreira Araujo R. Communities of attention networks: introducing qualitative and conversational perspectives for altmetrics. *Scientometrics* 2020; 124: 1793–1809.

[48] Kelly BS, Redmond CE, Nason GJ et al. The use of Twitter by radiology journals: an analysis of Twitter activity and impact factor. *J Am Coll Radiol* 2016; 13: 1391–1396.
[49] Panahi S, Watson J and Partridge H. Information encountering on social media and tacit knowledge sharing. *J Inf Sci* 2016; 42: 539–550.
[50] Robinson-Garcia N, van Leeuwen TN and Ràfols I. Using altmetrics for contextualised mapping of societal impact: from hits to networks. *Sci Publ Policy* 2018; 45: 815–826.
[51] Zhou X, Sun X, Wang Q et al. A context-based study of serendipity in information research among Chinese scholars. *J Doc* 2018; 74: 526–551.
[52] Campbell DT. Blind variation and selective retentions in creative thought as in other knowledge processes. *Psychol Rev* 1960; 67: 380–400.
[53] Simonton DK. Creativity as blind variation and selective retention: is the creative process Darwinian? *Psychol Inq* 1999; 10: 309–328.
[54] Caws P. The structure of discovery. *Science* 1969; 166: 1375–1380.
[55] Race TM. Resource discovery tools: supporting serendipity. In: Pagliero Popp M and Dallis D (eds) *Planning and implementing resource discovery tools in academic libraries*. Hershey, PA: IGI Global, 2011, pp. 139–152.
[56] Erdelez S. Information encountering: it's more than just bumping into information. *Bull Am Soc Inf Sci Technol* 1999; 25: 26–29.
[57] Kim J-J and Rebholz-Schuhmann D. Categorization of services for seeking information in biomedical literature: a typology for improvement of practice. *Brief Bioinform* 2008; 9: 452–465.
[58] Crestani F, Lalmas M, Rijsbergen CJV et al. 'Is this document relevant? . . . Probably': a survey of probabilistic models in information retrieval. *ACM Comput Surv* 1998; 30: 528–552.
[59] Alotaibi F and Johnson F. Why we like Google Scholar: postgraduate students' perceptions of factors influencing their intention to use. *Aslib J Inform Manag* 2020; 72: 587–603.
[60] Nisonger TE. The '80/20 Rule' and core journals. *Serials Libr* 2008; 55: 62–84.
[61] Abramo G, D'Angelo CA and Soldatenkova A. The dispersion of the citation distribution of top scientists' publications. *Scientometrics* 2016; 109: 1711–1724.
[62] Néda Z, Varga L and Biró TS. Science and Facebook: the same popularity law! *PLoS ONE* 2017; 12: e0179656.
[63] Bornmann L, Devarakonda S, Tekles A et al. Are disruption index indicators convergently valid? The comparison of several indicator variants with assessments by peers. *Quant Sci Stud* 2020; 1: 1242–1259.
[64] Leydesdorff L, Tekles A and Bornmann L. A proposal to revise the disruption index. *Prof Inform* 2021; 30: e300121.
[65] Wu L, Wang D and Evans JA. Large teams develop and small teams disrupt science and technology. *Nature* 2019; 566: 378–382.
[66] Lu Z, Kim W and Wilbur WJ. Evaluation of query expansion using MeSH in PubMed. *Inf Retr* 2009; 12: 69–80.
[67] Zhang Y, Chen Q, Yang Z et al. BioWordVec, improving biomedical word embeddings with subword information and MeSH. *Sci Data* 2019; 6: 52.
[68] Corcoran J. Addressing the 'bias gap': a research-driven argument for critical support of plurilingual scientists' research writing. *Writ Commun* 2019; 36: 538–577.
[69] Shchemeleva I. 'There's no discrimination, these are just the rules of the game': Russian scholars' perception of the research writing and publication process in English. *Publications* 2021; 9: 8.
[70] Guan J, Yan Y and Zhang JJ. The impact of collaboration and knowledge networks on citations. *J Informetr* 2017; 11: 407–422.
[71] Holman L, Stuart-Fox D and Hauser CE. The gender gap in science: how long until women are equally represented? *PLoS Biol* 2018; 16: e2004956.
[72] Hardt H, Smith AE, Kim HJ et al. The gender readings gap in political science graduate training. *J Polit* 2019; 81: 1528–1532.
[73] Pacher A, Heck T and Schoch K. Open editors: a dataset of scholarly journals' editorial board positions. *SocArXiv* 2021.