Information Legacy of Eugene Garfield: from the Chicken Coop to the World Wide Web

A report by Wendy Warr on a symposium presented at the 255th ACS National Meeting and Exposition in New Orleans, LA, on March 19, 2018

Eugene Garfield: the man and his legacy

Bonnie Lawlor. Retired, Radnor, Pennsylvania, United States

Eugene Garfield was a complex and remarkable man. He played the role of employer, mentor, friend, and role model to many people around the world and throughout his lifetime he was unwaveringly loyal to those who came to know him well. He also left a legacy far beyond the concept of citation indexing and bibliometrics, and the ideas that he developed during the latter half of the last century continue to fuel advances in cheminformatics and information science.

The son of immigrants from Europe, Garfield was born in 1925 in New York. He was raised by a Jewish mother. He graduated with a degree in chemistry from the University of Columbia in 1948, and afterwards took up a post as a laboratory assistant with Prof. Louis P. Hammett. Later, by accident, he stumbled on the sessions of the Division of Chemical Literature (now ACS CINF) at the spring 1951 ACS meeting in New York, where the work of IBM on punch cards fascinated him. An encounter at that meeting led to Garfield's taking a job at the Welch Medical Library at Johns Hopkins University.

At Johns Hopkins he became steeped in the details of *Chemical Abstracts* and other abstracting services. At the Welch project he worked on the chemical nomenclature used in the Medical Subject Headings (MESH), and he grew an understanding of the needs for new approaches to retrieving chemical information. During his two years there (1951-1953) he developed three concepts behind the future Institute for Scientific Information (ISI): content page services, chemical indexing (abstracting and indexing services had a three- year backlog at that time), and the use of references for indexing (i.e., citations).

Shepard's Citations was already used in United States legal research, providing a list of all the authorities citing a particular case, statute, or other legal authority, but it was not thought that a science citation index was feasible. A "Eureka moment" occurred when further funding for the Welsh project was denied, and Garfield was fired for his out-of-hours indexing efforts. He attended library school at Columbia from 1953-1954, and then wrote his seminal paper on citation indexes in science.[1]

In 1954 he became a documentation consultant and adopted the business name Eugene Garfield Associates. In 1955 he launched *Management DocuMation Preview*, the precursor of *Current Contents*. In 1956 he incorporated as DocuMation, Inc., and gained a contract with Bell Laboratories which gave him enough money to develop other products. Garfield became involved with the pharmaceutical industry, and began *Current Contents of Pharmacomedical, Chemical and Life Sciences* in 1956. In 1958 *Current Contents* was made a subscription service: the first science-based contents page service.

Garfield signed up for a Ph.D. at the University of Pennsylvania in 1955, and this culminated in his 1961 dissertation: *An Algorithm for Translating Chemical Names to Molecular Formulas* (http://www.garfield.library.upenn.edu/essays/v7p441y1984.pdf).[2] He realized that novel compounds could be easily located in the chemical literature, and proposed to launch a current awareness service based on compounds rather than citations. The National Science Foundation and Smith Kline & French would not supply upfront money, but 12 pharmaceutical companies offered to subscribe if a product were launched. *Index Chemicus* was launched in 1960, and Garfield's company was incorporated as the Institute for Scientific Information.

Garfield's initial efforts to obtain government funding to create a citation index were not successful, in part, because he was not affiliated with an academic institution, but Gene did attract the interest of Joshua Lederberg who had recently received the Nobel Prize, and the *Genetics Citation Index*, was eventually supported with public money. Due to later changes in the rules that prohibited NIH giving grants to companies, the money had to be transferred to NSF and converted to a contract. Gene was tenacious once he had an idea. *Genetics Citation Index* was launched in 1963; *Science Citation Index* followed.

When *Genetics Citation Index* was published in 1963, all three of Garfield's concepts, contents page services, chemical Indexing, and citation indexing, had become a reality. Later, three vice-presidents of ISI asked Garfield to leave, but eventually *they* left and Garfield carried on. He began writing *Essays of an Information Scientist* in 1962: ISI Press was a "vanity press". The 1970s saw the dawn of the online era, with Dialog, SDC Orbit, BRS, DataStar, DIMDI, and ESA. ISI launched *Social Sciences Citation Index*, and *Journal Citation Report*, and the journal Impact Factor was born. ISI gained its own office building in Market Street, Philadelphia, with punched card decoration on the outside.

The 1980s saw the dawn of personal computing, and ISI launched *Current Contents* on Diskette, *Index Chemicus* Personal Database, and Citation Indexes on CD ROM. The *Scientist* magazine was a departure from the three concepts. ISI's fortunes changed and the board insisted on a chief operating officer being brought in. Vice-presidents were fired, but revenues did not increase. Eventually Garfield bought the company back. Four years later it was sold to Thomson, and Garfield became president emeritus.

Lawlor concluded with a personal tribute to Garfield. She was an early and long-term employee of ISI who ultimately had the blessing of counting Eugene Garfield among her friends. ISI was an energizing and adventurous place to work with a whimsical side as well. Garfield was not a bureaucratic manager. The place was crazy. People parked their motorcycles at their desk. There was no dress code. One person used to love to wear baby-doll pajamas to work, one of the senior directors had a sapphire on his forehead, and a teddy bear on his belt, and those two people, Baby Doll and Sparkles, streaked at one of the company events. But beneath that surface, it was a very energizing, intellectually challenging environment. Garfield encouraged every one of his employees us to do their best with the talents God gave them, and to contribute what they could to the betterment of the company.

From the Index Chemicus Registry System to SciFinder and beyond

Wendy A. Warr. Wendy Warr & Associates, Holmes Chapel, Cheshire, United Kingdom

Warr's first job in chemical information was in the Experimental Information Unit at the University of Oxford where she learned to code structures into Wiswesser Line Notation (WLN)[3] as part of a project looking into the feasibility of making the Index Chemicus Registry System (ICRS) available to U.K. universities. She did not meet Gene Garfield until 1976, at an ICRS Users meeting in London, but after that kept in regular touch with him in person or by email, until he was in his eighties. She paid tribute to his inspiring and charismatic character, and was honored to have been mentioned as a colleague in one of his publications.[4]

Garfield is best known as a pioneer of citation analytics, but some of us remember him as an early leader in the indexing of chemical information. *Index Chemicus* (*IC*) was launched in 1960, four years before the official launch of the Science Citation Index. It began as a hard copy current awareness service with fragment codes, molecular formulas, and structural diagrams. In 1962, WLNs were added to *IC*. ICRS was launched in 1968, followed by Chemical Substructure Index (CSI), a permuted index of WLNs, in 1971. Other products in the ISI chemistry line followed: Automatic New Structure Alert (ANSA) in 1971 and *Current Chemical Reactions* (*CCR*) in print in 1979.

From 1984 to January 1987 *IC* was substructure searchable online under the system "Description, Acquisition, Retrieval, and Correlation" (DARC).[5,6] CCR was first delivered as an in-house database under the Reaction Access System REACCS[7] in 1986. After Thomson Reuters acquired ISI in 1992, Index Chemicus data from 1993 onwards have been made available in Web of Science from 1993 until today. CCR data from 1993 onwards are also in Web of Science. The same IC and CCR data are also offered in-house through BIOVIA Direct (http://accelrys.com/products/collaborative-science/biovia-direct/).

Garfield has described some of the fundamentals behind ICRS.[8] In 1958 he had shown that chemical names could be converted to molecular formulas (http://www.garfield.library.upenn.edu/essays/v7p441y1984.pdf). He said that there was no need to name compounds. (CAS, however, was laboriously naming compounds.) Journals provided molecular formulas for new compounds, so these compounds could be easily located. At that time there had been publicity about some key intermediates that CAS had missed, and indexes to the literature were very out of date, but Garfield foresaw that monthly MF molecular formula indexes for new compounds could easily be produced. He also realized that printing the structures for these compounds was essential. In addition, he knew that he did not need to abstract all the known chemical literature since Bradford's Law suggested that just 100 journals contained 95% of the new compounds. (See the later talk by Jim Testa for comments on Bradford's Law.)

Two early methods for storing and retrieving WLNs were ICI's Computerized Retrieval of StructureS Based On Wiswesser (CROSSBOW)[9,10] and ISI's Retrieval and Automatic Dissemination of Information from the *Index Chemicus* and Line notation (RADIICAL).

From 1969 to 1973 ICRS tapes were used in-house for a current awareness service by Literature Services Section at ICI Pharmaceuticals. In1973 the use of ICRS tapes transferred to Data Services Section. There were several reasons for the move. Literature Services Section preferred *Chemical Abstracts* to ICRS. Many ICI chemists swore by *Chemical Abstracts* because it covered more journals;

they did not understand ISI's novel compound advantages. Literature Services Section had little expertise in WLN and it was time consuming to formulate ICRS searches. Chemists did not understand the WLNs output by ICRS, and had to go to the library to get a chemical structure. Chemists also disliked the false drops that arose from searching notations. Moreover, they were already scanning *Current Abstracts of Chemistry & Index Chemicus* (*CAC&IC*) in hard copy and they could see no added value in monthly ICRS tape searches.

Data Services Section, on the other hand, was highly experienced in the use of WLN and CROSSBOW, and in that section searches of in-house compounds, and the literature (as in ICRS), and commercially available compounds could be done with the same system. CROSSBOW was used with ICRS from 1973 until about 1980, and ICI worked closely with ISI on software improvements, WLN checking, error correction etc. A full CROSSBOW database for ICRS 1960-1979 was, however, never built, because graphics-based systems arrived on the scene.

The CROSSBOW system carried out three types of structure search: search of fragment codes, line notation strings, and the full atom by atom topology represented by a connection table. Fragment codes and a CROSSBOW connection table were generated from each WLN. Full atom by atom search of the whole database was too slow on the computers used in that era, so initial fragment screening and WLN search were used to screen out compounds before the atom by atom search.

A key feature of the CROSSBOW search was structure display: the chemists were given a set of cards, each the size of about half an A4 page, with a computer-generated structure on it (albeit crude by today's standards), plus all the relevant bibliographic information, the molecular formula and the WLN. Moreover, an information professional could remove false drops from the card deck before it was handed over to the end user chemist. In 1960 Garfield had had the inspiration to deliver the structures that chemists wanted to see; his staff cut the required sections out of journals to produce *CAC&IC*. The electronic version, ICRS, did not have chemical structures. There was also no atom by atom search in RADIICAL and its fragment code was less sophisticated than that used by CROSSBOW.

Warr summarized the current state of the art related to Garfield's ICRS fundamentals: chemical nomenclature; structure display; currency, and Bradford's Law. Generating names from structures, and structures from names, is now a solved problem, but sixty years ago, Gene sowed the seeds (see the talk by Sayle elsewhere in this report). Garfield knew well the disadvantages of having to name compounds manually. (He was, incidentally, a volunteer abstractor for *Chemical Abstracts* in the early 1950s when he was working at Johns Hopkins.) Few chemists want to, or are able to name compounds any more. Even CAS now uses software to help its nomenclature experts. IUPAC realized years ago that the IUPAC International Chemical Identifier (InChI, https://www.inchi-trust.org/) had advantages over traditional nomenclature. There are no InChIs stored in CAS REGISTRY, but InChI entry is an option in SciFinder, and CAS has now joined the InChI Trust.

All today's systems are graphics- and structure-based. CAS has carried out a lengthy exercise improving all the structures displayed by SciFinder^{n}. Chemical structure input has moved from PSIDOM in STN Express, through CASDraw to ChemDoodle.

CAS now prides itself on its currency, especially with patents. The core patent authorities (US, WO, EP, DE, GB, FR, RU, JP, CA) have their first page bibliographic data online within 2 days, and all of the indexed concepts and compounds have been analyzed, registered and indexed by CAS within 27

days of the publication date. Currency for CN, KR and IN is also strong. With regards to the journal literature, a search on STN suggests that 50% of articles in major journals in January 2018 were indexed by March 1, and 75% were indexed by May 31.

As for Bradford's Law, Warr would have liked to discuss the choice of core journals for full, manual indexing in Reaxys, but time did not permit such digressions. Just as ICI discovered in 1973, chemists still have the ingrained (if rather misguided) belief that "if CAS doesn't have it, it doesn't exist".

Another theme that has developed over the years is integration and linking. Structures are virtually useless unless they are linked to data and bibliographies. Back in the 1970s ICI was able to carry out substructure search of the in-house chemical database, supplier catalogs, and literature information with just one software system, albeit the three separate databases had to be separately searched. From the 1980s on, graphics-based systems gradually came into common use, and substructure searching was opened up to end users, but the integration of in-house and external databases has continued to be a thorn in the flesh. APIs for integration of in-house data with Reaxys and SciFinder started to be adopted by some companies in the 2010s but these solutions are expensive, have limitations, and are not widely used. In the wider world, the Semantic Web has made a significant impact; and chemical structures can be searched by InChI in Google.

Garfield was the pioneer of citation searching, but CAS followed through later: CAplus now contains over 505 million cited references. Cited references are included for journals, conference proceedings, and basic patents from the USPTO, EPO, WIPO, and German patent offices added to the CAS databases from 1997 to the present. Also included are patent examiner citations from British and French basic patents (2003 to the present), Canadian patents (2005 to the present) and Japanese patents (2011 to the present). In addition, nearly 300,000 existing patent records from 1982-2008 have been supplemented with information for cited patents. Moreover SciFinder^{n} now has a nice citation mapping feature that Warr illustrated (acknowledging help from CAS) using a paper written by Garfield himself[11] as the focus of the map.

So what of the future of chemical databases and substructure searching? Edgar Fiedler said that he who lives by the crystal ball soon learns to eat ground glass. He also advised "*Give them a number or give them a date, but never both*." Nevertheless, Warr ventured into a few observations.

Substructure searching was a "solved problem" by the 1980s, and 3D and similarity searching followed. Yet, there are still frontiers to conquer. Very fast similarity searching is one. Dalke Scientific's chemfp project (http://chemfp.com/) began perhaps seven years ago. It is a set of command-line tools and a Python library for fingerprint generation and high-performance similarity search. MadFast (https://chemaxon.com/products/madfast) is a ChemAxon engine for fast similarity searching. It also provides fast calculation of descriptors. It uses efficient in-memory data storage and optimized multithreaded implementation. (Note that the use of memory is not novel: Daylight software took advantage of search in memory years ago.)

NextMove Software's SmallWorld (https://www.nextmovesoftware.com/smallworld.html) uses a graph database constructed by the decomposition of molecules one atom at a time. It uses Graph Edit Distance (GED) but GED is a generalization of calculating Maximum Common Subgraph (MCS), which is computationally hard. SmallWorld does efficient MCS searching of large chemical

databases. The sub-linear behavior of SmallWorld's nearest neighbor calculation makes it faster than fingerprint-based similarity methods.

"Big data" is a current buzzword, and huge databases of virtual chemical compounds are now appearing. Searching them fast will require new techniques. NextMove Software's Arthor (https://www.nextmovesoftware.com/arthor.html) technology builds on the company's Patsy chemical pattern matching engine. It outperforms current chemical cartridges, scaling to handle the hundreds of millions of compounds to be found in next generation chemical databases.

Of course, SciFinder is already capable of searching the 142 million or more compounds in CAS REGISTRY. For years parallelization of the atom-by-atom searches was the secret. Little is known about the technology behind the "exponentially better" features of SciFinder^{n}, and no mention has been made about any improvements in structure searching. Warr's quirky, personal suggestion for the future is to get rid of the tiresome superscripted "n".

Eugene Garfield: the father of chemical text mining and artificial intelligence (AI) in cheminformatics

Roger A. Sayle. NextMove Software, Cambridge, United Kingdom

The computational challenge of locating and resolving or classifying chemicals in text is commonly called "chemical named entity recognition". Today this field is an active area of artificial intelligence (AI) research, a branch of natural language processing, with international community-wide competitions (e.g., BioCreative) to assess progress, but this whole field started from Garfield's revolutionary Ph.D. thesis from the University of Pennsylvania's Department of Linguistics in 1961, reprinted as *Essays of an Information Scientist*, 1984, 7, 441-513 (http://www.garfield.library.upenn.edu/essays/v7p441y1984.pdf).

Sayle showed Garfield's list of morphemes for acyclic organic chemistry:

1. a	11. di	21. in	31. on**
2. acid	12. e*	22. iod	32. ox**
3. al	13. en	23. it	33. pent
4. am	14. eth	24. ium	34. sulf**
5. an	15. fluor	25. meth	35. tetr
6. at	16. hept	26. nitr	36. thi***
7. az	17. hex	27. o*	37. tri
8. brom	18. hydr	28. oct	38. y*
9. but	19. id	29. oic	39. yl
10. chor	20. im	30. ol	40. yn

The flow of Garfield's algorithm was as follows.

1. Ignore all locants.
2. Retain all parentheses.
3. Replace all morphemes by a dictionary value.
4. Resolve ambiguity of any penta-octa occurrences.
5. Place + between all morphemes except multipliers.
6. Carry out the multiplications and additions.
7. Calculate hydrogen using the formula H=2+2C+N-X-2DB.

The equation states that the number of hydrogens is (two, plus twice the number of carbons, plus the number of nitrogens, minus the number of halogens (X), minus twice the number of double bonds (DB).) Values for C, N, O, X and DB for certain "lexemes" of IUPAC names are given in the following table. For example, a triple bond, such as in a nitrile, is considered two double bonds, for the purposes of hydrogen counting using the equation above.

morpheme	C	N	O	DB
al			1	1
amino		1		
but	4			
en				1
eth	2			
hydroxy			1	
meth	1			
nitrile		1		2
nitro		1	2	1
ol			1	
one			1	1
prop	3			
yn				2

morpheme	mult
bis	2
di	2
hexa	6
penta	5
tetra	4
tri	3

Sayle showed some example calculations:

- methylaminoethane
  - → C+N+2C
  - → C_3H_9N
- (3-(diethylamino)propyl)ethyl-3-amino-1,4-butanedioic acid
  - → ((2(2C)+N)+3C)+2C+N+4C+2(2O+DB)
  - → $C_{13}H_{26}N_2O_4$
- bis(bis[diethylamino]propylamino)butane
  - → 2(2(2(2C)+N)+3C+N)+4C
  - → $C_{26}H_{50}N_6$
- hexanitrohexatriene
  - → 6(N+2O+DB)+6C+3DB
  - → $C_6H_2N_6O_{12}$

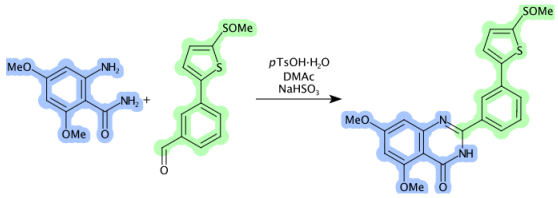
Modern "name-to-structure" software and chemical text mining tools not only owe their ancestry to algorithms first described over 50 years ago, but the ability of those original approaches to semantically resolve chemicals, and to handle ambiguous or generic structures, places them at what is considered the state-of-the-art even today.

Name-to-structure software includes Daniel Lowe's open parser for systematic IUPAC nomenclature (OPSIN)[12] used in NCI cactus services (https://cactus.nci.nih.gov/) and ChemDoodle (https://www.chemdoodle.com/); OpenEye Scientific Software's LexiChem (https://www.eyesopen.com/lexichem-tk) used by BIOVIA, (http://accelrys.com/products/collaborative-science/biovia-pipeline-pilot/component-collections/index.html); ChemAxon's Name to Structure (https://chemaxon.com/products/chemical-name-conversion); ACD/Labs' ACD Name (https://www.acdlabs.com/products/draw_nom/nom/name/index.php); Perkin-Elmer's ChemDraw Name>Struct (http://www.cambridgesoft.com/support/DesktopSupport/Documentation/N2S/); InfoChem's ICN2S (http://www.infochem.de/mining/annotator.shtml); and NextMove Software's Sugar & Splice (https://www.nextmovesoftware.com/sugarnsplice.html).

NextMove Software has considerable experience in automated chemical text mining.[13,14] The company's LeadMine (https://www.nextmovesoftware.com/leadmine.html) product is a text mining tool for the identification and annotation of chemicals, protein targets, genes, diseases, species, named reactions, company names, cell lines, etc. in the text of documents. Whilst initially developed to identify molecules of interest to medicinal chemists in patent applications, its functionality has been extended to handle also arbitrary entity types specified by dictionaries, ontologies, regular expressions or formal grammars. Sayle presented an example of a document marked up after extraction of a reaction and the melting point of the product:

[0835] To a solution of 2-amino-4,6-dimethoxybenzamide (0.266 g, 1.36 mmol) and 3-(5-(methylsulfinyl)thiophen-2-yl)benzaldehyde (0.34 g, 1.36 mmol) in N,N-dimethylacetamide (17 mL) was added $NaHSO_3$ (0.36 g, 2.03 mmol) and p-toluenesulfonic acid monohydrate (0.052 g, 0.271 mmol) at rt. The reaction mixture was heated at 120° C. for 12.5 h. After that time the reaction was cooled to rt, concentrated under reduced pressure and diluted with water (20 mL). The precipitated solids were collected by filtration, washed with water and dried. The product was purified by flash column chromatography (silica gel, 95:5 chloroform/methanol) to give 5,7-dimethoxy-2-(3-(5-(methylsulfinyl)thiophen-2-yl)phenyl)quinazolin-4(3H)-one (0.060 g, 10%) as a light yellow solid: mp 289-290° C.; ^{1}H NMR (400 MHz, DMSO-d_6) δ 12.19 (br s, 1H), 8.48 (s, 1H), 8.18 (d, J=7.81 Hz, 1H), 7.90 (d, J=8.20 Hz, 1H), 7.72 (d, J=3.90 Hz, 1H), 7.55-7.64 (m, 2H), 6.77 (d, J=2.34 Hz, 1H), 6.54 (d, J=1.95 Hz, 1H), 3.88 (s, 3H), 3.84 (s, 3H), 2.96 (s, 3H); ESI MS m/z 427 $[M+H]^+$.

NMSR/2222951



Product	MF	MW	Amount	Mass/Volume	Yield
5,7-dimethoxy-2-(3-(5-(methylsulfinyl)thiophen-2-yl)phenyl)quinazolin-4(3...	$C_{21}H_{18}N_2O_4S_2$	426.51		0.060 g	10 %
Reactant	**MF**	**MW**	**Amount**	**Mass/Volume**	
2-amino-4,6-dimethoxybenzamide	$C_9H_{12}N_2O_3$	196.20	1.36 mmol	0.266 g	
3-(5-(methylsulfinyl)thiophen-2-yl)benzaldehyde	$C_{12}H_{10}O_2S_2$	250.34	1.36 mmol	0.34 g	
Agent	**MF**	**MW**	**Amount**	**Mass/Volume**	
NaHSO3	$HO_3S.Na$	104.06	2.03 mmol	0.36 g	
p-toluenesulfonic acid monohydrate	$C_7H_8O_3S.H_2O$	190.22	0.271 mmol	0.052 g	
N,N-dimethylacetamide (Solvent)	C_4H_9NO	87.12		17 mL	

Info	
Document	US20140140956A1 [p0835]
Date	22-May-2014
Cite	David Fairfax *et al*.. Biaryl Derivatives As Bromodomain Inhibitors. ***U.S. Patent Application*** (2014)
Assignee	Rvx Therapeutics
NameRxn	Quinazolinone synthesis (4.1.40)

NextMove produces two separate databases: one for the reactions themselves, and the second for the product melting points. The arrow in the diagram above shows the linking of the product to its melting point, but the same paragraph of text can also be used to export the reaction or ELN page below.

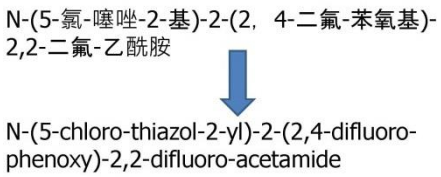
Software for handling English text often cannot handle the nonstandard use of whitespace, hyphenation, punctuation, Greek characters, italics and even superscripts found in chemical names. Likewise, the unusual letter combinations that occur in IUPAC, Chemical Abstracts, Beilstein and traditional names can trip up the trigram analysis frequently used in spell checking software. NextMove's CaffeineFix (https://www.nextmovesoftware.com/caffeinefix.html) overcomes the limitations of existing solutions by using novel algorithms for handling chemistry nomenclature. Sayle presented a spelling check example:

di-terf-butyl (4S)-/V-(fert-butoxycarbonyl)-4-{4-[3-(tosyloxy)propyl]benzyl}-L-glutamate

CaffeineFix corrected to:

di-tert-butyl (4S)-N-(tert-butoxycarbonyl)-4-{4-[3-(tosyloxy)propyl]benzyl}-L-glutamate

LeadMine can rapidly convert Korean, Chinese and Japanese chemical names to English as a preprocessing step, for example:



Furthermore, Sugar & Slice bridges the gulf between cheminformatics and bioinformatics by providing functionality for integrating the representations used in each domain. Sayle presented an example of generating molecular formulas for “biologics” from their common names and IUPAC names:

Oxytocin

L-cysteinyl-L-tyrosyl-L-isoleucyl-L-glutaminyl-L-asparagyl-L-cysteinyl-L-prolyl-L-leucyl-glycinamide (1->6)-disulfide

$C_{43}H_{66}N_{12}O_{12}S_2$

Mipomersen

O2'-(2-methoxyethyl)-P-thio-guanylyl-(3'->5')-O2'-(2-methoxyethyl)-5-methyl-P-thio-cytidylyl-(3'->5')-O2'-(2-methoxyethyl)-5-methyl-P-thio-cytidylyl-(3'->5')-O2'-(2-methoxyethyl)-5-methyl-P-thio-uridylyl-(3'->5')-O2'-(2-methoxyethyl)-5-methyl-P-thio-cytidylyl-(3'->5')-2'-deoxy-P-thio-adenylyl-(3'->5')-2'-deoxy-P-thio-guanylyl-(3'->5')-P-thio-thymidylyl-(3'->5')-2'-deoxy-5-methyl-P-thio-cytidylyl-(3'->5')-P-thio-thymidylyl-(3'->5')-2'-deoxy-P-thio-guanylyl-(3'->5')-2'-deoxy-5-methyl-P-thio-cytidylyl-(3'->5')-P-thio-thymidylyl-(3'->5')-P-thio-thymidylyl-(3'->5')-2'-deoxy-5-methyl-P-thio-cytidylyl-(3'->5')-O2'-(2-methoxyethyl)-P-thio-guanylyl-(3'->5')-O2'-(2-methoxyethyl)-5-methyl-P-thio-cytidylyl-(3'->5')-O2'-(2-methoxyethyl)-P-thio-adenylyl-(3'->5')-O2'-(2-methoxyethyl)-5-methyl-P-thio-cytidylyl-(3'->5')-O2'-(2-methoxyethyl)-5-methyl-cytidine

$C_{230}H_{324}N_{67}O_{122}P_{19}S_{19}$

Sometimes everything old is new again. Recently, attention has returned to molecular formulas and the challenges of turning them into chemical structures (connection tables). “Traditional” molecular formulas include line formulas such CH3CH2CH2Cl (a complete molecule), CH2CH2 (a linker), and CH3CH2 (a substituent); molecular formulas of inorganic salts, such as $MgSO_4$ and $AuCl_2$; and sum formulas, such as $C_{20}H_{25}NO_6$. There is, however a new twist. Consider:

peptide formulas:

- Cys(1)-Tyr-Phe-Gln-Asn-Cys(1)-Pro-Arg-Gly-NH2
- [N(Me)Leu15]orexin B (1-25)

oligosaccharides:

α-L-Fuc*p*-(1→4)-[β-D-Gal*p*-(1→3)]-β-D-Glc*p*NAc-(1→3)-β-D-Gal*p*-(1→4)-D-Glc-ol

oligonucleotides:

- 3'-AATG-5'
- sP-cl2Ade-Ribf.

The meaning of AI has changed. "I don't know what I do any more, but it used to be called artificial intelligence", quipped Sayle. Once upon a time, AI covered many broad disciplines, for example, problem solving and planning (A*-search, proof-number search, Monte Carlo search, genetic algorithms); natural language processing (NLP); propositional logic and reasoning (PROLOG); and optical character recognition (OCR). Definitions evolve over time: AI is now almost synonymous with (supervised) machine learning ("self-driving cars and finding kittens in Google").

Sayle listed some of Garfield's insights:

- chemical text mining to solve a real world problem pragmatically
- determining the probability that a word in text represents a chemical is far less useful than resolving it to a chemical formula for indexing
- a chemical formula is sufficiently useful for document indexing, and finesses issues of structural representation
- $C_{10}H_{10}Fe$ will retrieve the multiple possible representations of ferrocene.

Eugene Garfield's pragmatic approach to indexing chemical documents is in many ways superior to the recent vogue (fad?) of applying deep learning using long short-term memory, recurrent neural networks, or conditional random fields to chemical text mining. Sayle concluded with a couple of comments made by Garfield himself.

"*Certainly if we are to find methods of analyzing chemical texts for indexing and other purposes, we cannot expect better than a 50% resolution of the indexing problem in chemistry*"

"*We will have reaped a very poor harvest if we are able to describe the text of a chemical article grammatically without a corresponding ability to deal with the problem of synonymy*".

Eugene Garfield's legacy and its impact on the culture of research

Svetla Baykoucheva. STEM Libraries, University of Maryland, College Park, Maryland, United States

Gene Garfield's achievements have had a wide impact in many areas:

- the effective retrieval of scientific information
- tools for measuring academic impact
- scientific communication and publishing
- globalization of science, networking, and collaborations
- career advancement (hiring, promotion, awards, and monetary rewards)
- the new disciplines of scientometrics and bibliometrics
- science policy and research funding.

Baykoucheva listed some of the key products of Garfield's company, the Institute for Scientific Information (ISI).

- Science Citation Index (SCI)
- Social Sciences Citation Index (SSCI)
- *Current Contents* (CC)
- *Essays of an Information Scientist* (through ISI Press)
- Index Chemicus (IC)
- Current Chemical Reactions (CCR)
- Web of Science (WoS)
- Journal Citation Reports (JCR)
- Essential Science Indicators (ESI).

Current Contents has issues for multiple disciplines: life sciences; physical, chemical and earth sciences; engineering, technology, and applied sciences; clinical practice; agricultural, biological, and environmental sciences; arts and humanities; and social and behavioral sciences. ESI reveals emerging science trends, as well as influential individuals, institutions, papers, journals, and countries in different fields of research.

ISI's citation indexes, which are now included in WoS, were the most important sources of bibliometric information until Scopus was launched by Reed Elsevier in 2004. Created by Eugene Garfield in the early 1960s, SCI evolved to become the basis of innovative concepts and products such as WoS, JCR, and ESI. SCI is an effective information retrieval tool, which sparks new ideas through unexpected associations. It can help users to avoid duplication of research effort, and to assess the multidisciplinary influence of papers.

The factors that led to a need for SCI were the growth of science; an explosive growth in the volume of literature; the need for researchers to get recognition by their peers; and the search for objective, quantifiable tools for evaluation of journals, and of individuals' work. SCI has been used in studying how scientific information is communicated; in comparing the research output of different countries, institutions, and research groups; in underwriting new information products such as WoS, JCR, and ESI; and in historical and sociological studies.

The journal Impact Factor (IF) is a by-product from SCI, created to measure a journal's performance (for inclusion in the SCI). IFs are published in JCR every year. In any given year, the Impact Factor of a journal is the number of citations, received in that year (in indexed journals), of *all articles* published in that journal during the two preceding years, divided by the total number of *citable* articles published in that journal during the two preceding years.

The denominator excludes editorials, letters, notes etc.; they are not deemed to be citable. This is a controversial aspect of IF, because the numerator includes all citations, including those to reviews published as editorials, but it does not include those editorials in the denominator.

IF is frequently used as a proxy for the relative importance of a journal within its field; journals with higher impact factors are often deemed to be more important than those with lower ones. Authors look at the IFs when deciding where to publish their articles, because scholars are often evaluated, hired, promoted, and funded on the basis of whether they have published in high-impact journals. Editors sometimes try to understand how the IF is calculated so that they can manipulate the content of their journals to increase their ranking. Publishers and editors can determine a journal's influence in the marketplace and review editorial functions. Librarians often make decisions about which journal subscriptions to drop or add on the basis of the journals' IFs. Funding agencies are placing an increased weight on the IFs of the journals in which the applicants for grants have published their papers. Administrators monitor bibliometric and citation patterns to make strategic and funding decisions.

Unfortunately, citations can be misused, for example in self citations, and in "citation padding" (in which scholars cite each other). False reporting, plagiarism, and negligence in citing also occur, with some scholars getting a lot of information without doing much work. Exclusions such as editorials can have an effect on IF; reviews can be published as editorials. An editor can suggest that an author cites a particular journal if his or her paper is to be accepted for publication in that journal. Some unfortunate consequences from SCI and IF are the need to publish quickly in high-impact journals; increased competition for publishing in high-impact journals; obsession with citations; scientific misconduct and article retractions; and disadvantages for young scientists.

Garfield thought that it was inappropriate to use IF as a proxy to evaluate researchers for hiring, promotion, awards, and monetary rewards, or to use IF for policy decisions and research funding. He said: "*The Science Citation Index…was designed to be used for information retrieval. It is unfortunate that it's been so successful…in science policy uses*". Baykoucheva quoted two more of Garfield's opinions on the misuse of SCI and IF:

"*The impact factor is a very useful tool for evaluation of journals, but it must be used discretely. Considerations include the amount of review or other types of material published in a journal, variations between disciplines, and item-by-item impact*."

"*It is not appropriate to compare articles by IF, because IF applies to an entire journal...The article may never be cited, but if it is published in a high-impact journal, it indicates a high level of quality by being accepted in that journal…It is the citation count for an individual article and not the IF of the journal, which matters most. The journal IF is an average for all articles published in that journal*."

Baykoucheva turned to some of the other many ramifications of Eugene Garfield's ideas and legacy and how they have changed the culture of research. WoS offers ease of moving from item to item, and allows the user to go from "cycling" to "hypersearch". It facilitates historical research and prevents duplicate research. Garfield said once: "…chemical information was…my first love." His Ph.D. thesis (http://www.garfield.library.upenn.edu/essays/v7p441y1984.pdf) was about translating chemical names to molecular formulas. He went on to create IC and CCR.

He also had an impact on the issue of English as the *lingua franca* for science, because of journal selection for SCI. There is an English-language bias in SCI and SSCI, which works against material published in Chinese, Japanese, Russian, and other languages. Some foreign language journals are covered in SSCI, but coverage is not as comprehensive. In the social sciences in most countries much of the material is published in the local language. Nevertheless, anything that is published in foreign-language journals can be cited in the journals covered in SCI and SSCI. The English-language bias in SCI journal selection served to promote English to scientists in other countries. The top journals in the world are now published in English, and the official language at conferences is usually English. Garfield once offended some French scientists by his promotion of English, and a new journal in French was launched, but he actually had a great interest in languages.

Garfield had other impacts on the culture of research. His weekly essays, published for many years in *Current Contents*, touched on themes of enormous interest to a broad audience of scientists, academic administrators, and even politicians. He raised awareness of citations. He had an impact on the globalization of science, and the increase in interdisciplinary research. Research became more visible through ResearcherID, an identifying system for scientific authors which was introduced in 2008 by Thomson Reuters. SCI preceded the search engines, which used the principle of citation indexing to create algorithms for relevancy of documents. "Citation linking", a concept that is central to SCI, was on the minds of Sergey Brin and Larry Page when they published the paper in which Google was first mentioned (http://ilpubs.stanford.edu:8090/361/).

Baykoucheva quoted another of Garfield's insights:

"*If everything becomes open access…then we will have access to…all the literature. The literature then becomes a single database…All published articles will be available in full text. When you do a search, you will see not only the references for articles, which have cited a particular article, but you will also be able to access the paragraph of the article to see the context for the citation…*"

When Eugene Garfield created SCI, he could not have foreseen the dramatic impact his brilliant ideas would have on science and scientists in decades to come. Baykoucheva said that presenting at a symposium honoring Garfield's contributions was both an honor and a very emotional event, as Garfield had a tremendous impact on her own professional life and career. She has written articles about him and did two interviews with him. The first one was published in the *Chemical Information Bulletin* (http://hdl.handle.net/1903/11412) in 2006 and the second one (http://hdl.handle.net/1903/19169) was done in February 2015 and became a chapter[15] in her book "Managing scientific information and research data". Baykoucheva ended with a quotation from that interview:

"*I had always envisaged a time when scholars would become citation conscious, and to a large extent they have, for information retrieval, evaluation, and measuring impact…I did not imagine that*

the worldwide scholarly enterprise would grow to its present size, or that bibliometrics would become so widespread."

Beyond citations. What are new ways to assess content that will extend the assessment toolbox?

Todd A. Carpenter. National Information Standards Organization (NISO), Baltimore, Maryland, United States

In responding to a letter from a scientific rival Robert Hooke, Sir Isaac Newton wrote: **"***What Descartes did was a good step. You have added much several ways, and especially in taking the colors of thin plates into philosophical consideration. If I have seen further it is by standing on the shoulders of giants*." Citations linking the scientific literature allow us to "stand on the shoulders of giants". Fifty-eight years ago, on July 15, 1955, Eugene Garfield published his groundbreaking paper[1] on citation indexing in *Science* magazine. This innovative paper envisioned information tools that allow researchers to expedite their research process, evaluate the impact of their work, spot scientific trends, and trace the history of modern scientific thoughts. With that paper, essentially, Garfield launched the field of bibliometrics. Three years later, in July 1958, he laid the foundations for the Institute for Scientific Information (ISI) by borrowing $500 from Household Finance. He hired his first full-time employee and began to build an organization that included more than 500 people when it was acquired by the Thomson Corporation in 1992.

The Impact Factor (IF) and its compilation the Journal Citation Reports (JCR) have become, over the past five decades since it was launched, *the* metric for assessing journal quality. Journals live and die by this metric. In some developing countries, authors are awarded bonuses if they have published in a highly ranked title. Many in the STM community regularly tout their publication's performance. At the end of June when the JCR is released, it is often accompanied by a stream of press releases announcing this or that title's IF. For all its importance and value, the IF is an imperfect measure, and the community has been arguing about its imperfections for years. These include the time delay of citation data, the inability to compare different domains, the lack of granularity, and the measure's overuse and misapplication. If there were one metric to which the scholarly community was interested in finding an alternative, it probably is this one. That said, it is an ingenious and valuable metric that really has stood the test of time.

Beyond the Impact Factor, citation-based assessment metrics do play an important role in our community, though not in the way you might think, and certainly well beyond the domain of STM publishing. About 40 years after Garfield's seminal publication, two students, Sergey Brin and Larry Page, were working on the Stanford Digital Library Project (SDLP). The SDLP's goal was "to develop the enabling technologies for a single, integrated and universal digital library". It was funded through the National Science Foundation and other federal agencies. Brin and Page were focused on the problem of finding out which Web pages link to a given page, considering the number and nature of such backlinks to be valuable information about that page (with the role of citations in academic publishing in mind). This research project was nicknamed "BackRub". They wrote a paper (http://ilpubs.stanford.edu:8090/361/), and they started a company, Google, in someone's garage. Whenever you use Google, you are using a variant of a bibliometric citation analysis, that is, a combination of reference linking and usage data, to provide your search results. Basically, this is an "alternative metric" by a different name.

Isaac Newton's famous quotation came from an exchange of letters between Robert Hooke and Newton about a paper that Newton had written on the properties of light. Hooke had taken umbrage over the paper as it was something he had explored some 10 years earlier in his seminal work *Micrographia*. Hooke had also been involved in a long-running feud with Newton over which one had discovered the inverse square law. Newton was not a man to dole out praise, particularly to men whom he disdained or with whom he had scientific disagreements. It is not likely that Newton viewed Hooke as an intellectual giant since it was in part Hooke's criticisms that led to Newton's self-initiated withdrawal from the Royal Society in 1674. It is very likely, as some scholars have argued, that the phrase "on the shoulders of giants" was a veiled insult of Hooke, who was a quite short man physically, but also one in Newton's eyes who lacked intellectual stature, and was also the supposed butt of a successful theater farce at the time entitled *The Virtuoso*. Likewise, citations are not always what we think they are. They are amazing, but they are not the perfect solution to all the world's assessment problems.

Citation data reflect the very last stages associated with the publication process. Usage data can reflect earlier stages, and reflect a wide range of scholarly communication activities; they serve as an early, and potentially more comprehensive, indicator. More importantly, research outputs today involve much more than journals. How do we measure the impact of these different forms of output, or, indeed, the impact of all science? No researcher focuses on only one data source or methodological approach, so we too should search for an alternative.

Some alternatives are usage-based metrics, relationship-based metrics,[16] metrics based on social media, interaction-based metrics (e.g., Connotea, citeulike, BibSonomy, and Mendeley), adoption-based metrics (e.g., Open Syllabus Explorer), and sentiment analysis (e.g., Twinword and SenticNet). There *is* value in counting Tweets and Facebook likes. This is more than just a question of popularity: research[17-19] is pointing to the fact that there is a modest positive correlation between early-signal metrics (altmetrics) and later-signal metrics (citations).

Rather than simply replacing citation-based metrics with another form of assessment, it is more important that we should consider why and how we use assessment. Assessment is used for discovery, filtering, trend spotting, review, and decision making. What we are lacking in this new world of assessment is trust. The elements of metrics around which we need to build trust are definition, identification, granularity, time scale and exchange. We need standards to define what is to be counted, to describe what to count, and to identify what to count. We need standards for procedures for counting or not, for aggregating counts from the network, and for exchange of what was counted.

The National Information Standards Organization (NISO) is a nonprofit industry trade association, accredited by the American National Standards Institute (ANSI), with over 230 members. Its mission is developing and maintaining technical standards related to information, documentation, discovery, and distribution of published materials and media. It is a volunteer-driven organization with over 400 volunteers spread out across the world. It is responsible for standards such as ISBN, ISSN, DOI, Dublin Core metadata, DAISY digital talking books, OpenURL, and MARC records.

NISO has a altmetrics project (http://www.niso.org/standards-committees/altmetrics) and has published on recommended practice (http://groups.niso.org/publications/niso-rp-25-2016-outputs-niso-alternative-assessment-metrics-project). In 2016 NISO published "NISO RP-25-2016, Outputs of

the NISO Alternative Assessment Project" (https://www.niso.org/publications/rp-25-2016-altmetrics). This recommended practice on altmetrics was developed by working groups that were part of NISO's altmetrics initiative, a project funded by the Alfred P. Sloan Foundation. It covers definitions and use cases; code of conduct; output types for assessment; data metrics; and persistent identifiers and assessment.

Citations, usage data, and altmetrics are *all* potentially important and potentially imperfect. It is inadvisable to use altmetrics as an uncritical proxy for scholarly impact, because the attention paid to a research output, or the rate of the output's dissemination, may be unclear until combined with qualitative information. Additionally, it is important to recognize that data quality and indicator construction are key factors in the evaluation of specific altmetrics. Indicators that do not transparently conform to recommended standards are difficult to assess, and thus may be seen as less reliable for purposes of measuring influence or evaluation.

Altmetrics is a broad term that encapsulates the digital collection, creation, and use of multiple forms of assessment that are derived from activity and engagement among diverse stakeholders and scholarly outputs in the research ecosystem, including the public sphere. The inclusion in the definition of altmetrics of many different outputs and forms of engagement helps distinguish it from more established citation based metrics. At the same time, it leaves open the possibility of the complementary use of these conventional metrics, including for purposes of gauging scholarly impact. The development of altmetrics in the context of alternative assessment, however, sets its measurements apart from conventional instances of citation-based scholarly assessment.

Use cases for altmetrics are driven by the different stakeholders in the research ecosystem, many of whom interact directly with one another, and some of whom overlap on an individual basis. The deployment of personas helps to highlight the different ways in which these stakeholders collect, develop, and consume altmetrics, as well as the potential commonalities between altmetrics stakeholders' needs, goals, and usages. NISO developed eight personas, with three themes for each persona. "Showcase achievement" indicates the stakeholder's interest in highlighting the positive achievements garnered by one or more scholarly outputs. "Research evaluation" indicates the stakeholder's interest in assessing the impact or reach of research. "Discovery" indicates the stakeholder's interest in discovering or increasing the discoverability of scholarly outputs and researchers. An example for one persona, an academic or researcher is as follows:

Persona	Use case	Theme(s)
As a researcher, I want to...	Assess the reach, engagement with, and influence of my own research outputs, by, for example, incorporating altmetrics into my portfolio to complement my other accomplishments.	Showcase achievements Research evaluation
	Assess the reach, engagement with, and influence of the research outputs of my peers, by, for example, writing an external letter in support of the tenure of a researcher at another university.	Research evaluation
	Comply with reporting requests or mandates from funders, department heads, research administrators, etc.	Research evaluation Showcase achievements
	Choose to publish in a journal that will provide the maximum exposure of my work to relevant audiences.	Discovery

Other personas are research administrators, members of funding agencies, members of hiring committees, media officers or public information officers, publishing editors, content platform providers, and librarians.

The literature of altmetrics is rich with terminology that requires or implies more specific definitions. The NISO document has a glossary that represents a selection of these terms, based on the contents of the document, and the related outputs of phase II of the NISO altmetrics initiative.

The code of conduct covers transparency, replicability, and accuracy. Altmetric data providers are encouraged, and altmetric data aggregators are expected to adhere to the code. Transparency is the degree to which information and details about the provided data are clear, well-documented, and open to all users (human and machine) for verification. Information should be offered about how data are generated, collected, and curated; how data are aggregated, and derived data are generated; when and how often data are updated; how data can be accessed; and how data quality is monitored.

Replicability is the degree to which a set of data is consistent across providers and aggregators, and over time. The code ensures that:

- the data provided are generated using the same methods over time
- changes in methods and their effects are documented
- changes in the data following corrections of errors are documented
- data provided to different users at the same time are identical or, if not, differences in access provided to different user groups are documented
- information is provided on whether and how data can be independently verified.

Accuracy means that the data represent what they purport to reflect; known errors are identified and corrected; and any limitations of the provided data are communicated.

By following the code of conduct, altmetric data providers and aggregators agree to provide a publicly available annual report documenting in detail how they adhere to the recommendations above. The report should follow the standard format provided in a self-reporting table, which NISO

has created to support the code of conduct. NISO has also, in conjunction with the community, created sample reports for a selection of altmetric data providers and aggregators, to serve as a sort of guidepost.

NISO's alternative outputs table is a current list of nontraditional research outputs. These outputs may fall within the scope of assessment when developing metrics to evaluate the impact of scholarly activity, with the acknowledgment that meaningful impact can go far beyond traditional publishing workflows, and often involves the rich array of scholarly products that are created during the research process. These output types are alphabetized with a brief description, and documentation of known current efforts (and by whom they are being undertaken). Relevant links are listed where available, and most entries have been assigned a focus area to group them by similar contextual uses. Focus areas include: life and biologic sciences; capacity; code and software; communications; data; education and training materials; events; gray literature; images, diagrams, and video; industry; instruments, devices, and inventions; methodologies; publications; regulatory, compliance, and legislation; standards; and other.

NISO has made recommendations about data metrics. Metrics on research data should be made available as widely as possible. Data citations should be implemented following the Force11 Joint Declaration of Data Citation Principles. In particular, providers should use machine-actionable persistent identifiers; provide metadata required for a citation; provide a landing page; and ensure that data citations go into the reference list or similar metadata. Standards for statistics of use of research data need to be developed, based on COUNTER, and considering special aspects of research data. Data download metrics should examine both human and nonhuman downloads. Research funders should provide mechanisms to support data repositories in implementing standards for interoperability and obtaining metrics. Data discovery and sharing platforms should support and monitor "streaming" access to data *via* API queries.

A "Persistent Identifiers in Scholarly Communications" document is an environmental scan of common persistent identifiers that are used across a variety of scholarly domains to identify research outputs of any known type. Persistent identifiers may be applied to content at multiple levels of granularity, from links to a subset of a dataset to links to aggregated content. The purpose of the document is to raise awareness of the scope and complexity of persistent identifier use across systems, in the hopes of promoting and facilitating the use of persistent identifiers.

Carpenter thanked the dozens of people on the working groups and the hundreds of people who participated in brainstorming and commenting on this effort. He gave some statistics for the NISO alternative metrics initiative:

- 39 presentation slides have been downloaded 39,282 times (as of March 15, 2018)
- there have been more than 50 articles, blogs, or papers about the initiative
- the Phase 1 report published in 2014 was downloaded 12,734 times
- the final report has been downloaded 14,399 times
- pages hosting content related to this project were accessed 68,737 times
- more than 2,200 people attended the 24 in-person presentations about the project.

A project related to increasing trust and confidence in altmetrics is about one half completed. There are plans to promote, "operationalize", and iterate the initiative in future. NISO has yet to determine

the role of alternative assessment metrics in research evaluation, and to identify what gaps exist in data collection around evaluation scenarios; to identify best practices for grouping and aggregating multiple data sources; to identify best practices for grouping and aggregating by journal, author, institution and funder, and to define data usage metrics.

Carpenter briefly presented a number of other initiatives. MakeDataCount (https://makedatacount.org/) is a new project, funded by the Alfred P. Sloan Foundation, to develop and deploy the social and technical infrastructure necessary to elevate data to a first-class research output alongside more traditional products, such as publications. It will run between May 2017 and April 2019. The Research Data Alliance has set up a related data usage metrics working group (https://www.rd-alliance.org/groups/data-usage-metrics-wg). Rodrigo Costas, who has been doing some really exciting work to improve the way we identify scholars on Twitter, has found over 387,000 scholars with a Twitter account, and is confident of a 94% accuracy among those based on the ORCID validation. Elsevier now calls itself an analytics company and has acquired Plum Analytics.

Carpenter had a few final thoughts. Early-state indicators are just that, and may or may not correctly predict the future. As much as we may want it to be, "it's not all about the number". These are heuristics, not gospel: you have to read the paper. Reading the CliffsNotes will not lead you to understand and appreciate Shakespeare.

Novel research and its scientific and technological impact

Jian Wang. Leiden University, Leiden, Netherlands

Wang began with a quotation[20] from Eugene Garfield:

"Citations are the formal, explicit linkages between papers that have particular points in common. A citation index is built around these linkages. It lists publications that have been cited and identifies the sources of the citations. Anyone conducting a literature search can find from one to dozens of additional papers on a subject just by knowing one that has been cited. And every paper that is found provides a list of new citations with which to continue the search."

Following the Mertonian view,[21] peer recognition is the building block for the reward system of science and citations embody peer recognition for original and relevant contributions to science. On the other hand, citations are also tools of persuasion and subject to misbehavior.

On the subject of novel research, Roger Kornberg, who won the 2006 Nobel Prize in Chemistry, said (*Washington Post*, May 28, 2007): *"If the work that you propose to do isn't virtually certain of success, then it won't be funded".* How can you measure novelty? What is the relationship between novelty and impact? Are commonly used bibliometric indicators biased against novel research?[22]

Scientific discovery can be viewed as a form of problem solving, the process for which involves a combinatorial aspect, such as integrating different perspectives for defining the problem space and assembling various methods and tools for solving the problem within the problem space. In this respect, the creation of new scientific knowledge builds on combining existing pieces of knowledge. This is termed "combinatorial novelty".

To measure novelty, Wang used a publication's references (i.e., the citations in an article) as traces of its knowledge integration, and journals as bodies of knowledge. The novelty of a publication is the

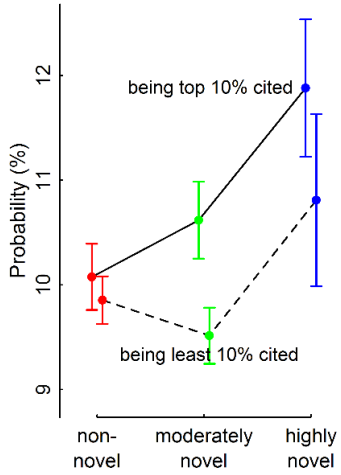
number of referenced journal pairs that are new, weighted by journal distance. Consider, for example, the dream of mapping the human brain. In 1986 Brenner's team[23] published a complete neuron wiring diagram of the worm, *C. elegans.* A long time elapsed before there was a breakthrough[24] involving an environmental scanning electron microscope plus ultra-microtome. Wang and his co-workers found that this article ranked in the top 1% for novelty in its field, using their measure.

They constructed their novelty indicator for all 661,643 research articles published in 2001 and indexed in the Web of Science Core Collection (WoS), based on their references. For each paper, they retrieved all of its referenced journals and paired them up (i.e., J_1-J_2, J_1-J_3, J_1-J_4…). They examined each journal pair to see whether it was new, that is, had never appeared in prior literature starting from 1980. For those new journal pairs (e.g., J_1-J_2), they assessed how easy it was to make this new combination, by investigating how many common "friends" the paired journals have. More precisely, they compared the co-citation profiles of the two journals (J_1 and J_2) in the preceding three years (i.e., 1998–2000).

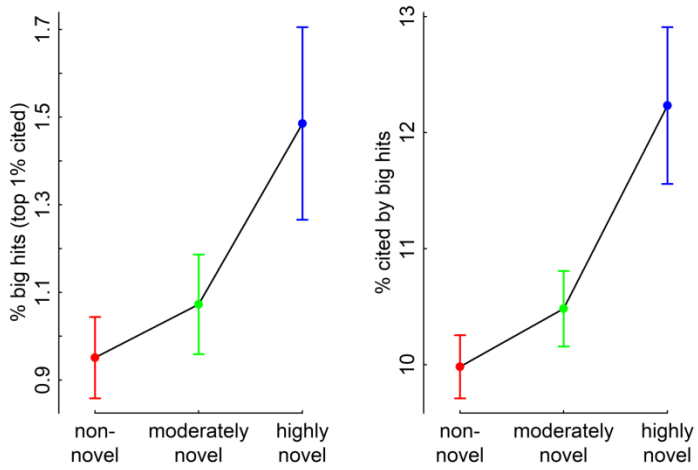
Novelty scores are highly skewed: 89% of articles do not have new combinations. Wang *et al.* used three categories for the novelty measure: non-novel articles have no new journal combinations; moderately novel ones are novel but not highly novel; and highly novel articles have a novelty score among the top 1%, in the same year and field.

Wang used a 15-year time window to count citations for the set of 2001 papers. They controlled for other confounding factors with potential influence on the relationship between novelty and impact. First, they controlled for specific scientific field effects, by including the complete set of dummies for the 251 WoS subject categories. Second, they controlled for the number of references made in the focal paper. Third, they took into account the size and nature of the collaborative effort, by including the number of authors and whether the paper is internationally coauthored as additional controls.

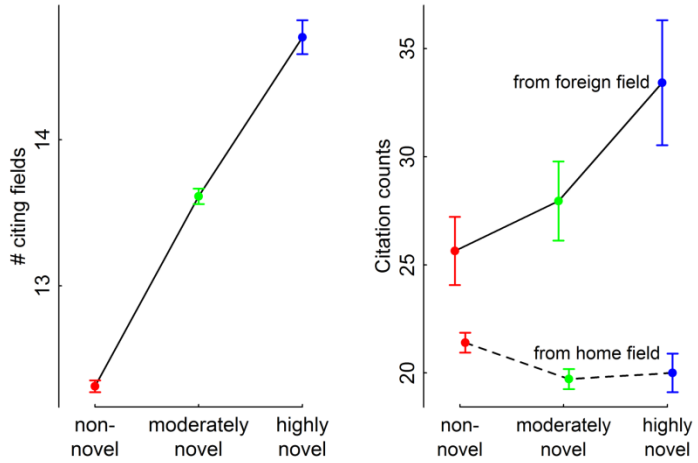
The high risk of novel research proved to be true:



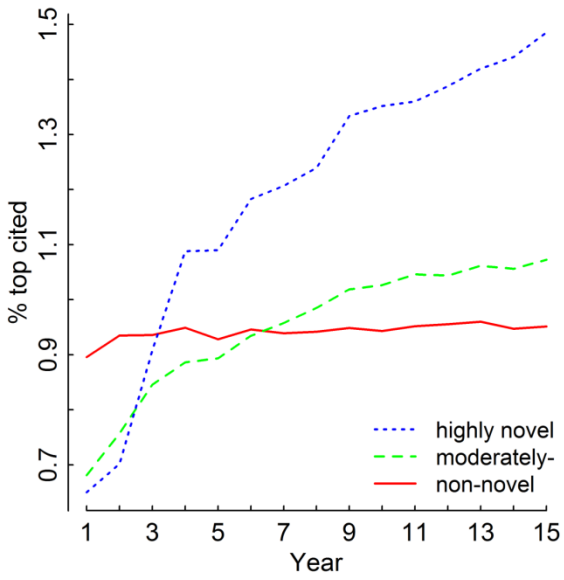
On the other hand, there is high gain from novel research:



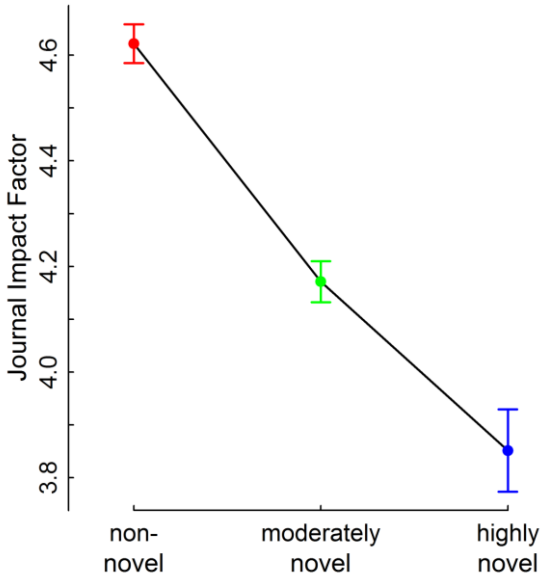
Novel research demonstrates a high risk/high gain profile: novel papers are more likely to be a top 1% highly cited paper in the long run, and to inspire follow-on highly cited research. They are also more likely to be cited in a broader set of disciplines, but at the same time display a higher variance in their citations. In addition, novel research is more highly cited in "foreign" fields than in the "home" field.



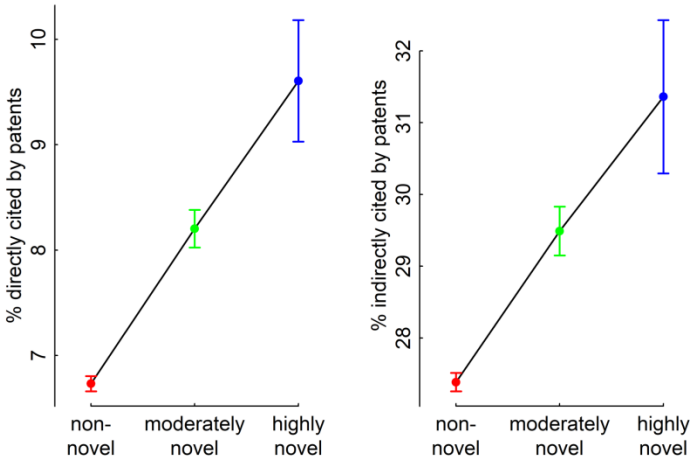
Wang and his colleagues also observed delayed recognition of novel papers: they are less likely to be top cited in the short term:



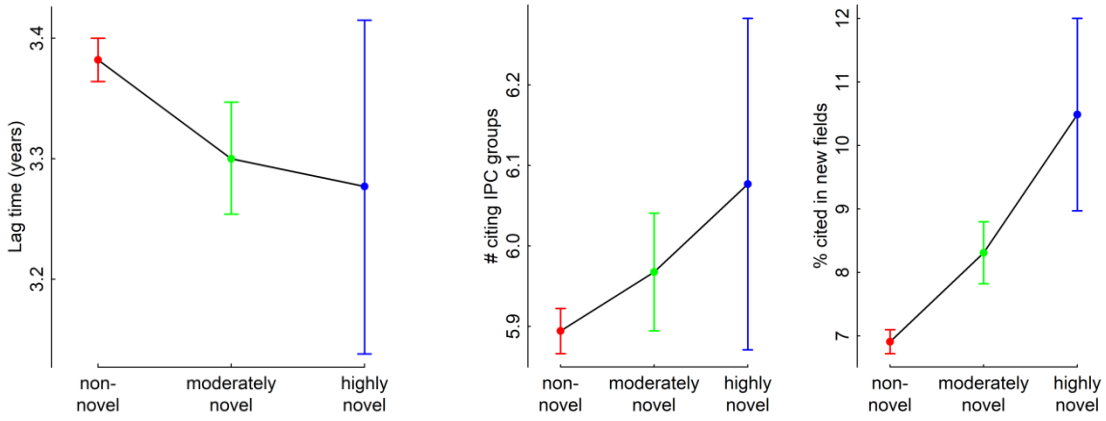
Finally, novel papers are published in journals with a lower Impact Factor, compared with non-novel papers, other things being equal:



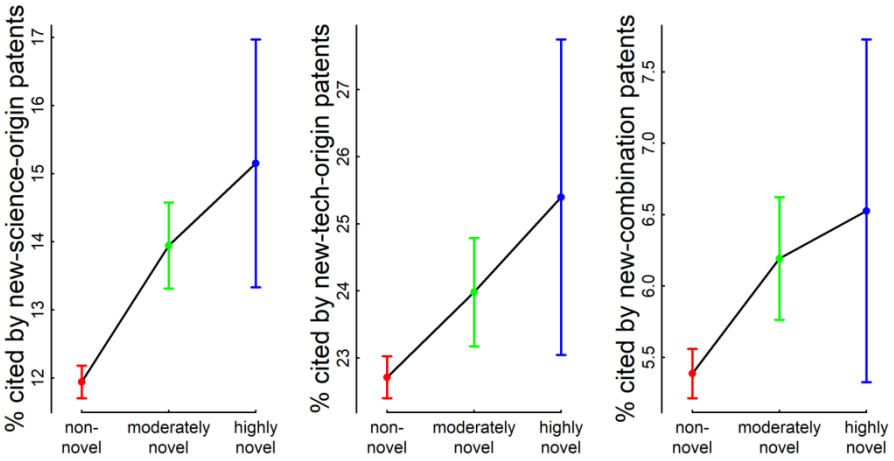
Universities and scientists face an increasing pressure to make a more direct contribution to the economy and society. What kind of science is more likely to be used for industrial innovation? Patent references to the scientific literature provide a paper trail of knowledge flow from science to innovation. Wang and his colleagues have thus investigated the relationship between novelty and technological impact. They find that novel publications are more likely to be directly cited by patents and also indirectly by other scientific publications which are cited by patents:



Within the set of scientific papers cited at least once by patents, there are no additional significant differences in the speed or the intensity of the technological impact between novel and non-novel scientific prior art. Wang plotted the time lag (the difference between the year of first patent citation and the year of paper's publication) and the technological fields impacted:



On the other hand, the technological impact from novel science is significantly broader, covering more diverse technological fields and reaching technology fields previously not impacted:



If we over-rely on Impact Factor and short-term citation, science policy will be biased against novel research. The monodisciplinary approach in peer review may fail to recognize the full value of novel

research. The pursuit of novelty (a self-organized science award system) does not conflict with the pursuit of economic and societal value.

Clarivate Analytics. Building on the Garfield legacy with Web of Science

James Testa. Clarivate Analytics, Philadelphia, Pennsylvania, United States

From the cosmic cataclysm referred to as the Big Bang, all things have come, and all things continue to move away from the original source: we exist in an ever expanding universe. The material universe and the infinite number of tangible and traceable relationships that exist among all of its elements cannot now be illustrated in any complete way. The universe of human thought and creative scholarly inquiry, on the other hand, can now be explored from any point in time forward, as well as from any point in time backward. This ability comes as a direct result of Eugene Garfield's pioneering work in citation indexing.

The universe of human thought and creative scholarly inquiry is the recorded scholarship of our age captured in the scientific literature. This scholarly universe has a core that consists of a small set of theories and analytical techniques which are considered a given. These theories and techniques are documented and developed further in the literature and, along with a relatively small number of extremely highly cited journals and articles, form the core literature. The universe also has a research frontier which is made up of all the current and prior work being done by all active researchers in a particular discipline, most of which is hardly noticed.

Each new paper has the potential to expand the core knowledge through evaluation by the community, a process that may result in citation. Like the sun in our solar system, the core literature is a tremendous source of energy and continually inspires new work on the research frontier. Works on the frontier may also gain critical mass by way of citation from the community, and as their mass increases they are drawn by intellectual gravity closer to the core. But unlike the sun which only *emits* energy, the scholarly core is continually cited, and thus energized, by work on the research frontier. As a result, the core literature continually grows in importance and influence. Only a relatively small number of items in the research frontier, however, will be referenced significantly. By analyzing citations from core literature covered in the Web of Science, Clarivate Analytics is able to detect and measure emerging science on the frontier and bring it into the collection.

Garfield conceived of and developed the first index of citations not only as a means of navigating the universe of scholarly research, but also to extract and make visible the order inherent in it and the interrelatedness of individual works of scholarship. It is possible, through citation analyses, to track and to visualize the relationships between the core literature and the literature of the research frontier.

Before Garfield founded the Institute for Scientific Information (ISI), the scientific literature was often organized by traditional subject indexing. By indexing citations, Garfield effectively replaced and amplified the indexers' expertise with the reference list created by the authors themselves. An alternative was "title word indexing"; Garfield wisely extracted all the value of title word indexing by creating the Permuterm Subject Index as part of the Science Citation Index (SCI).

In his seminal publication[1] Garfield proposed that citation indexing would help in the elimination of fraud and obsolete data. This approach was typical of Garfield: he saw a real problem and pursued a

practical solution to it through citation indexing. He defined the relationships among source articles and the papers they cite, or are cited by, as an "association of ideas". He created a "thought" index: he made visible each article's association with prior art and, in time, its influence on subsequent studies.

Garfield was inspired by *Shepard's Citations*, used by the legal profession, which used the concept of citation indexing, but applying it to the literature of science was a much more complex undertaking. All the bibliographic elements of each paper and each cited reference needed to be captured in a standardized manner, and then abbreviated meaningfully and efficiently. To imagine the monumental energy, courage and tenacity required to actually *begin* this work is overwhelming, but Garfield's vision was clear and he moved it forward.

He understood that a comprehensive view of world scholarship did not require indexing every single scholarly journal. On the other hand, he also understood that a citation index must include every issue of every journal it covers, every item published in each journal, and every cited reference in each of those sources. The citations that did not refer to sources covered by SCI had great value to the team charged with building coverage in SCI. These citations revealed the existence of emerging journals, journals that were being cited by the core literature, but that were not yet covered. Additionally, Garfield understood that the problem of coverage is also one of practical economics.

Because SCI is a multidisciplinary resource, it was necessary to identify, in a cost-effective manner, the core literature for each of the 175 subjects covered. Garfield's thought on coverage was profoundly influenced by the work of S.C. Bradford, famous for Bradford's Law of Scattering.[25] One formulation of Bradford's Law is that if journals in a field are sorted by number of articles into three groups, each with about one-third of all articles, then the number of journals in each group will be proportional to $1:n:n^2$. There are a number of related formulations of the principle. As a practical example, suppose that a researcher has five core scientific journals for his or her subject. Suppose that in a month there are 12 articles of interest in those journals. Suppose further that in order to find another dozen articles of interest, the researcher would have to go to an additional 10 journals. Then that researcher's Bradford multiplier (bm) is 2 (i.e. 10/5). For each new dozen articles, that researcher will need to look in bm times as many journals. Bradford's principle is often illustrated as a comet whose nucleus represents the core literature, and whose ever-widening tail represents additional journals that have some occasional relevance to the subject

Unfortunately, between 500 and 1,000 different journals were required to cover a given field fully in SCI, and with 175 subject categories in SCI, the total number of journals would have been unmanageable. It turns out, however, that there is a very significant degree of overlap among different fields. In 1972 Garfield did a study using SCI data that showed that 75% of all references captured, regardless of field, identified fewer than 1,000 journals, and that 84% of these references are to just 2,000 journals. The study also showed that only 500 SCI journals published 70% of the total articles in a given year. In addition, nearly half of the 3.85 million references published in SCI that year came from only 250 journals.

This phenomenon has become known as Garfield's Law of Concentration which states that the tail of the literature of one discipline consists, in large part, of the cores of the literature of other disciplines. As a result, the core literature of *all* scientific disciplines in 1972 involved a group of not

more than 1,000 journals and may even possibly be reduced to as few as 500 journals. So the 1972 coverage of some 3,000 journals in SCI far exceeded the core of all scientific literature of the time. In today's Web of Science Core Collection this principle continues to hold although, since the literature of science is growing continually, the overall numbers are larger.

By 1975 Garfield had extended the principles of citation indexing to the literature of the Social Sciences and the Arts and Humanities as well.

SCI consists of a number of interrelated indexes: the Citation Index, the Source Index, the Corporate Index and the Permuterm Subject Index. The Permuterm Subject Index uses words appearing in the titles of articles as indexing terms. All significant title words are permuted to create all possible pairs. Each pair then becomes a separate entry in the index. Title words are divided into three groups: primary terms, stop words, and semi-stop words. Next to these primary terms and their co-terms is the list of authors who used them in the titles of their articles. A reader can look up the author in the Source Index and find the article of interest. The Source Index is an alphabetic listing of all authors including all their papers published during the period covered by the index. In the Source Index we could find a current paper that cited a paper by an author we found in the Citation Index. The Citation Index is an alphabetical listing by first author of all cited papers, books, etc. that occur as references found in footnotes and bibliographies of the journals covered in the current SCI. These are the cited references found in the articles listed in the Source Index.

Garfield had solved the depth *versus* cost problem, by simply substituting the authors' cited references for the subject indexer's judgment. In the process, he created a contextual view of each paper and enabled the searcher, through citation analysis and ranking, to see clearly which papers were the most "important", at least as measured by volume of citations received. So we may start with a cited author in the Citation Index and navigate to the Source Index to find the citing author and article published in the years covered by that edition of the SCI.

The Corporate Index identifies all papers published at a specific institution in a specific geographic location. It consists of two complementary parts: the Geographic Index and the Organization Index. The Geographic Index is subdivided by country, city, institution, department, etc. The alphabetic Organization Index cross-references each institution with its geographic location.

Developing standardized abbreviations for every bibliographic element that might be found in any scholarly publication or article is the key to successful indexing. The point of standardization is to convey maximum meaning with minimum characters. The labor- and space-saving goals are in service to the greater purpose of achieving correct attribution of citation to the right journal, institution and person. Standardization of institutional and corporate names and addresses is an ongoing work demanding continuous review and revision.

A simple search might involve accessing one or more volumes of SCI. At the end of the process it would be possible to request a photocopy or a tear sheet of a source item by mailing in a form with information about the journal issue and the article title or author. It might take a week for delivery of the article. Conducting a search in the print SCI was a laborious and often time-consuming operation. Each SCI literature search usually involved consulting multiple volumes, often with the help of an expert librarian. The print version of SCI became gigantic and unwieldy; the 2014 print SCI, for example, filled 36 volumes, and around 125,000 pages of very small print.

With the advent of the Web of Science in 1996 the entire process of searching the scholarly literature became a lightning fast operation where everyone was an expert searcher from the beginning. All the richness of indexed and standardized metadata, including access to full text through a library's holdings, the publisher, or Google Scholar is now available instantly.

In Web of Science you can also view a citation map of the items that any article has cited, and the items that cited their work. Clicking on any one of these cited or citing items reveals all its bibliographic details. This gives a picture of the article's ancestors and its descendants, so to speak. You could also have gone straight to the lists of cited and citing articles. This contextualization of each source by way of the network of cited references in which it exists is the present realization of the vision of Eugene Garfield.

For over 50 years now the steady application of Garfield's journal selection process has allowed what is now Clarivate Analytics to build an index that provides direct access to the core literature. While the company has made some movement in recent years to include many works from the research frontier, the core will continue to grow fairly slowly. Efforts to illuminate the work that continues to emerge on the frontier will not stop. The entire scope of Web of Science, however, now has room for many more publications that may not yet be exerting influence on the surrounding literature, but that have specific importance to Web of Science users. Clarivate Analytics is listening very closely to Web of Science users worldwide to understand better the literature that is important to them.

The company can clearly see the emergence of important work on the frontier and through the Web of Science journal selection process may determine if this new work belongs in the Core Collection. Because the journal coverage in Web of Science is founded on Garfield's Law of Concentration, it is carefully managed by the Editorial Development Department at Clarivate Analytics whose members apply the principles set forth in the Garfield's journal selection process.

The future of the Web of Science is in partnership with the community of primary scholarly publishers, not only as the source of content, but also in collaborative efforts to improve continually the effectiveness and efficiency of scholarly publishing. Clarivate Analytics is now able to have a meaningful dialogue with publishers whose content has importance and influence to a specific community, importance and influence that are not necessarily measured in citation impact.

Along with all of the traditional indicators of good scholarly publishing, the company is also focusing attention on ethical publishing practices, and demonstration of impartial and thorough peer review for every journal selected, regardless of citation impact or any other factor. The future of Web of Science will, therefore, be established on and guided by the treasured legacy of Eugene Garfield, a legacy that is defined by integrity, collaboration, and innovation.

References

(1) Garfield, E. Citation indexes for science; a new dimension in documentation through association of ideas. *Science* **1955,** *122* (3159), 108-111.
(2) Garfield, E. Chemico-linguistics: computer translation of chemical nomenclature. *Nature* **1961,** *192*, 192.
(3) Warr, W. A. Representation of chemical structures. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2011,** *1* (4), 557-579.

(4) Garfield, E. How I learned to love the Brits. *J. Inf. Sci.* **2008,** *34* (4), 623-626.
(5) Dubois, J. E.; Viellard, H. DARC system. VII. Theory of generation-description. 1. General principles. *Bull. Soc. Chim. Fr.* **1968,** *3*, 900-904.
(6) Attias, R. DARC substructure search system: a new approach to chemical information. *J. Chem. Inf. Comput. Sci.* **1983,** *23* (3), 102-108.
(7) Moock, T. E.; Nourse, J. G.; Grier, D. L.; Hounshell, W. D. The implementation of atom-atom mapping and related features in the Reaction Access System (REACCS). In *Chemical Structures;* Warr, W. A., Ed.; Springer Verlag: Berlin, 1988; pp 303-313.
(8) Garfield, E. From laboratory to information explosions... the evolution of chemical information services at ISI. *J. Inf. Sci.* **2001,** *27* (2), 119-125.
(9) Ash, J.; Hyde, E. System for chemical retrieval. *Pure Appl. Chem.* **1977,** *49* (12), 1845-1853.
(10) Eakin, D. R.; Hyde, E.; Parker, G. Use of computers with chemical structural information. ICI [Imperial Chemical Industries Ltd.] CROSSBOW system. *Pestic. Sci.* **1974,** *5* (3), 319-326.
(11) Garfield, E. The history and meaning of the Journal Impact Factor. *JAMA, J. Am. Med. Assoc.* **2006,** *295* (1), 90-93.
(12) Lowe, D. M.; Corbett, P. T.; Murray-Rust, P.; Glen, R. C. Chemical name to structure: OPSIN, an open source solution. *J. Chem. Inf. Model.* **2011,** *51* (3), 739-753.
(13) Tetko, I. V.; Lowe, D. M.; Williams, A. J. The development of models to predict melting and pyrolysis point data associated with several hundred thousand compounds mined from PATENTS. *J. Cheminf.* **2016,** *8*, 2/1-2/18.
(14) Schneider, N.; Lowe, D. M.; Sayle, R. A.; Tarselli, M. A.; Landrum, G. A. Big Data from Pharmaceutical Patents: A Computational Analysis of Medicinal Chemists' Bread and Butter. *J. Med. Chem.* **2016,** *59* (9), 4385-4402.
(15) Baykoucheva, S. From the Science Citation Index to the Journal Impact Factor and Web of Science: Interview with Eugene Garfield. In *Managing Scientific Information and Research Data*; Chandos Publishing: Witney, Oxfordshire, 2015; pp 115-121.
(16) Bollen, J.; van de Sompel, H.; Hagberg, A.; Bettencourt, L.; Chute, R.; Rodriguez, M. A.; Balakireva, L. Clickstream Data Yields High-Resolution Maps of Science. *PLoS ONE* **2009,** *4* (3), e4803.
(17) Brody, T.; Harnad, S.; Carr, L. Earlier Web usage statistics as predictors of later citation impact. *J. Am. Soc. Inf. Sci. Technol.* **2006,** *57* (8), 1060-1072.
(18) Costas, R.; Zahedi, Z.; Wouters, P. Do “altmetrics” correlate with citations? Extensive comparison of altmetric indicators with citations from a multidisciplinary perspective. *J. Am. Soc. Inf. Sci. Technol.* **2014,** *66* (10), 2003-2019.
(19) Thelwall, M.; Haustein, S.; Lariviere, V.; Sugimoto, C. R. Do altmetrics work? Twitter and ten other social Web services. *PLoS One* **2013,** *8* (5), e64841.
(20) Garfield, E. *Citation Indexing - Its Theory and Application in Science, Technology, and Humanities*; Wiley: New York, 1979.
(21) Merton, R. K. The Matthew Effect in Science: The reward and communication systems of science are considered. *Science* **1968,** *159* (3810), 56-63.
(22) Wang, J.; Veugelers, R.; Stephan, P. Bias against novelty in science: A cautionary tale for users of bibliometric indicators. *Research Policy* **2017,** *46* (8), 1416-1436.
(23) White, J. G.; Southgate, E.; Thomson, J. N.; Brenner, S. The structure of the nervous system of the nematode Caenorhabditis elegans. *Philos Trans R Soc Lond B Biol Sci* **1986,** *314* (1165), 1-340.
(24) Denk, W.; Horstmann, H. Serial block-face scanning electron microscopy to reconstruct three-dimensional tissue nanostructure. *PLoS Biol.* **2004,** *2* (11), 1900-1909.
(25) Bradford, S. C. Sources of information on specific subjects. *Engineering* **1934,** *137*, 85-86.