

AD No. 31804
ASTIA FILE COPY

DOCUMENTATION
INCORPORATED

2521 CONNECTICUT AVENUE, N. W.
WASHINGTON 8, D. C.
COLUMBIA 5-4577

THE PREPARATION OF MANUAL DICTIONARIES OF ASSOCIATION

TECHNICAL REPORT NO. 5

PREPARED UNDER

CONTRACT NO. Nonr-1305(00)

for

THE OFFICE OF NAVAL RESEARCH

APRIL
1954

THIS REPORT HAS BEEN DELIMITED
AND CLEARED FOR PUBLIC RELEASE
UNDER DOD DIRECTIVE 5200,20 AND
NO RESTRICTIONS ARE IMPOSED UPON
ITS USE AND DISCLOSURE.

DISTRIBUTION STATEMENT A

APPROVED FOR PUBLIC RELEASE;
DISTRIBUTION UNLIMITED.

TECHNICAL REPORT NO. 5

THE PREPARATION OF MANUAL DICTIONARIES OF ASSOCIATION

In the previous report, we explained the logical significance of the symbol "*" to be read "is associated with" and the symbol "," to be read as "and" between classes. In this paper, we will describe the method of preparing manual dictionaries of association for any system of information.

Let us consider a system of information, S , comprising N items of information (documents, reports, books, etc.) indexed by the terms $t_1, t_2, \dots, t_n$. The sets of terms used to analyze or index the items of information will be designated by $I_1, I_2, \dots, I_N$, the subscript of any I indicating the number of the document indexed by that set of terms. The set of terms associated with any given term t_r we will call P_r : $t_k \in P_r \equiv t_k * t_r$. Thus P_1, P_2 , etc. will constitute the "pages" of our dictionary. We shall write $t_k * P_r$ if t_k is associated with each term of P_r .

The three elements which determine the size of any system are:

The number of items, N

The number of terms, n

The average number of terms used to analyze

each item, $\frac{\sum_k (\text{number of terms in } I_k)}{N}$

" $\in$ " means "belongs to"

As in Technical Report No. 3, we shall assume that S contains 50,000 items analyzed by $I_1, I_2, \dots, I_{50,000}$; and that each I is a unique set of 10 terms chosen from $t_1, t_2, \dots, t_{5,000}$.

There are many ways to organize an S so that the actual associations of the system can be exhibited. But it will be recalled that the number of actual associations in a system of 50,000 items each analyzed by 10 terms will range from 25,000,000 to 50,000,000. This means that although we could arrange all the associations in a linear alphabetical card file, the file would contain upwards of 25,000,000 cards and would be too big to be useful. We could also set up a file in which each I was recorded just once. Such a file would contain only 50,000 cards. But we would not know how to arrange it, since each I has 1023 potential filing positions. As a random IBM file, the search for any t_r would involve the sorting of 50,000 items. Such a search would yield a group of I 's containing t_r .

Let us assume that such a search yielded 200 I 's. The number of terms in such a group would be 2000 but many of these might be duplicates. We would thus have to examine the 200 cards to list the different t 's and eliminate duplicates. If now we asked for all t_k 's for which $t_k * t_r * t_s$, we could resort our 200 I 's to determine which ones contain both t_r and t_s . This whole process is also too laborious to be a practical method.

The method we have chosen for the initial preparation of dictionaries which exhibit the association of any group is as follows:

Suppose any term t_r appears in sets $I_1, I_2, \dots, I_{200}$, we then construct P_r , which includes all distinct terms contained in any set which contains t_r , that is, all the terms of $I_1, I_2, \dots, I_{200}$. Using materials we had indexed for another project, we selected three terms from the vocabulary set, namely, Waves (t_a); Boundary (t_b); and Lamination, Laminar (t_c).

For each of these terms we constructed a "page" by listing in sequence all other terms associated with them. t_a appeared on 52 items and in association with 247 other terms. This indicates that many terms were duplicated in analyzing the 52 items. Otherwise the P_a associated with t_a would have contained 468 terms. t_b appeared on 43 items, associated with 204 different terms; and t_c appeared on 40 items associated with 160 terms. The three "pages" are presented as Exhibits 1, 2, and 3.

If we enumerate the members of P_a , we have $t_a * (t_1, t_2, t_3, \dots)$. Since $A * A$, $t_a \in P_a$. Further, if $t_a * t_b$, then both P_a and P_b will contain t_a and t_b . Thus, if "waves" is associated with "laminar" in S , we would discover on the page for "waves" and on the page for "laminar" the terms "waves" and "laminar".

Suppose we write out (Exhibit 4) all the terms common to P_a and P_b . The terms common to P_a and P_b constitute the logical product of the two sets and we can always write $(t_a, t_b) * (P_a \times P_b)$, but not $(t_a * t_b) * (P_a \times P_b)$. That is, the terms which are common to the two lists are associated with "waves" and with "laminar", but not necessarily with both at once (in the same document).

Suppose now we write out the words which are common to our three pages (Exhibit 5), that is $(t_a, t_b, t_c) * \overline{[(P_a \times P_b) \times P_c]}$. Our result will be

$$(t_a, t_b, t_c) * (t_a, t_b, t_c, \dots)$$

since each is associated with the other two. Here again, it is important to note that we still have not advanced beyond a chain of pairs of associations; $t_a * t_b$; $t_a * t_c$; $t_b * t_c$, etc. The fact that t_a , t_b and t_c belong to $P_a \times P_b \times P_c$ does not imply that $t_a * t_b * t_c$.

The same information is given by a simple punched card system. We can design a card for each term or "page" with 5000 dedicated positions for $t_1, t_2, \dots, t_{5,000}$, numbered from 1 to 5000. The seventh hole, for example, is punched on every card representing a P to which t_7 belongs, including, of course, P_7 . Thus any P can be represented by a card on which all the members of P are punched in fixed pre-assigned positions.

The set of associations

$$(t_a, t_b, t_c) * (t_a, t_b, t_c, \dots)$$

will be given immediately by the punched holes common to all three cards P_a , P_b , and P_c since each common hole represents a member of $P_a \times P_b \times P_c$. But suppose we wished to use this same simple method of superimposition to find $t_a * t_b * t_c$. We would then have to construct P_1 , P_2 , etc. by listing on any P_k all the 2-term combinations rather than the single terms associated with t_k . On any single "pages" the number of actual 2-termed combinations $(t_1 * t_2)$, $(t_2 * t_3)$, $(t_1 * t_3)$, etc. might not be appreciably larger than the number of single terms, but in order to use superimposition of punched cards to find any associated pair associated with a third term, $(t_a * t_b) * t_c$, we would require a dedicated position on each card, not for 5,000 terms but for each of the $\frac{5000 \times 4999}{2}$ possible or potential pairs. A card with that many dedicated positions is a practical absurdity.

We certainly will find it necessary to go beyond such chains and to discover whether in any specific case the associations $t_a * t_b$, $t_a * t_c$, $t_b * t_c$ are in S because t_a , t_b , t_c are contained in a single set I_1 , that is, $(t_a * t_b * t_c)$ or

whether the associations represent three separate sets

$$I_1 (t_a * t_b)$$

$$I_2 (t_a * t_c)$$

$$I_3 (t_b * t_c)$$

It appears, at this stage of our investigations that the indexing machine we have designed for other purposes can also be used to answer such questions quickly and automatically, and the next report will describe the indexing machine with particular reference to its use in problems involving the association of ideas.

BOUNDARY (t_a)

Aerodynamics	Ejectors, jet	Jets	Q/VF technique	Ternary
Air	Electromagnetic			Tests
Airfoils	Errors	Laminar	Radiation	Theory
Alloys	Estimation	Layers	Radioactive	Thermodynamics
Alpha	Etches	Lift	Recovery	Tracer
Aluminum	Exchange	Liquid	Reflection	Transfer
Analysis	Experiments, Nikuradse	LOW	Refractive	Transformations
Angle			Regions	Transition
Anisotropy	Factor	Measurements	Revolution	Troposphere
Arrangements, atomic	Feshbach-Weisskopf-	Mechanics		Tubes
"Atlas"	Peaslee theory	Medium	Schlieren	Tunnels
Atomic	Flat	Metals	Secondary	Turbomachines
	Flow	Method	Sections	Turbulence
Bends	Fluctuations	Molybdenum	Secular ice	Turning
Blades	Fluids	Momentum	Self-induced	
Bodies	Formulas	Motion	Separation	USSR
Boundaries	Friction	Movement	Ships	Velocity
			Shock	Viscosity
Calculation	Gas	Nickel	Silicon	Visual
Cascades	Gradient	Nikuradse experiments	Silver	Visualization
Cobalt	Grain	Nozzles	Skewed jets	Volume
Coefficient			Skin	
Compressible	Heat	Optics	Slamming	Wake
Conditions	Helium II	Optimum	Smoke	Water
Conduction	Hermes		Smooth	Waves
Configurations	High	Parabolic	Snell's law	Weak
Continuous	Hydraulics	Parachutes	Soil	Weisskopf theory
Convection		Peaslee theory	Solid	Wind
Copper	Ice	Penetration	Solutions	
Creep	Impulsive	Perturbation	Speeds	
Crystals	Incidence	Phases	Stability	X-ray
	Index	Photographs	Stall	
	Induced	Pipes	Stress: stressing	Zinc
	Infinite	Pitot	Subsonic	
	Incompressible	Plane	Suction	
	Inhomogeneous	Plates	Supersonic	
	Integral	Point	Surface	
	Interaction	Poiseuille flow	Systems	
	Ionization	Precipitation		
	Iron	Prediction	Tandem	
	Isothermal	Pressure	Tank	
	Isotropic	Prevention	Technique, Q/VF	
		Propagation	Temperature	

WAVES (tp)

Acoustics	Damped	Heat	Meteorological	Radiation	Table
Aerodynamics	Data	Helium	Method	Radio	Tail
Air	Debye	High	Micrococcus pyogenes	Radio	Tapered
Airborne	Delta	Hobbing, hot	(var. aureus)	Ratio	Technique, Q/VF
Aircraft	Design	"Horizon" (Ship)	Microwaves	Rayleigh	Temperature
Airfoils	Detonation	Horizontal	Millimeter	Reflection	Tests
Analysis	Dielectric	Humidity	Mixing, jet	Reflex	Thickness
Antennas	Diffusers, supersonic	Hydraulics	Model	Oscillator	Theory
Antimony	Digital	Hydrogen	Modes	Research	Timoshenko's
Apparatus	Discontinuities	Hyperfine structure	Moisture	Resonators	equations
"Atlas"	Dispersion	Incompressible flow	Molecular	Revolutions	Transfer
Attenuation	Disturbance	Inhomogeneous	Momentum	Revolution	Transmission
	Domes	Integral	Mounted	Schlieren	Travelling waves
Bacteria	Drag	Intensity	Nonlifting	Seismographs	Troposphere
Basin	Earthquakes	Interaction	Nozzles	Shell	Tubes
Beams	Ejectors, jet	Interdigital	Nuclear	Ships	Tunnels
Bending	Elastic; Elasticity	Interference	Observations	Shock	Turbulence
Blocking	Electromagnetic	Ionization	Oceans	Sine wave	Ultra-High Frequency
Bodies	Electrostatic	Isotropic	"Ordvac"	Single	Unsteady
Boundary	Elevator	Jets	Oscillators	Sirens	Upper
Cadmium	Equations	Kamchatka peninsula	Pacific Ocean	Skewed jets	Utilization
Calculations	Estimation	Laminar	Palisades seismograph	Slamming	Varying
Caniliver beam	Exchange	Layers	Parachutes	Slope	Vibrations
Cavities	Exciters	Lethality	Pattern	Slots	Visualization
Cerenkov radiation	Fabrication	Limits	Photography	Spectrometers	Wake
Charts	Field	Linear	Piezoelectric	Spherical	Water
Circular	Filled, Dielectric	Lines	Plane	Strength	Waveguides
Climatology	Fine	Load	Plates	Structure	Waves
Clouds	Flat	Long	Poiseuille flow	Stress; Stress-	Weak
Coating	Flow	Low	Polar	ing	Westerlies
Cockpit	Fluids	Magnetrons	Pond	Subsonic	Whistles
Combustion	Flush-mounted	Maneuvering	Pressure	Superrefraction	Wind
Computer; Computing	Fourier integral	Mantle	Propagation	Supersonic	Wing
Compressible	Frequency	Mathematics	Q/VF technique	Surface	Yacht
Conducting	Generators	Mechanics	Quadrupole lines	Suspension	Zinc
Convection	Glow	Medium		Synoptic charts	
"Crest" (ship)	Great Britain				
Crystals	Grinding				
Curves					

LAMINATION: LAMINAR (t_c)

Acetylene	Filters	Natural	Surface
Adhesives	Flames	Neoprene	Symmetric
Air	Flat	Nylon	Temperature
Analysis	Flow	Oxygen	Tests
Annealing	Flowmeters	Parallel	Thickness
Araldite	Fluids	Phenol	Transfer
Axial	Friction	Pitot	Transformations (Math)
		Plastic	Transition
Benzyl alcohol	Gases	Plasticizers	Treatment
Bonding	Glass	Plates	Tubes
Boundaries	Heat	Plexiglas	Tunnel
Breakdown, dielectric	High	Point	Turbulence
Burners	Hydrogen	Polyethylene	Urea
Burning	Hypersonic	Polyesters	Velocities
		Polytetrafluoroethylene	
Calibration	Improvement	Polyvinylbutyral	
Carbon	Insulating	Porous	Walls
Channels	Integral	Pressure	Waves
Cloth	Interaction	Rain	Weak
Coatings	Interlayers	Reflection	Wettability
Compressible	Jets	Regions	Wind
Constant	Layers	Resins	
Convection	Loads	Resistance	
Crazing		Rocket	
Curing			
Deterioration	Machine	Separation	
Development	Materials	Shock	
Dielectric	Mathematics	"Silastic 240"	
Displacement	Meter	Silicones	
Distillation	Methacrylate	Skin	
Dowahols	Method	Speeds	
	Mica	Stability	
Effects	Mixing	Streams	
Elastomer	Moisture	Strength	
Electrical	Molding	Stress, stressing	
Epon	Momentum	Subsonic	
Erosion	Monoxide	Supersonic	
Errors	Motor		
Estimation			
Ethylurea			

BOUNDARY (t_g) and WAVES (t_b)

Aerodynamics	Incompressible flow	Schlieren
Air	Inhomogeneous	Shock
Airfoils	Integral	Skewed jets
Analysis	Interaction	Soil
"Atlas"	Ionization	Stress: stressing
Bodies	Isotropic	Subsonic
Boundaries	Jets	Supersonic
Calculations	Lamination; laminar	Surface
Convection	Layers	Technique, Q/VF
Crystals	Low	Temperature
Data	Mechanics	Tests
Diffusers, supersonic	Medium	Theory
Drag	Method	Transfer
Ejectors, jet	Momentum	Troposphere
Electromagnetic	Motion	Tubes
Estimations	Parachutes	Tunnels
Exchange	Plates	Turbulence
Flat	Poiseuille flow	Visualization
Flow	Pressure	Wake
Fluids	Propagation	Water
Heat	Q/VF technique	Weak
High	Radiation	Wind
Hydraulics	Reflection	Zinc
	Revolution	

BOUNDARY (t_a) and WAVES (t_b) and LAMINATION; LAMINAR (t_c)

Air	Method
Analysis	Momentum
Boundaries	Plates
Compressible	Pressure
Convection	Reflection
Estimation	Shock
Flat	Stress; stressing
Flow	Subsonic
Fluids	Supersonic
Heat	Surface
High	Temperature
Integral	Tests
Interaction	Transfer
Jets	Tubes
Lamination; laminar	Tunnels
Layers	Turbulence
	Wave
	Wave
	Wind