150 YEARS OF NATURE: A DATA GRAPHIC CHARTS OUR EVOLUTION

Methodology for our analysis of the archive.

Nature has published more than 407,000 separate items in more than 300 different article types, from news reports, book reviews and obituaries to scientific articles and letters. Those items are grouped into eight different categories (see 'Content'). There were about 146,000 research contributions from scientists mostly in the categories 'Letters and reviews' and 'Articles'. Early in *Nature*'s history, letters were sometimes relatively short, and resemble the Correspondence articles published in the journal in recent decades.

We conducted deeper analyses on the 146,000 research contributions. The London-based technology firm Digital Science used its machine-learning software to analyse titles and abstracts, and assigned each article to one or more of 22 different fields of research and 157 subfields. Manual examination of these assignments showed that the software placed too many items in the research field 'History and archaeology'. We therefore manually analysed selected years to correct for that bias (see 'Scientific focus').

Digital Science analysed titles and abstracts to identify the most frequently used scientific terms. Words appearing in more than 85% of documents were not considered, to avoid common words such as 'the' and 'a'. Words that appeared in fewer than three documents were considered too rare and were omitted. The five words shown for each decade are the ones that appeared most often in each document's top ten scientific terms (see 'Key words').

'Author list' shows the median number of authors per item.

Software assigned genders to authors on the basis of their names. 'Gender split' charts only the authors whose presumed gender could be confidently assigned — around three-quarters of names in later years. For earlier years, too few names could be assigned.

The analysis in 'By country' assigned one unit to each country named in an affiliation on a research contribution. Many items in the first half of the time series did not have affiliations and could not be assigned.