

Fig. 39.5 Diagram of an eccentric tablet press

at a maximum, and so its use is limited to relatively small scale production or development work. A diagram of this type of press is given in Fig. 39.5.

The rotary tablet press

In this type, there are a number of dies and sets of punches. The former are set in a rotating disc or 'table', and the punches are set in tracks mounted above and below the table. The table and tracks rotate together, so that one die is always associated with one pair of punches. The vertical position of the lower punch in the die is governed by passage above cams, and the force is applied by the punches passing over and under pressure rolls. This is illustrated in Fig. 39.6.

Outputs of over 10,000 tablets per minute can be achieved by this type of press, the output being governed by the speed of rotation of the table and the number of sets of punches.

The cross-section of the die is usually circular, but is not necessarily so, and non-circular tablets

are becoming more common, primarily as an aid to product identification.

Whilst punch faces may be flat, thereby giving cylindrical tablets, this is unusual. Bevelled and convex tablets, made with concave punches, are more frequently encountered, and the latter are essential if the tablets are to be coated. In addition, punches may be embossed, so that an identification mark, product name or manufacturer's logo appears on the tablet.

FUNDAMENTALS OF POWDER COMPRESSION

Measurement of force in a tablet press

Though Brockedon's patent for the compression of medicinal substances was issued in 1843, it was not for over 100 years that significant research into the production of tablets from powders took place. This was initially carried out by T. Higuchi and his group at the University of Wisconsin, and their series of publications entitled 'The physics of tablet compression' (Higuchi *et al.*, 1954) were the foundation of much of the research on compaction which has been carried out since then.

The reason for the delay in commencing fundamental research on tableting was probably an inability to measure accurately the compressing force. This parameter is a major influence on many tablet properties, e.g. strength and disintegration time, and so without knowledge of the applied force, meaningful studies on these tablet properties become difficult if not impossible.

Tableting research was transformed by the introduction of the so-called 'instrumented tablet machine' by Higuchi in 1954, in which strain gauges are attached to various parts of the press, enabling force to be measured accurately. In its simplest form, the strain gauge is a network of wires through which an electric current is passed. The wires are bonded very securely to, for example, the upper punch of a press. If a force is applied to the punch it deforms, the magnitude of deformation (i.e. strain) being governed by the applied force and the value of Young's modulus for the punch. The wire of the strain gauge is also deformed, and hence its electrical resistance changes. This results in a small voltage change.

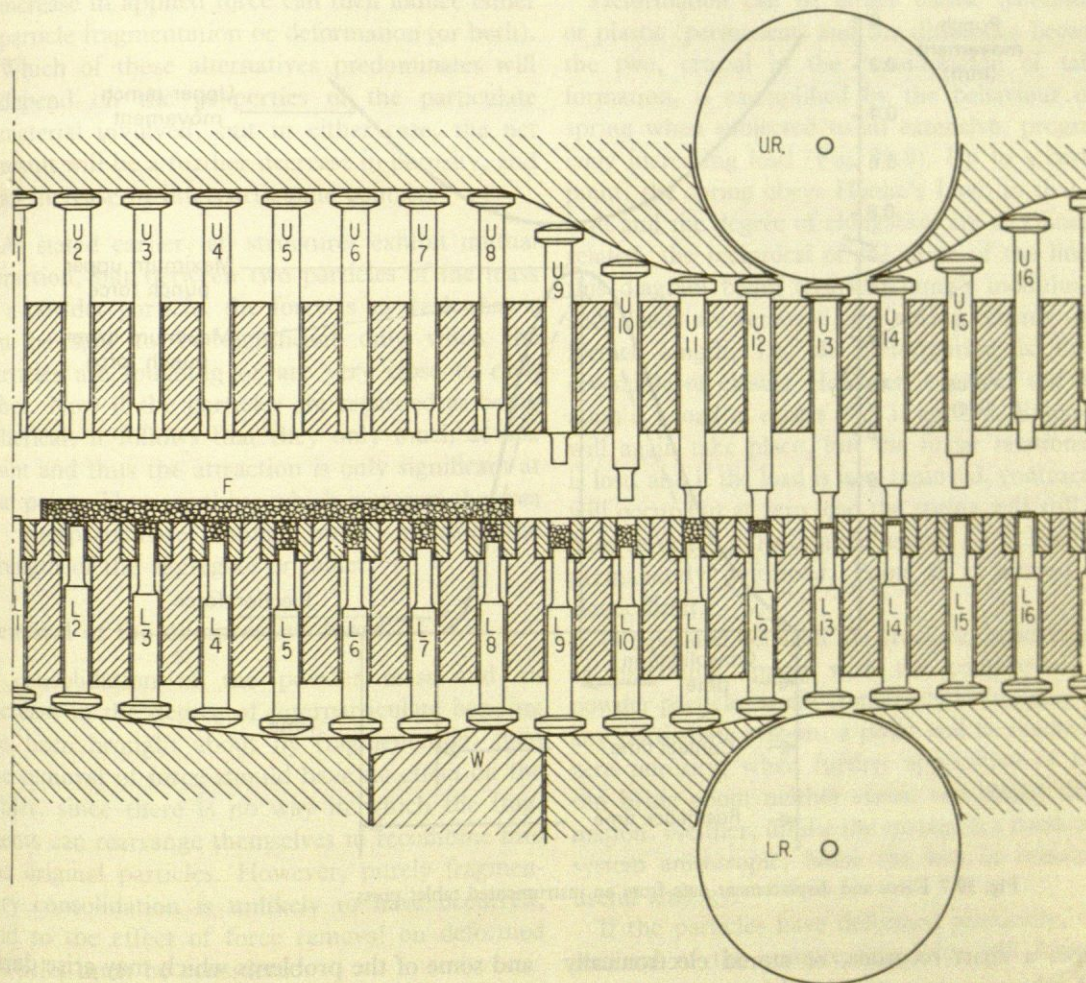


Fig. 39.6 Diagram of punch tracks of a rotary tablet press. U R, upper roller; L R, lower roller; W, capacity adjuster; F, feed frame with granules.

U1 to U8, upper punches in raised position; L1, lower punch at top position, tablet ejected; L2 to L7, lower punches dropping to lowest position and filling die with granules to an overfill at L7; L8, lower punch raised to expel excess granules giving correct capacity; U9 to U12, upper punches lowering to enter die at U12; L9 to L12, synchronized with U9 to U12 lower punches rising prior to compression; L13 and U13, upper and lower punches pass between rollers, and granules are compressed to a tablet; U14 to U16, lower punch rising to completely eject tablet at L16; U1 and L1, beginning of cycle

which can be amplified and recorded. The size of the signal from the strain gauge is proportional to the amount of deformation which in turn is a function of the applied force. Hence after appropriate calibration, the electrical signals can be expressed in terms of the applied force.

This concept is the foundation of much research on powder compaction. Rotary tablet presses have also been instrumented, and this, in addition to providing a research tool, permits automatic weight control in a production environment, since

an incorrect fill of granules into the die gives an unacceptably high or low force. That particular tablet can thus be rejected. A further development has been the use of piezoelectric crystals as an alternative to strain gauges. The former emit an electric charge when compressed, the magnitude of which is proportional to the compressing force. Also the attachment of displacement transducers, which measure distance, enables punch position to be accurately determined. The signals from the various transducers may be fed into an oscillo-

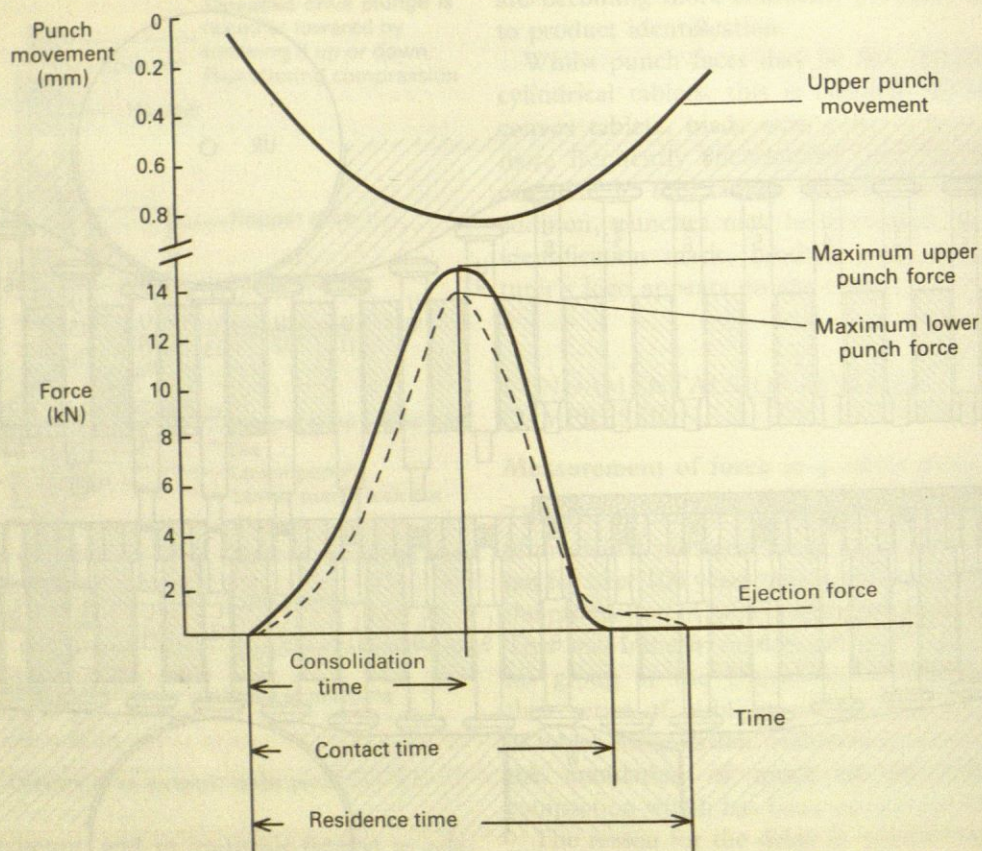


Fig. 39.7 Force and displacement data from an instrumented tablet press

scope, a chart recorder, or stored electronically and subsequently manipulated by computer.

A typical trace from an instrumented eccentric press is given in Fig. 39.7, which shows representations of upper and lower punch forces and the distance separating the punch faces. Considerable information can be derived from such a diagram, and it will be referred to several times in the remainder of this section.

The reader will already be aware that attractive forces exist between two particles. These forces may be non-specific, e.g. van der Waal's forces, or may be more specific in nature, e.g. brought about by molecules exhibiting intermolecular hydrogen bonding. However, irrespective of their nature, it is these forces which enable a coherent tablet to be formed, and an appreciation that their magnitude depends on the interparticulate distance is the key to understanding how tablets are formed,

and some of the problems which may arise during their manufacture.

Application of a force to particles in a die

Consider a number of particles present in a die and to which a force is applied. A series of events can then occur, perhaps sequentially, but there is a greater likelihood that some overlap will occur.

- 1 The particles will undergo rearrangement to form a less porous structure. This will take place at very low forces, the particles sliding past each other. This stage will usually be associated with some fragmentation, as the rough surfaces move relative to one another and rough points are abraded away.
- 2 The particles have now reached a state where further relative movement is impossible, though the porosity may still be considerable. A further

increase in applied force can then induce either particle fragmentation or deformation (or both). Which of these alternatives predominates will depend on the properties of the particulate material involved, but in either case, the net result will be a further decrease in porosity, and an increase in interparticulate contact.

As stated earlier, all structures exhibit mutual attraction, but between two particles of the mass of a powder particle, the force is so weak that it can be said to be significant only when the particles are touching or are very close to each other. Now if the particles are regarded as being spherical, it follows that they only touch at one point and thus the attraction is only significant at that point. Thus anything which increases the area of interparticulate contact must increase the strength of the aggregate or tablet.

Removal of the compressive force

If consolidation of the powder mass and an increase in the degree of interparticulate bonding has been brought about by fragmentation, then the removal of force should have no effect on the tablet, since there is no way in which the fragments can rearrange themselves to recombine into the original particles. However, purely fragmentary consolidation is unlikely to have occurred, and so the effect of force removal on deformed particles must be considered.

Deformation can be either elastic (reversible) or plastic (permanent) and the difference between the two, crucial in the consideration of tablet formation, is exemplified by the behaviour of a spring when subjected to an extensive, progressively increasing load (Fig. 39.8). Up to a certain point, the spring obeys Hooke's Law, in that the load and the degree of elongation are rectilinearly related, the reciprocal of the slope of the line in this diagram being termed Young's modulus. If such loads are removed, the spring returns to its former length, i.e. the deformation is totally reversible or elastic. However, consider the situation if a load in excess of E is applied. Extension will again take place, but the linear relationship is lost, and if the load is now removed, contraction will occur but at zero load the spring will still not have returned to its original length, i.e. it has been permanently deformed. Point E is termed the elastic limit.

The parallel between the Hookean behaviour of an extending spring with the compression of powder particles is not completely valid since with any particulate system, a point will be reached at zero porosity, when further application of force can bring about neither elastic nor plastic deformation. Neither, unlike the spring, is a particulate system anisotropic. None the less, it remains a useful analogy.

If the particles have deformed plastically, then removal of the compressing force will have no

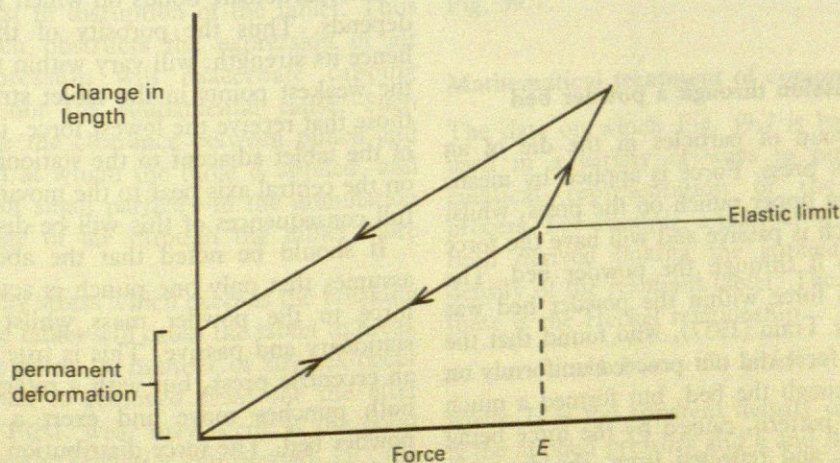


Fig. 39.8 Deformation of a spring by an applied force

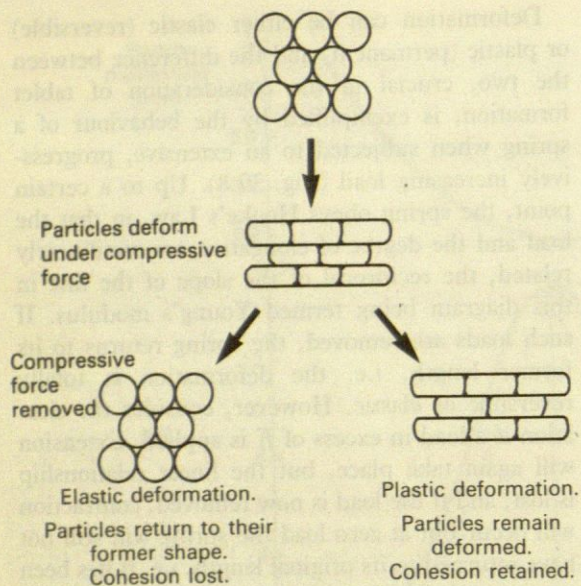


Fig. 39.9 Plasticity and elasticity in a particulate mass. (Reproduced from Armstrong, 1982 with permission of the copyright holder)

effect, since interparticulate bonds which are formed as a result of increased particle contact will not be disrupted. However, if particles tend to revert to their former shape, coherence will be lost as a consequence of the reduction in the area of interparticulate contact (Fig. 39.9). Most particulate systems show some elastic and some plastic properties but it is obviously desirable that plastic deformation should predominate. How the degree of elasticity or plasticity present in a system can be measured and altered will be discussed later in this chapter.

Force transmission through a powder bed

Consider a group of particles in the die of an eccentric tablet press. Force is applied by means of a descending upper punch on the press, whilst the lower punch is passive and will have the force transmitted to it through the powder bed. The distribution of force within the powder bed was investigated by Train (1957), who found that the diminution of force did not proceed uniformly on descending through the bed, but formed a much more complex pattern, caused by the force being transmitted to, and reflected from, the die wall. Significant features are zones of high force at the

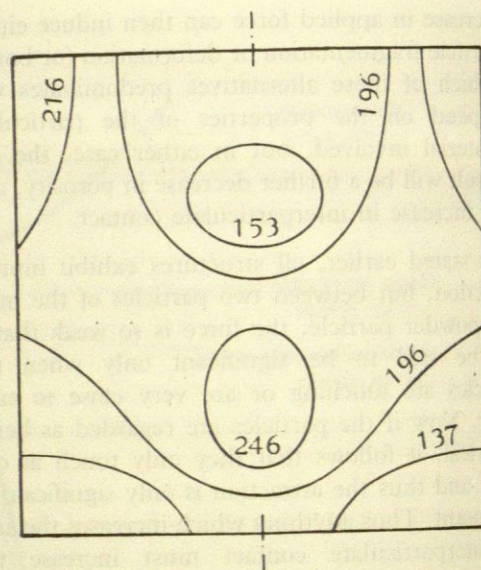


Fig. 39.10 Pressure distribution in a particulate mass. Contours are in N mm^{-2} (Reproduced from Train, 1957 with permission of the copyright holder)

periphery near the moving punch and much lower in the powder mass on its vertical axis. On the other hand, lower force zones occur on the same axis but much nearer the moving punch (Fig. 39.10).

The consequences of this on tablet strength can be profound. Particle deformation, whether elastic or plastic, will be proportional to the force applied, and as has been discussed, this deformation is an essential preliminary to the formation of interparticulate bonds on which tablet integrity depends. Thus the porosity of the tablet, and hence its strength, will vary within the tablet, and the weakest points in the tablet structure will be those that receive the lowest force, i.e. on the face of the tablet adjacent to the stationary punch and on the central axis near to the moving punch. The full consequences of this will be discussed later.

It should be noted that the above discussion assumes that only one punch is actively applying force to the powder mass whilst the other is stationary and passive. This is true in the case of an eccentric press, but with a rotary tablet press, both punches move and exert a force on the powder bed. The force distribution so obtained is thus different from that shown in Fig. 39.10, and results in two low density zones near the faces of

the tablet, and a high density zone in approximately the centre of the powder mass.

The foregoing can be summarized by stating that because of its non-uniform density, some areas of a tablet are stronger than others. Furthermore after the compressing force is removed, elastic recovery will occur to a greater or lesser extent, which will result in a reduction in the strength of interparticulate bonds and an overall weakening of the tablet. It therefore follows that if the tablet is to be disrupted by elastic recovery, this is most likely to occur at its weakest point. This is just below the top surface, and is a phenomenon often encountered in tablet manufacture known as capping or lamination. With this explanation in mind, some observed effects associated with capping, and some pragmatic causes and solutions to the problem, can now be explained.

Capping was for many years considered to be due to the entrapment of air in the tablet, and even the production *in vacuo* of tablets exhibiting capping did little to dispel this theory. Neither did this theory explain why air should cause tablet fracture at one particular zone. However, by considering the non-uniform density distribution in the tablet, it can be seen that the weakness is not caused by the presence of air *per se*, but rather the relative absence of solid material in those parts of the tablet which have high porosity. If this air is compressed, it follows that the pores are now filled with air at elevated pressure, which will obviously assist in disruption of the tablet. Thus anything which obstructs the expression of air during compression will exacerbate capping though it is not the fundamental cause. Such factors include the clearance between punch and die, the speed at which the force is applied, and the presence of small particles in the granulation making passage of air through the tablet more tortuous.

Similarly, any stress which exceeds the breaking strength of the tablet will cause the tablet to break at its weakest point. A number of stresses occur when the tablet is removed from the die after compression. First, wear may occur at the point in the die where the tablet is compressed, i.e., the die is fractionally wider at this point than elsewhere. Thus when the tablet is ejected, it is forced

through an aperture of diameter slightly less than the tablet itself. This will obviously stress the tablet, and the interparticulate bonds may be overcome at their weakest point. As the tablet is extruded from the die, elastic expansion will occur not just in an axial direction but also radially. The latter occurs progressively, i.e., one segment of tablet is free to expand whilst the one below it is still constrained by the die. Disruption of some interparticulate bonds is an inevitable consequence.

Assessment of lubricant action

The reduction in applied force with descent through the powder bed is primarily due to force losses at the die wall and by interparticulate friction. Both of these factors are reduced by the presence of a lubricant, and so comparison of the force applied by the upper punch with that received by the lower punch affords a measure of lubricant efficiency. This method, first suggested by Higuchi and co-workers (1954), defines the force ratio (R) as the ratio between lower punch maximum force and upper punch maximum force, and can be derived from the data represented in Fig. 39.7. The maximum value of R is unity, and lubricants based on stearates usually exhibit R values greater than 0.95.

An alternative method of measuring lubricant action is to measure the force required to eject the tablet from the die after the compressing force is removed. This too can be obtained from Fig. 39.7.

Mathematical treatment of compression data

The data on which Fig. 39.7 is based have been used in a variety of ways to provide a mathematical representation of the compression process. For example, a number of equations have been derived linking the applied force to the density of the resultant tablet. The best known of these is the Heckel relationship:

$$\ln 1/(1 - D) = kP + A$$

where D is the apparent density of the tablet, P is the applied pressure and k and A are constants. The apparent density is derived from a knowledge of the tablet dimensions, and the latter can be

obtained in turn from the displacement data in Fig. 39.7.

If the Heckel relationship is valid, then a graph of $\ln 1/(1 - D)$ vs P should be a straight line of slope k and intercept A on the ordinate. The equation has been used to distinguish substances which consolidate by fragmentation rather than deformation, and also as a means of assessing plasticity.

A totally different treatment is using the data from Fig. 39.7 to construct the so-called 'force-displacement curve'. In this, the force is plotted as ordinate and the corresponding punch position as the abscissa (Fig. 39.11). Calculation of the area enclosed by the curve has the units of force \times distance which dimensionally is equivalent to 'work'. Therefore the force-displacement curve has been used to calculate the work expended in the compression of a solid. Further refinement of this technique enables a measurement of elasticity and plasticity to be made.

This has shown that the presence of a granulating agent causes a marked increase in the plasticity of the particulate mass, with a consequent increase in cohesion and tablet strength. The film of granulating agent between the particles can be regarded as a highly viscous liquid with a large yield value. Application of a force in excess of the

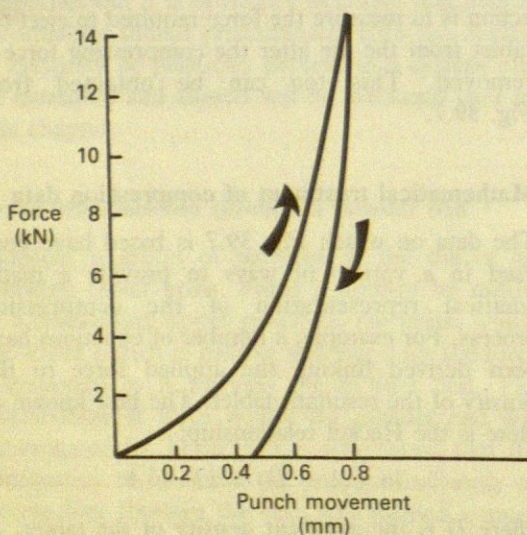


Fig. 39.11 Applied force as a function of punch movement in an eccentric press; a force-displacement curve

yield value causes granules to deform. Reduction of the force to below the yield value leads to permanent deformation. A somewhat similar mechanism is believed to account for the properties of some direct-compression diluents, e.g. spray-dried lactose, which consists of small crystalline masses embedded in an amorphous and more easily deformed matrix.

STANDARDS OF QUALITY FOR COMPRESSED TABLETS

Pharmacopoeial tests

Like all other dosage forms, the tablet is subjected to those pharmacopoeial standards which deal with 'added substances' with respect to their toxicity, interference with analytical methods, etc. However, there are a number of procedures which apply specifically to tablets, and which are designed, in the main, to ensure that the patient receives a product containing the required amount of drug substance in a form which enables the latter to exert its full pharmacological action.

Such standards in the *British Pharmacopoeia* are: uniformity of diameter, uniformity of weight, content of active ingredient, uniformity of content, disintegration, and dissolution. In addition there are a number of quality control procedures, which, though widely applied, are not defined by the Pharmacopoeia. These will be discussed later.

Uniformity of diameter

If tablets containing the same amount of drug substance but made by different manufacturers differ greatly in size, the consumer may well doubt whether tablets of such dissimilar appearance are of the same potency. The purpose of this standard is to help to remove this doubt. For details of the test, the stipulated diameters for specific tablets and permitted deviations, the reader is referred to the current edition of the *British Pharmacopoeia*. Though only the diameter is specified and not the tablet weight, it is reasonable to expect that tablets of the same diameter will not differ markedly in weight. This test was

introduced into the *British Pharmacopoeia* of 1958, replacing a guide of recommended tablet diameters which had been issued for some years by the Association of British Pharmaceutical Industry. It does not apply to tablets which are enteric, film or sugar coated. The usefulness of this standard is currently the subject of debate: the USP has never had such a standard.

Uniformity of weight

This test is carried out by removing a sample of 20 tablets from a batch, weighing them individually and calculating the mean weight, which in turn governs the permitted deviations from the mean. These are given in Table 39.2.

Not more than two tablets are permitted to differ from the mean by greater than the stated percentage and no tablet by more than double that percentage. Other national pharmacopoeias have similar standards, perhaps differing in minor detail. This standard applies to uncoated and compression-coated tablets.

Failure to comply with this standard, with consequent rejection of the batch, may be due to

uneven feeding of granules into the die, or irregular movement of the lower punch, producing a die space of varying capacity.

Where the drug substance forms the greater part of the tablet mass, dosage is obviously linked to tablet weight, and compliance with this standard helps to ensure that uniformity of dosage is achieved. However, in the case of highly potent substances, where the bulk of the tablet is diluent, this is not the case. This point is exemplified by the work of Airth *et al.* (1967), some of whose findings are illustrated in Fig. 39.12. When the tablet contained some 90% of active ingredient, then a perfectly linear relationship was found between tablet weight and drug content. Where the active ingredient comprised only 23% of the weight of the tablet, the relationship was much less significant.

Content of active ingredient

To carry out this test, 20 tablets are chosen at random from a batch, powdered together, and an assay carried out on an aliquot of the resultant mixture, according to the method given in the relevant pharmacopoeial monograph. The latter also states the range for the content of active ingredient which is permissible, and this range can be modified if fewer than 20 tablets are available.

There are two points to note about this standard. The first is that though the test is called 'content of active ingredient', there is no test for activity *per se*, the assay measuring, by chemical

Table 39.2 Uniformity of tablet weight

Average weight of tablet	Percentage deviation
80 mg or less	10
More than 80 mg and less than 250 mg	7.5
250 mg or more	5

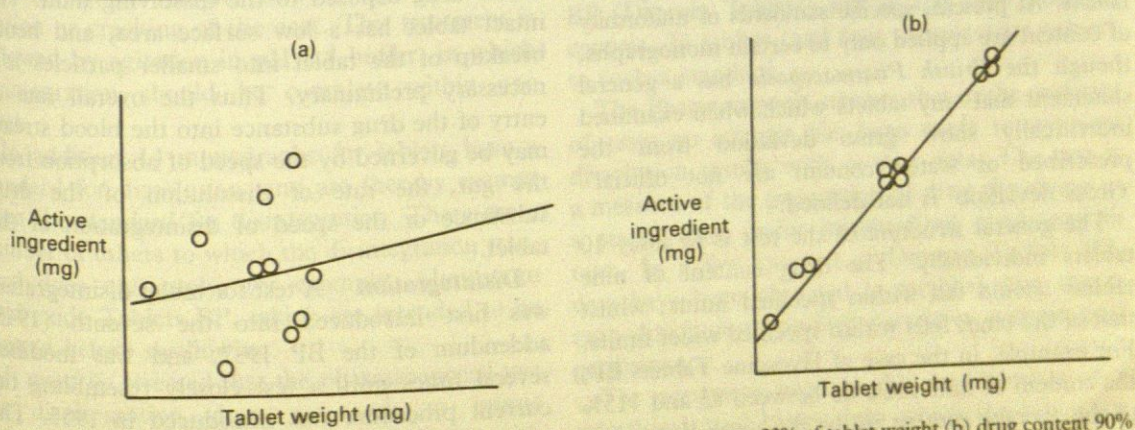


Fig. 39.12 The relationship between tablet weight and drug content: (a) drug content 23% of tablet weight (b) drug content 90% of tablet weight. (Reproduced from Airth, 1967 with permission of the copyright holder)

or physicochemical means, the amount of drug substance present. Possible therapeutic consequences of this are discussed elsewhere in this text (see particularly Chapters 8 and 9).

The second point is that the assay is carried out on a mixture obtained from 20 tablets, and thus the calculated content is the average content of those 20 tablets. Consider a hypothetical example, in which tablets were supposed to contain 100 mg of drug substance. Therefore 20 tablets would yield powder containing 2000 mg of drug. However, the same powder could be obtained from a sample of 10 tablets containing 150 mg plus 10 containing 50 mg or, in the extreme case, 10 tablets containing 200 mg plus 10 containing no drug whatsoever. Such high-dose tablets would almost certainly fail the uniformity of weight test described earlier. However, if the above calculation is now repeated, using a dose per tablet of 100 μ g rather than 100 mg, then the potential therapeutic hazard of such a test will be apparent. Such tablets will comprise mainly diluent, and it is unlikely that a 100% variation in the weight of drug substance per tablet will necessarily be reflected in tablet weight. Consideration of some of the points in Fig. 39.12 will illustrate this. It is for this reason that a standard for uniformity of content was introduced.

Uniformity of content

This standard is designed to guard against the variability in drug content within a sample of tablets. At present, specific standards of uniformity of content are applied only to certain monographs, though the *British Pharmacopoeia* has a general statement that 'any tablets which when examined individually show gross deviation from the prescribed or stated content are not official'. 'Gross deviation' is not defined.

The general structure of the test is to assay 10 tablets individually. The drug content of nine tablets should fall within specified limits, whilst that of the tenth falls within specified wider limits. For example, in the case of Hyoscine Tablets BP, the content of nine must lie between 85 and 115% of the average content whilst the tenth lies within 80 and 120% of the average. The average content is determined by the assay procedure on a bulked

sample of 20 tablets, and in the case of Hyoscine Tablets BP should lie between 90 and 110% of the stated value. It thus follows that an acceptable sample can contain one tablet with an actual content as low as 72% or as high as 132% of its stated content.

The 1980 *British Pharmacopoeia* contains 14 monographs for tablets in which uniformity of content is specified, without exception those containing potent substances of dose a few milligrams or less. Non-compliance with this standard and that of 'content of active ingredient' will be due to incorrect weighing of ingredients, failure to achieve satisfactory mixing at the blending stage, or subsequent segregation of the components of the tablet formulation.

The USP has a general rule that if the stated dose of a drug in a tablet is less than 50 mg, then a content uniformity test shall be applied. The structure of the test is similar to that in the BP, except that if one tablet lies outside the narrower limits, a further 20 tablets are assayed, all of which must lie within that limit.

Tablet disintegration and dissolution

Establishing the accuracy of the dose of a drug in a tablet is meaningless unless the drug can carry out its therapeutic function. In the majority of cases, this can only occur when the drug substance has dissolved in the fluids of the gastrointestinal tract. Dissolution rate depends on the surface area of the drug exposed to the dissolving fluid. The intact tablet has a low surface area, and hence breakup of the tablet into smaller particles is a necessary preliminary. Thus the overall rate of entry of the drug substance into the blood stream may be governed by the speed of absorption from the gut, the rate of dissolution of the drug substance or the speed of disintegration of the tablet.

Disintegration A test for tablet disintegration was first introduced into the seventh (1945) addendum of the BP 1932, and was modified several times until a test closely resembling the current procedure was introduced in 1955. The test provides a uniform means of agitating the tablet in an aqueous medium at body temperature, and a reasonably non-subjective end-point.

The disintegration chamber consists of a glass tube closed at the lower end by 2.00 mm aperture steel mesh. The tube is raised and lowered in a water bath at a constant frequency, so that at its highest point, the mesh remains below the surface of the water. For full details, the reader is referred to the current edition of the Pharmacopoeia. Commercial versions of this apparatus normally have six tubes.

The test is carried out by choosing a random sample of six tablets, placing one in each tube, adding a cylindrical plastic disc and agitating the tube in a water bath for 15 minutes. Disintegration is defined as 'that state in which no residue of the tablet, except fragments of undissolved coating, remains on the screen of the test apparatus or, if any other residue remains, it consists of a soft mass having no palpably firm, unmoistened, core', and to comply with the standard, all tablets must normally disintegrate within 15 minutes.

There are, however, a number of exceptions. Coated tablets are expected to disintegrate within 60 minutes, either in water or 0.1 M hydrochloric acid, soluble tablets and dispersible tablets within 3 minutes in water at 19–21 °C, whilst effervescent tablets should disintegrate within 3 minutes when placed in a beaker of water at room temperature.

The disintegration properties of enteric coated tablets are studied by agitating the tablet for 2 hours in 0.1 M hydrochloric acid, during which time the tablets should show no sign of disintegration or cracking of the coat. This treatment is followed by agitation in pH 6.8 buffer, in which disintegration should be complete within one hour.

In addition 14 monographs for tablets have a standard for dissolution, and are thereby exempt from the standard for disintegration. There are a number of others to which the disintegration standard does not apply, for example, Aluminium Hydroxide Tablets BP, which are intended to be chewed before swallowing.

It must be stressed that the pharmacopoeial test for disintegration does not seek to mimic conditions in the human gastrointestinal tract with respect to fluid composition or intensity of agitation. Thus compliance with the standard is

no guarantee of clinical efficacy. However, the converse is probably true in that a preparation which fails the pharmacopoeial test is unlikely to be fully efficacious. The disintegration time of a tablet is controlled by a number of experimental variables which are often interdependent. These include the type of granulating agent, the use of water-repellent lubricants, the type and amount of disintegrating agent and the force used to compress the tablet.

Dissolution Whilst the test for tablet disintegration gives some control over those drugs whose bioavailability from tablets is governed by the rate at which the tablet disintegrates, it gives no information regarding those cases where the tablet disintegrates satisfactorily, but the rate-limiting step is the rate at which the active drug substance dissolves in the fluids of the gastrointestinal tract.

The realization that drug dissolution could affect bioavailability occurred in the 1960s when several instances were reported in which tablets, whilst meeting all pharmacopoeial requirements, failed to produce the expected therapeutic response. The British Pharmacopoeia Commission selected from all the tablet monographs in the BP 1973 those substances which might pose dissolution problems, perhaps because of low solubility, or which had been the subject of allegations of inequivalence, or where if inequivalence arose, serious therapeutic consequences might ensue. The first monograph stipulating a dissolution standard appeared in the 1977 addendum to the 1973 BP (Digoxin Tablets) and in the 1980 Pharmacopoeia, 14 tablets (and four capsules) are subject to such a standard.

The Pharmacopoeia stresses that in the majority of cases no attempt has been made to correlate dissolution results with *in vivo* data. The test is a measure of the proportion of drug dissolving in a stated time under standardized conditions *in vitro*. In a few cases, e.g. Digoxin Tablets BP, data have been obtained to establish that the *in vitro* dissolution standard correlates with *in vivo* performance.

The apparatus used in this test consists of a cylindrical steel basket into which the tablet (or tablets) is placed, and the basket is then rotated in a bath of dissolution fluid, the constitution of

which is specified in the appropriate monograph. For details of the apparatus the reader is referred to the current edition of the BP. A sample is withdrawn from the dissolution fluid after a specified time (usually 45 minutes) and analysed by the method described in the monograph. Unless otherwise specified, not less than 70% of the stated content of the tablet should have dissolved. Five replicates are carried out, and if one of these fails to reach the standard, then five more tablets are tested, and now none should fail. Tablets which are subject to a dissolution standard need not be subjected to the test for disintegration.

The dissolution fluid is chosen bearing in mind the nature of the drug substance, and also the sensitivity of the assay procedure. Thus for example, 0.1 M hydrochloric acid is specified for bases such as chloroquine phosphate and quinine sulphate, a less acidic medium (pH 6.8 phosphate buffer) is specified for acidic substances such as phenoxymethylpenicillin, whilst water is suitable for neutral molecules such as digoxin. The latter also provides an example where the test is influenced by assay sensitivity, since 600 ml of water is used as a dissolution medium for six tablets. In this case, the dissolution data so obtained are obviously an average of the six tablets.

Whilst it is the intention of the *British Pharmacopoeia* to introduce a dissolution test only where dissolution problems are anticipated, the USP has adopted a totally different approach. Stating that the dissolution behaviour of oral solid dosage forms has been shown to be a useful criterion for controlling formulation and process variables, it proposes that except where such a standard would be inappropriate, all solid dosage forms shall be subject to a dissolution standard. The tablet disintegration test would thus ultimately be totally replaced.

Two methods are specified for carrying out the USP dissolution test. The first is a rotating basket method very similar to that of the BP. In the second, a paddle rotates in a bath of dissolution fluid. The composition of the latter, together with the method to be used, is specified in each monograph. Each monograph also specifies the amount (Q) which should dissolve in a stated time. Q is a percentage of the stated content of the

Table 39.3 Dissolution standards of the USP

Stage	Number of tablets to be tested	Criterion of acceptance
S ₁	6	Each unit not less than Q
S ₂	6	Mean of S ₁ and S ₂ not less than Q . None less than $Q - 15\%$
S ₃	12	Mean of S ₁ , S ₂ and S ₃ not less than Q . Not more than 2 less than $Q - 15\%$.

Q is the amount which should dissolve in the specified time.

tablet, and the USP permits three stages of sampling, using up to 24 tablets. Details of this scheme are given in Table 39.3. If the tablets meet the specified standard at stage 1, then it is not necessary to proceed to stages 2 and 3.

The general dissolution requirement of the USP is 75% dissolved in 45 minutes, which is very close to the normal BP standard. In an attempt to reduce interlaboratory variation in dissolution testing, the USP suggests the use of standard discs of salicylic acid and prednisone as calibration devices for dissolution apparatus.

Non-pharmacopoeial tests

The previous section dealt with those standards which are mandatory if a tablet formulation is to be the subject of an official monograph. However, it must be clearly understood that the manufacturer may well apply tests to his product which are not stipulated by the Pharmacopoeia or may apply the pharmacopoeial tests but with higher standards. Thus, for example, a manufacturer may have his own 'in-house' standard for content uniformity or dissolution, even though there may be no such specification in the pharmacopoeia. It would be unlikely, in fact, for a product licence to be granted for a tableted drug unless such data were presented. There are also a number of tests, frequently applied to tablets, for which there is no pharmacopoeial requirement, but which will often form part of a manufacturer's own product specification. The two most important tests in this category involve the measurement of the tablet's ability to retain its physical integrity.

Crushing strength

This is often referred to as tablet 'hardness', but the latter term is a misnomer in relation to the manner in which this test is carried out. 'Hardness' implies that the tablet possesses a surface which is resistant to penetration, and whilst penetration tests have been carried out on tablets, this test normally consists of breaking or crushing the tablet by application of a compressive load. A number of devices have been designed to measure crushing strength. The simplest are hand-operated, the tablet being held between a fixed and a movable jaw. Typical examples of this type are the Monsanto tablet hardness tester, in which the force is applied via a screw-driven spring, and the Pfizer tester, in which a gripping action transfers force to the tablet. In experienced hands, these can give fairly reproducible results, but it has been found that the crushing strength so obtained is in part governed by the rate at which the force is applied. Therefore in more sophisticated testers, the load is applied at a more uniform rate by mechanical or electromechanical means, and so greater reproducibility is obtained. Examples in this category include the Schleuniger, Erweka, CT40 and Casburt testers. Even if the load is applied at a uniform rate, the variation in strength within a batch of tablets can be considerable. Also since crushing strength is dependent on tablet dimensions, comparing the strengths of tablets of different sizes is difficult. In an attempt

to overcome these problems, the parameter 'tensile strength' is frequently used. This is given by the formula:

$$S_t = 2p/\pi \cdot d \cdot t \cdot (1 - e)$$

where S_t is the tensile strength, p is the crushing force, d is the tablet diameter, t is the tablet thickness, and e is the tablet porosity. This definition also compensates for the fact that porous tablets, having proportionally less interparticulate contact, will be expected to be weaker. A further development is to determine tablet 'toughness' in which the flexing of the tablet as a load is applied is also measured, analogous to force-displacement data obtained during tablet compression.

Resistance to abrasion

It is unlikely that in the normal life of a tablet it will be subjected to a compressive load large enough to fracture it. However, the tablet may well be subjected to a tumbling motion, e.g. during coating, packaging or transport, which whilst not severe enough to break the tablet, may abrade small particles from its surface. To examine this, tests to measure resistance to abrasion or 'tablet friability' have been devised. In all of these, the tablets are subjected to a uniform tumbling action for a specified time, and the weight loss from the tablets is measured. The Roche Friabilator is probably the most frequently encountered abrasion tester.

REFERENCES

- Airth, J. M., Bray, D. F. and Radecka, C. (1967) Variability of uniformity of weight test as an indicator of the amount of active ingredient in tablets. *J. pharm. Sci.*, **56**, 233-235.
- Armstrong, N. A. (1982) Causes of tablet compression problems. *Mfg Chem.* October, 64-65.
- Higuchi, T., Nelson, E. and Busse, L. W. (1954) The physics of tablet compression. III, The design and construction of an instrumented tablet press. *J. Am. Pharm. Ass. (sci. Edn)*, **43**, 344-348.
- Lowenthal, W., (1973) Mechanism of action of tablet disintegrants. *Pharm. Acta Helv.*, **48**, 589-609.
- Schwartz, J. B. (1979) Optimisation techniques in pharmaceuticals. In *Modern Pharmaceutics* (eds G. S. Banker and C. T. Rhodes), pp. 711-734, Marcel Dekker, New York.
- Train, D. (1957) Transmission of forces through a powder

mass during the process of pelleting. *Trans. Inst. Chem. Engrs.*, **35**, 258-266.

BIBLIOGRAPHY

- Armstrong, N. A. and Morton, F. S. S. (1979) An evaluation of the compression characteristics of some magnesium carbonate granulations. *Pharm. Weekblad*, **114**, 1450-1459.
- Augsberger, L. L. and Shangraw, R. F. (1966) Effect of glidants in tableting. *J. pharm. Sci.*, **55**, 418-423.
- Banker, G. S., Peck, G. E. and Baley, B. (1980) Tablet formulation and design. In *Pharmaceutical Dosage Forms: Tablets, Volume 1* (eds H. A. Lieberman and L. Lachman), pp. 61-107, Marcel Dekker, New York.
- de Blaeij, C. J. and Polderman, J. (1970) The quantitative

- interpretation of force-displacement curves. *Pharm. Weekblad*, **105**, 241-250.
- Chowhan, Z. T. and Chow, Y. P. (1980) Compression behaviour of pharmaceutical powders. *Int. J. Pharm.*, **5**, 139-148.
- Cooper, A. R. and Eaton, L. E. (1962) Compaction behaviour of several ceramic powders, *J. Am. Ceram. Soc.*, **45**, 97-101.
- David, S. T. and Augsberger, L. L. (1977) Plastic flow during compression of directly compressible fillers and its effect on tablet strength. *J. pharm. Sci.*, **66**, 155-159.
- Goodhart, F. W., Draper, J. R., Dancz, D., and Ninger, F. C. (1973) Evaluation of tablet breaking strength testers. *J. pharm. Sci.*, **62**, 297-304.
- Hess, H. J. E. (1978) Tablets under the microscope. *Pharm. Tech.*, **2**, 36-57.
- Hiestand, E. N., Wells, J. E., Peot, C. B. and Ochs, J. F. (1977) The physical processes of tableting. *J. pharm. Sci.*, **66**, 510-519.
- Higuchi, T. and others, The physics of tablet compression, a series of 18 articles which appeared in *J. pharm. Sci.* from 1953 onwards.
- Jarosz, P. J. and Parrott, E. L. (1982) Factors influencing axial and radial tensile strength of tablets. *J. pharm. Sci.*, **71**, 607-614.
- Jones, T. M. (1978) Formulation studies to predict the compaction properties of materials used in tablets and capsules. *Acta Pharm. Tech.*, **6**, 141-159.
- Lachman, L. and Sylwestrowicz, H. D. (1964) Experiences with unit-to-unit variations in tablets. *J. pharm. Sci.*, **53**, 1234-1242.
- Matsuda, Y., Minamida, Y. and Hayashi, S. (1976) Comparative evaluation of tablet lubricants. *J. pharm. Sci.*, **65**, 1155-1160.
- Pietsch, W. B. (1969) The strength of agglomerates bound by salt bridges. *Can. J. Chem. Engng*, **47**, 403-409.
- Rippie, E. G. and Danielson, D. W. (1981) Viscoelastic stress-strain behaviour of pharmaceutical tablets. *J. pharm. Sci.*, **70**, 476-482.
- Rudnic, E. M., Rhodes, C. T., Welch, S. and Bernardo, P. (1982) Evaluation of the mechanism of disintegrant action. *Drug Dev. Ind. Pharm.*, **8**, 87-109.
- Scott, M. W., Lieberman, H. A., Rankell, A. S. and Battista, J. V. (1964) Continuous production of tablet granulations on a fluidised bed. *J. pharm. Sci.*, **53**, 314-320 and 320-324.
- Sheth, B. B., Bandelin, F. J. and Shangraw, R. F. (1980) Compressed tablets. In *Pharmaceutical Dosage Forms: Tablets, Volume 1* (eds H. A. Lieberman and L. Lachman), pp. 109-185, Marcel Dekker, New York.
- Shotton, E. and Obiorah, B. A. (1975) Effect of physical properties on compression characteristics. *J. pharm. Sci.*, **64**, 1213-1216.
- Stenlake, J. B. (1981) The British Pharmacopoeia, 1980, scientific innovations. *Pharm. J.*, May 1981, 497-499.
- Train, D. and Lewis, C. J. (1962) Agglomeration of solids by compaction. *Trans. Inst. Chem. Engrs*, **40**, 235-240.

ATTACHMENT 3b



Keyword Search
[Back to results](#) » [Pharmaceutics](#) : » [Holdings](#)

Pharmaceutics : the science of dosage form design /

Other Contributors: [Aulton, Michael E.](#); [Cooper, John W. 1896-](#)
 Format(s): **Book**
 Language: **English**
 Published: **Edinburgh ; New York : Churchill Livingstone, 1988.**
 Edition: **1st ed.**

Tools




[Add to Book Bag](#)
[Permalink](#)
[Cite](#)
[Text](#)
[Email](#)
[Export](#)
[Report error](#)

[Holdings](#) | [Description](#) | [Staff View](#)

LEADER	01266nam a2200373 a 4500
001	492900
003	SIRSI
008	880203s1988 stka b 00100 eng d
010	a 86025888
020	a 0443036438 (pbk.) : c £19.95 (U.K.)
035	a (CSTRLIN)NJR88-B3337
035	a (OCoLC)ocm14272258
040	a DNLM/DLC c DLC d NJR
050	0 a R5420 b .P48 1988
090	a R5420 b .P48 1988 i 08/16/88 CTZ
245	0 0 a Pharmaceutics : b the science of dosage form design / c edited by Michael E. Aulton.
250	a 1st ed.
260	0 a Edinburgh ; a New York : b Churchill Livingstone, c 1988.
300	a xv, 734 p. : b ill. ; c 25 cm.
500	a Replaces: Cooper and Gunn's tutorial pharmacy, 6th ed. 1972.
504	a Includes bibliographies and index.
596	a 18
650	0 a Drugs x Design.
650	0 a Drugs x Dosage forms.
650	0 a Biopharmaceutics.
650	0 a Pharmaceutical technology.
650	0 a Pharmaceutical chemistry.
650	2 a Biopharmaceutics.
650	2 a Chemistry, Pharmaceutical.
650	2 a Dosage Forms.
650	2 a Technology, Pharmaceutical.
700	1 a Aulton, Michael E.
700	1 a Cooper, John W. q (John William), d 1896- t Tutorial pharmacy.
999	a R5420.P48 1988 STACKS t BOOK-Y

More on this subject

[Drugs > Design.](#)
[Drugs > Dosage forms.](#)
[Biopharmaceutics.](#)
[Pharmaceutical technology.](#)
[Pharmaceutical chemistry.](#)
[Biopharmaceutics.](#)
[Chemistry, Pharmaceutical.](#)
[Dosage Forms.](#)
[Technology, Pharmaceutical.](#)

Contact us Website Feedback Privacy Policy	NEWARK Camden Health Sciences	973-555-3901 856-225-6034 973-972-4580	  	  
www.libraries.rutgers.edu		Copyright © 2017 Rutgers, The State University of New Jersey		

ATTACHMENT 3c

Libraries that Own Item

- This screen shows libraries that own the item you selected.

[Home](#)
[Databases](#)
[Searching](#)
[Results](#)
[Staff View](#)
[My Account](#)
[Options](#)
[Comments](#)
[Exit](#)
[Hide tips](#)
[List of Records](#)
[Detailed Record](#)
[Marked Records](#)
[Saved Records](#)

 Go to page

 Current database: **WorldCat** Total Libraries: **88**

 Title: **Pharmaceutics : the science of dosage form design** Author: **Aulton, Michael E** Accession Number: **14272258**
Libraries with Item: "Pharmaceutics : / the sci..." ([Record for Item](#) | [Get This Item](#))

Location	Library	Local Holdings	Code
US,IL	ABBVIE		ITB
US,CA	ALIBRIS		ALBRS
US,CA	CALIFORNIA STATE UNIV, DOMINGUEZ HILLS		CDH
US,CT	BAYER CORP, PHARM DIV		XML
US,CT	UNIV OF CONNECTICUT		UCW
US,DC	COVINGTON & BURLING LLP		DCO
US,DC	LIBRARY OF CONGRESS		DLC
US,DC	WILLIAMS & CONNOLLY LIBR		DCY
US,FL	FLORIDA A&M UNIV		FCM
US,FL	UNIV OF FLORIDA, HEALTH CTR LIBR		FUH
US,IA	UNIV OF IOWA LIBR		NUI
US,KS	UNIV OF KANSAS		KKU
US,MA	PFIZER C/O SENTINEL		YLL
US,MD	NATIONAL LIBR OF MED		NLM
US,NC	BAKER & TAYLOR INC TECH SERV & PROD DEV		BTCTA
US,NH	YBP LIBRARY SERVICES		YDX
US,NJ	BRISTOL-MYERS SQUIBB PHARM RES INST		SQU
US,NJ	RUTGERS UNIV		NJR
US,NJ	STEVENS INST OF TECH	Local Holdings Availa...	NNO
US,NY	ADELPHI UNIV		VJA
US,NY	NEW YORK STATE LIBR		NYG
US,NY	ST JOHNS UNIV LIBR NETWORK		ZSJ
US,OH	OHIO STATE UNIV, THE		OSU
US,OH	PROCTER & GAMBLE, TECH LIBR		PGT
US,PA	UNIV OF THE SCIENCES IN PHILADELPHIA		PCP
US,RI	UNIV OF RHODE ISLAND		RIU
US,TX	UNIV OF HOUSTON		TXH
US,VA	VIRGINIA COMMONWEALTH UNIV		VRC
US,WI	UNIV OF WISCONSIN, MADISON, EBLING LIBR		GZH
Australia	CURTIN UNIV OF TECH		LC0
Australia	MONASH UNIV LIBR		LM1
Australia	ST MARYS COLL & NEWMAN COLL ACAD CENTRE	Local Holdings Availa...	ATSMN
Australia	UNIV OF ADELAIDE		LE1
Australia	UNIV OF S AUSTRALIA		LS0
Australia	UNIV OF SYDNEY		LS1
Australia	VICTORIA UNIV		LR0
Barbados	NLB MAIN LIBRARY		BXQHQ
Barbados	OISTINS BRANCH		BXQOI
CA,BC	UNIV OF BRITISH COLUMBIA LIBR		UBC
CA,ON	RYERSON UNIV		RRP
CA,ON	SENECA COL LIBR		CNSEN
CA,ON	UNIV OF TORONTO GERSTEIN SCI INFO CTR		CNGSI
CA,QC	MERCK FROSST CANADA RES LIBR		MFQ
China	SHANGHAI LIBR		SLY
France	BIBLIOTHEQUE UNIVERSITAIRE LYON-SANTE		FLB
Germany	UNIV ULM		DEULU
Hong Kong	UNIV OF HONG KONG		HUA
Ireland	TRINITY COLL DUBLIN		ERD

Jamaica	UNIV OF TECH OF JAMAICA		U@J
Malaysia	INTERNATIONAL ISLAMIC UNIV MALAYSIA IIUM		MYIIU
Malaysia	UNIV OF MALAYA LIBR		MYUML
Malaysia	UNIV OF MALAYSIA SARAWAK		MYMAS
Malaysia	UNIVERSITI SAINS MALAYSIA		MYUSM
Netherlands	UNIV OF GRONINGEN	Local Holdings Availa...	GRU
New Zealand	UNIV OF AUCKLAND LIBR		UV0
New Zealand	UNIV OF OTAGO LIBR		UZ0
South Africa	CAPE PENINSULA UNIV OF TECHNOL CAPE TOW		CQ\$
South Africa	DURBAN UNIV OF TECHNOL-STEVE BIKO CAM 5		T2N
South Africa	ETHEKWINI MUNICIPAL LIBR		Z5I
South Africa	NELSON MANDELA METROP UNIV 7500		OG\$
South Africa	NORTH W UNIV POTCHEFSTROOM CAM 3170		Y@Y
South Africa	RHODES UNIV LIBR 7090		OH#
South Africa	SEFAKO MAKGATHO HEALTH SCIS UNIV 2530		OE\$
South Africa	TSHWANE UNIV OF TECHNOL-ARCADIA CAM 2158		Y6I
South Africa	UNIV OF KWAZULU NATAL-HOWARD COLL 5450	Local Holdings Availa...	Z5F
South Africa	UNIV OF LIMPOPO TURFLOOP CAM 3500		Y#N
South Africa	UNIV OF PRETORIA 2840	Local Holdings Availa...	P4A
South Africa	UNIV OF THE WESTERN CAPE 6680		OD\$
South Africa	UNIV OF THE WITWATERSRAND HLTH SCI LIBR		X2#
Thailand	MAHIDOL UNIV LIBR & KNOWLEDGE CTR		MHDOL
United Arab Emirates	AJMAN UNIV OF SCI & TECH		UAAUS
United Kingdom	BANGOR UNIV		EQF
United Kingdom	BRITISH LIBR		BRI
United Kingdom	CARDIFF UNIV		RDF
United Kingdom	KINGS COL LONDON		KIJ
United Kingdom	NATIONAL LIBR OF SCOTLAND		NLE
United Kingdom	PUBLIC AUTH FOR APPLIED EDUC & TRAINING		KPE
United Kingdom	ROYAL ARMY MED COL		EUR
United Kingdom	SAINT BARTHOLOMEW HOSPITAL		EQM
United Kingdom	UNIV COL OF LONDON		LUN
United Kingdom	UNIV OF LONDON LIBR		ELU
United Kingdom	UNIV OF MANCHESTER LIBR THE		EUM
United Kingdom	UNIV OF OXFORD		EQO
United Kingdom	UNIV OF SHEFFIELD		SHS
United Kingdom	UNIV OF STRATHCLYDE		ESU
United Kingdom	WYETH RES, UK, LTD		WR4
United Kingdom	WYETH RES, UK, LTD, RCON		WY4
Zimbabwe	UNIV OF ZIMBABWE		W80

Record for Item: "Pharmaceutics : / the sci..."([Libraries with Item](#))

GET THIS ITEM	
Access:	http://www.gbv.de/dms/bowker/toc/9780443036439.pdf
Availability:	Check the catalogs in your library. <ul style="list-style-type: none"> Libraries worldwide that own item: 88 Search the catalog at the Library of University of Illinois at Urbana-Champaign
External Resources:	<ul style="list-style-type: none"> Discover UIUC Full Text Interlibrary Loan Request Cite This Item
FIND RELATED	
More Like This:	Search for versions with same title and author Advanced options...
Find Items About:	Pharmaceutics (677); Cooper, John W. (max: 4)
Title: Pharmaceutics : the science of dosage form design /	
Author(s):	Aulton, Michael E. Cooper, John W. ; 1896- ; (John William); Tutorial pharmacy.
Publication:	Edinburgh ; New York : Churchill Livingstone, Edition: 1st ed.
Year:	1988
Description:	xv, 734 pages ; 25 cm
Language:	English
Standard No:	ISBN: 0443036438; 9780443036439; National Library: 8610116; LCCN: 86-25888
Access:	Materials specified: Table of contents http://www.gbv.de/dms/bowker/toc/9780443036439.pdf
SUBJECT(S)	
Descriptor:	Drugs -- Design. Drugs -- Dosage forms. Biopharmaceutics.

[Biopharmaceutics](#)
[Pharmaceutical technology](#)
[Pharmaceutical chemistry](#)
[Biopharmaceutics](#)
[Chemistry, Pharmaceutical](#)
[Dosage Forms](#)
[Technology, Pharmaceutical](#)
[Médicaments -- Formes pharmaceutiques](#)
[Biopharmacie](#)
[Techniques pharmaceutiques](#)
[Chimie pharmaceutique](#)
[Médicaments -- Design](#)
[Biopharmaceutics](#)
[Drugs -- Design](#)
[Drugs -- Dosage forms](#)
[Pharmaceutical chemistry](#)
[Pharmaceutical technology](#)

Identifier: 21030; [pharmaceutics](#); Biopharmaceutics; Chemistry, Pharmaceutical; Dosage Forms; Drugs; Design; Drugs; Dosage forms; Pharmaceutical chemistry; Pharmaceutical technology; Technology, Pharmaceutical

Note(s): Replaces: Cooper and Gunn's tutorial pharmacy. 6th ed. 1972./ Includes bibliographical references and index.

General Info: **National bibliography no:** GB8701425

Class Descriptors: **LC:** [RS420](#); **Dewey:** [615.5/8](#); **NLM:** QV 785

Responsibility: edited by Michael E. [Aulton](#).

Vendor Info: Baker and Taylor YBP Library Services (BTCP YANK) £19.95 (U.K.)

Material Type: Internet resource (url)

Document Type: Book; Internet Resource

Date of Entry: 19860912

Update: 20160219

Accession No: **OCLC:** 14272258

Database: WorldCat



Current database: **WorldCat** Total Libraries: **88**



[English](#) | [Español](#) | [Français](#) | [عربي](#) | [日本語](#) | [한국어](#) | [中文\(繁體\)](#) | [中文\(简体\)](#) | [Options](#) | [Comments](#) | [Exit](#)

ATTACHMENT 3d

Statewide Illinois Library Catalog

UNIV OF ILLINOIS
[Ask A Librarian](#)

Libraries that Own Item

- This screen shows libraries that own the item you selected.

Home
Databases
Searching
Results

[Staff View](#) | [My Account](#) | [Options](#) | [Comments](#) | [Exit](#) | [Hide tips](#)

List of Records
Detailed Record
Marked Records
Saved Records

Go to page

E-mail
Print
Return
Help

Current database: WorldCat Total Libraries: 10

WorldCat

Title: **Pharmaceutics** Author: **Aulton, Michael E** Accession Number: 15591608

Libraries with Item: "Pharmaceutics" / ([Record for Item](#) | [Get This Item](#))

Location	Library	Code
India	TAMIL NADU VETERINARY & ANIMAL SCI LIBR	INTNV
Ireland	LIBRARY INST OF TECH, SLIGO	LQI
Saudi Arabia	UNIV OF DAMHAM	SAUOD
United Kingdom	BRIDGEND LIBR AND INFO SRV	UKBRD
United Kingdom	BRITISH LIBR	BRI
United Kingdom	BRITISH LIBR	UKM
United Kingdom	BRITISH LIBR REFERENCE COLLECTIONS	BLSTP
United Kingdom	ESSEX CNTY LIBR	ESSEX
United Kingdom	PORTSMOUTH CITY COUNCIL	UKPCC
United Kingdom	WOLVERHAMPTON CITY LIBR	WOLVL

Record for Item: "Pharmaceutics" / ([Libraries with Item](#))

E-mail
Print
Return
Help

Current database: WorldCat Total Libraries: 10

WorldCat

[GET THIS ITEM](#)

Availability: **Check the catalogs in your library.**

- [Libraries worldwide that own item:](#) 10
- [Search the catalog at the Library of University of Illinois at Urbana-Champaign](#)

External Resources:

- [Discover Full Text](#) Discover UIUC Full Text
- [Interlibrary Loan Request](#)
- [Cite This Item](#)

[FIND RELATED](#)

More Like This: [Search for versions with same title and author](#) | [Advanced options...](#)

Find Items About: [Pharmaceutics](#) (648)

Title: **Pharmaceutics /**

Author(s): [Aulton, Michael E](#)

Publication: Churchill Livingstone,

Year: 1988

Description: 734 pages ; 25 cm

Language: English

Standard No: ISBN: 0443036438; 9780443036439 LCCN: 86-25888

SUBJECT(S)

Descriptor: [Pharmaceutical chemistry](#)
[Pharmaceutical chemistry](#)

Identifier: **Pharmaceutics**

Class Descriptors: LC: [RS403](#); Dewey: [615/19](#)

Responsibility: edited by Michael E. [Aulton](#).

Document Type: Book

Entry: 19880217

Update: 20150328

Accession No: OCLC: 15591608

Database: WorldCat

E-mail
Print
Return
Help

Current database: WorldCat Total Libraries: 10

WorldCat

[English](#) | [Español](#) | [Français](#) | [عربي](#) | [日本語](#) | [한국어](#) | [中文\(繁體\)](#) | [中文\(简体\)](#) | [Options](#) | [Comments](#) | [Exit](#)

ATTACHMENT 3e

Auton: Pharmaceutics: T x

Secure | https://scholar.google.com/scholar?hl=en&as_sdt=400005&scioldt=0%2C14&cites=9194937792602916924&scipsc=&as_ylo=1988&as_yhi=1995

Apps WTS Document Delivery - BL on Demand » Home Southern Regional Lib: TIB Login - Technische Inf Penco Information Re Document Services University of Texas Lib Other bookmarks

include citations

Create alert

Pharmaceutical Formulations—Suspensions and Solutions
 P EDMAN - Journal of Aerosol Medicine, 1994 - online.liebertpub.com
 ABSTRACT Solutions and suspensions of drugs are used widely in the pharmaceutical industry for production of dosage forms for different routes of administration; for example, oral, parenteral and inhalation. Pharmaceutical solutions and suspensions might appear to
 Cited by 8 Related articles All 3 versions Cite Save

Practice research: Strategies for stability studies on hospital pharmaceutical preparations
 AC Mehta - International Journal of Pharmacy Practice, 1993 - Wiley Online Library
 IN the past five years the hospital pharmaceutical service has seen rapid changes, one of which is the way medicines are prepared and supplied to wards and other locations. Many pharmacy de- partments now provide preparations in ready-to- use form. Intravenous (IV) admixtures are
 Cited by 8 Related articles Cite Save

Serendipitous preparation of crystals of methotrexate and attempts to modify its crystal habit
 HK Chan, I Gonda - Journal of Crystal Growth, 1989 - Elsevier
 ABSTRACT A number of techniques were tried to obtain crystals of methotrexate (MTX), but without any success. Tetragonal crystals of this substance were finally obtained while attempting to prepare a complex of MTX with thymidine. Crystal habit modification of MTX
 Cited by 7 Related articles All 4 versions Cite Save

Development of sulphamethoxazole-trimethoprim spheroidal granules: factors affecting drug release in vitro
 GC Athanassiou, DM Rekkas, NH Choulis - International journal of ..., 1991 - Elsevier
 Abstract In the present study spheroidal granules of sulphamethoxazole (SMZ) and trimethoprim (TMP) were produced by the process of wet granulation of powders in a rotating pan with concomitant addition of a binder solution. Initially, preliminary trials were
 Cited by 4 Related articles All 4 versions Cite Save

[CITATION] 15. Suppositories and pessaries
 DM Collett - Pharmaceutical Practice, 1990 - Churchill Livingstone
 Related articles Cite Save

[CITATION] 16. Powders and granules
 DM Collett - Pharmaceutical Practice, 1990 - Churchill Livingstone
 Related articles Cite Save

Windows taskbar: 3:42 PM 3/27/2017

ATTACHMENT 3f

SERENDIPITOUS PREPARATION OF CRYSTALS OF METHOTREXATE AND ATTEMPTS TO MODIFY ITS CRYSTAL HABIT

Hak-Kim CHAN * and Igor GONDA **

Department of Pharmacy, University of Sydney, Sydney, NSW 2006, Australia

Received 15 September 1988; manuscript received in final form 21 October 1988

A number of techniques were tried to obtain crystals of methotrexate (MTX), but without any success. Tetragonal crystals of this substance were finally obtained while attempting to prepare a complex of MTX with thymidine. Crystal habit modification of MTX was investigated using different solvents, surfactants, “tailor-made” additives, dyes and miscellaneous other substances as well as physical factors (supersaturation, rate of cooling, degree of agitation, temperature, and growth on substrates and in a gel). Two other solid forms, in addition to the tetragonal crystals and the original powder, were found: one which has the same unit cell as the tetragonal crystals but presents itself as spheres which consist of aggregates of small tetragons, and an amorphous form.

1. Introduction

Methotrexate (MTX) (fig. 1) is a widely used chemotherapeutic agent. We have studied the solid forms of MTX [1–4] with the view to optimize its aerodynamic properties for the direct administration into the human respiratory tract in the form of an aerosol for the treatment of lung cancer. The sites of deposition of the inhaled drug particles depend on the size, shape and density of these particles [5]. Together with the drug solubility, the regional distribution dictates the concentration of released drug able to exert therapeutic and toxic effects [6]. Therefore, the design of suitable respirable particles of MTX and the knowledge of their properties are essential prerequisites of meaningful biological studies.

The commercially available MTX is a powder with a low degree of crystallinity [3,4]. Initially, we had investigated methods to crystallize MTX. Having succeeded in the preparation of well-developed tetragonal crystals of MTX [1–4], we attempted to modify the crystal habit of this sub-

stance. Numerous techniques were found accidentally, developed, or proposed, in the past to change the external shape of crystals of the same polymorph. Among the “chemical factors”, much attention has been paid to the influence of the solvent on the crystal habit of the solute. Examples of such studies are given in table 1. Solute–solvent interactions have usually been used to interpret these effects [7–11]. Watson [12] reported that the surface morphology of a solute depended on solvent–solute association and Ananikyan et al. [13] found that the best formed crystals of potassium iodate and pentaborate were grown from the least associated solutions. There is, perhaps, even more extensive literature on the effect of “additives”, or “impurities”, on the crystal habit, as it is especially relevant to industrial crystallization (e.g., refs. [20–26]). Specific chemical interactions have been used successfully to modify the crystal habits of organic molecules with “tailor-made” additives; these have molecular structures similar to the major component of the crystal and their effect can be interpreted as crystal growth inhibition in specific directions [11,27–33]. Earlier investigators [34–36] used dyes for the same purpose. Many dye molecules have functional groups capable of hydrogen bonding or

* Present address: College of Pharmacy, University of Minnesota, Health Science Unit F, 308 Harvard Street S.E., Minneapolis, Minnesota 55455, USA.

** Correspondence to this author.

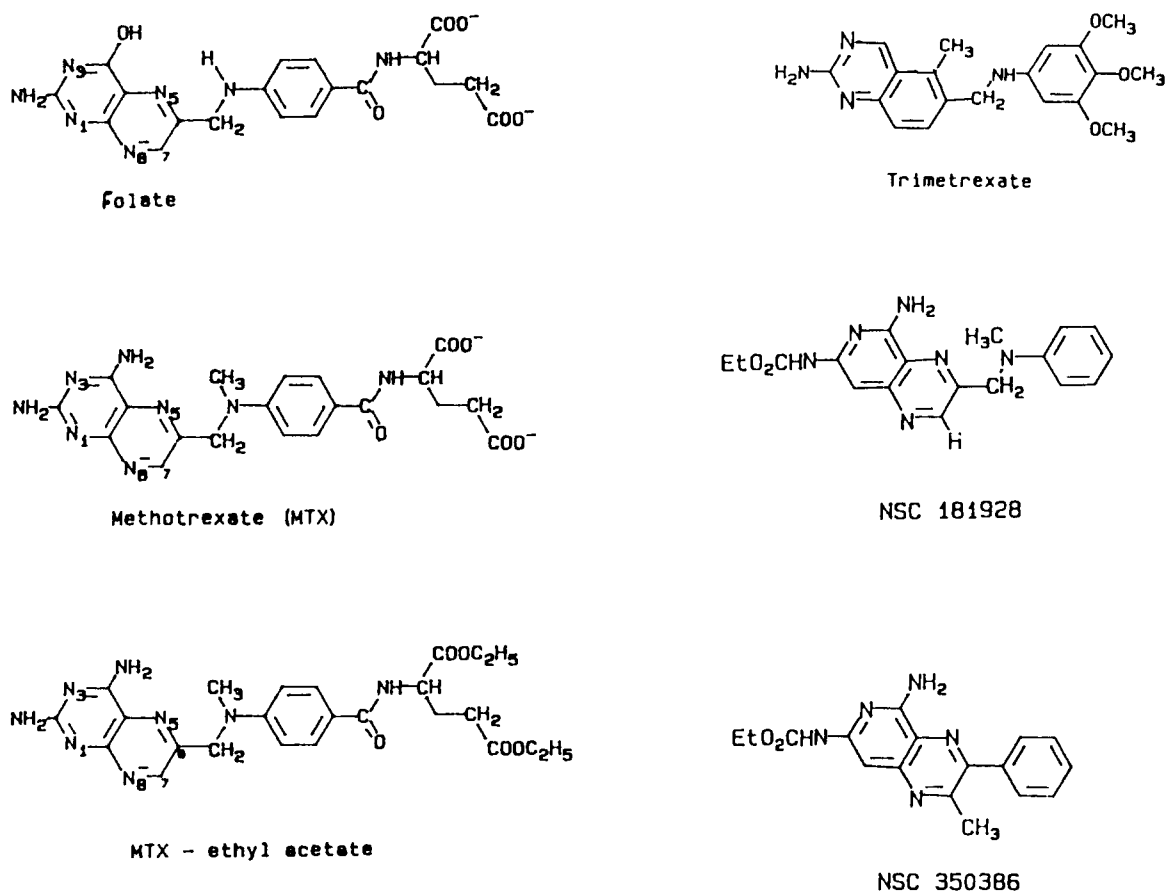


Fig. 1. The chemical structures of methotrexate and some of the additives used in the habit modification studies.

interactions via the electron cloud of the π electrons of the aromatic rings. They could be easily visualized if they adsorb on the crystal and specific adsorption to selected crystal faces could be readily detected. Similar interactions have been postulated to explain the effects of surfactants on the habit of adipic acid [14] and gypsum [37] while Garti et al. [38] suggested that the surfactants changing the crystal habit of stearic acid were causing the formation of different polymorphs. Physical factors have been known for a long time to affect the habit of crystals. Thus, the degree of supersaturation often plays an important part in the shape of crystals [14,15,17,20,39]. An empirical equation suggesting that elongated crystals are more likely to form at high supersaturation was reviewed by Haleblan [39]. Related to this prob-

ably is the effect of high rate of cooling which often leads to the formation of extreme shapes of crystals such as long needles or dendrites (e.g., refs. [12,17,20]). Temperature is an implicit parameter in factors such as supersaturation and rate of cooling but it was suggested that the sole effect of the crystallization temperature on the habit is unimportant if phase changes are excluded [40]. In contrast, Khamskii [41] did report promotion of non-isometric forms purely as a result of high temperature. It should be also mentioned that temperature may have a hidden effect when it is linked to the influence of impurities [40]. Increased agitation which promotes mass and heat transfer, is thought to enhance interfacial supersaturation and nucleation. This would be likely to encourage the formation of small, rela-

Table 1
Survey of reported effects of solvents on crystal habit of various substances

Compound	Habit	Solvent	Reference
Iodoform	Hexagonal bipyramids	Aniline	[7]
	Hexagonal prisms	Cyclohexane	
Resorcinol	Compact crystals	Water	[7]
	Very fine needles	Benzene and carbon tetrachloride	
Adipic acid	Needles	Vapour or non-polar solvents	[14]
	Hexagonal plates	Water	
Salol	Compact crystals or plates ^{a)} Only plates	Alcohols benzene, carbon tetrachloride acetone	[15]
Phthalic acid	Various habits	Polar solvents	[9]
Oxalic acid	Prismatic crystals	Acetone-water mixture	[16]
	Tabular habits	Water	
Nitrofurantoin	Tabular to needles	Formic acid	[17]
	Plates ^{b)}	Formic acid/H ₂ O or ethanol	
Succinic acid	Plates	Water	[18]
	Needles	Isopropanol	
Acetanilide	Long needles Less elongated crystals	Benzene Dimethylsulphoxide, acetone, alcohols	[19]

^{a)} Also depending on supersaturation and temperature.

^{b)} Also depending on solvent volume ratio, stirring, etc.

tively isometric crystals as observed with nitrofurantoin [17] and ammonium chloride [41]. Other physical factors reported to have caused changes of crystal habit are, e.g., pH [42] and the presence of solid substrates [43]. It should be emphasized that the spectrum of opportunities for the changes of the crystal habit is much increased if different polymorphic forms are taken into consideration. The distinction between different crystal habit due to purely external changes of shape as opposed to polymorphism needs to be appreciated [44–47] as the latter has important implications for the physical and chemical stability of the product.

2. Materials and methods

All chemicals used were of analytical grade, or better, unless stated otherwise.

Preliminary experiments on crystal habits of MTX in different solvents. 2 mg of MTX (American

Cyanamid Company, Pearl River, NY, USA) was dissolved in 1 ml of solvents and heated up if necessary. The volume of the solvent was increased up to 10 ml if dissolution was difficult. The solutions were allowed to cool and evaporate at room temperature for crystallization. The solvents used were double distilled water, methanol, ethanol, butanol, octanol, cyclohexane, carbon tetrachloride, benzene, dimethylsulphoxide (DMSO) and 0.1N HCl.

Attempts to obtain MTX crystals from the vapour phase. MTX in a round-bottom flask was heated in an oil-bath. The flask was connected to a vertical condenser tube with circulating cold water to allow condensation of any vapour evolved. The system was under vacuum (~ 50 μ Torr). Temperature of the bath was increased gradually and the condenser was observed carefully for sublimation. No sublimes were formed. The temperature was then fixed at 150–160 °C for 1 h. Afterwards, the temperature was further increased until the MTX melted with decomposition at > 200 °C. Any sub-

limates formed were collected and dissolved in 0.1N NaOH for characterization by UV-visible spectroscopy (Lambda 5 UV/VIS, Perkin-Elmer, USA).

Crystallization in the presence of thymidine. MTX was crystallized from aqueous solutions containing different amounts of thymidine (Calbiochem, La Jolla, CA, USA) as follows: 2 mg of MTX powder was weighed into a specimen tube (2 inch \times 1 inch). 1 ml of double distilled water and thymidine as required were added and heated in a 90–100 °C water bath to dissolve the MTX. The weight ratios of MTX/thymidine were 1:0.5, 1:1 and 1:5. The hot solutions were then allowed to cool to room temperature for spontaneous nucleation and crystallization. Afterwards (overnight), drops of solution containing the solid were observed under an ordinary optical microscope. A zoom stereomicroscope (Kyowa, Trinocular model SDZ-Tr-P, Japan) was subsequently used to observe the crystals inside the specimen tubes in-situ.

2.1. Systematic studies of habit modification of MTX

Solvent effect. 2 mg MTX was dissolved in hot solvent, volume 1–10 ml as necessary, and allowed to cool to room temperature for crystallization, unless described otherwise. The preliminary work (see above) showed MTX to be insoluble in non-polar solvents. Therefore, only polar solvents, or their mixtures with water, were used.

Surfactants. Cationic, anionic and non-ionic surfactants (table 2) were used. The quantities employed were adjusted so that the effects below

the critical micelle concentration could be also observed.

"Tailor-made" additives. The compounds with structures similar to MTX were used (fig. 1). They were folic acid (Hopkin and Williams Ltd., Essex, UK), 1-deaza-7,8-dihydropteridines (NSC 181928 and 350386) (Southern Research Institute, Alabama, USA), trimetrexate isethionic acid (Warner-Lambert Company, Michigan, USA) and methotrexate ethylacetate (synthesized by acidic esterification of MTX in ethanol). Because of the low solubility of NSC 181928, in addition to its saturated solution, different dilutions of this compound were also prepared from a stock solution of 1 mg/ml in dimethylformamide (DMF).

Dyes. Water-soluble dyes methyloange, methylene blue and amaranth were used.

Other additives. These were selected for a variety of reasons: capability to disturb hydrogen bonding (urea), or their known effect on crystal habits of other crystals (Al^{3+} , Cr^{3+} and Fe^{3+}) [48]. The optically active aminoacids were used for their potential to interact selectively with the chiral glutamate moiety in the MTX [27]; the L- and D-glutamic acids were employed for similar reasons.

2.2. Physical factors

Supersaturation. 0.2, 0.3, 0.5 and 1 mg/ml hot aqueous solutions of MTX were allowed to cool to room temperature for crystallization.

Rate of cooling. 1 ml of 2 mg/ml MTX in a 15 ml specimen tube dissolved at 90 °C was rapidly cooled in an ice bath. The nuclei, which formed immediately, were allowed to grow in the ice bath for several hours.

Agitation. The nuclei were prepared as in the previous experiment. The suspension was stirred for various length of time (0.5–3 min) with a magnetic bar and then allowed to stand undisturbed for crystallization.

Temperature effect. Crystallization of MTX aqueous solutions, 5 ml of 2 mg/ml dissolved at 90 °C, was allowed to proceed by slow evaporation in oil baths kept at 50 and 70 °C.

Growth on substrates. Slow cooling of hot saturated MTX aqueous solutions in the presence

Table 2
Surfactants used in the crystal habit modification studies of methotrexate

Surfactant	Molecular weight (Dalton)	Critical micelle concentration (g/dm^3)
Cetrimide	364	1.06
Cetylpyridinium chloride (CPC)	340	0.17
Dodecylpyridinium chloride (DPC)	284	0.80–8.26
Sodium lauryl sulfate (SLS)	288	2.36
Tween 80	1240	0.014

of the following surfaces was tested: cellophane sheet, dialysis membrane, sintered glass, porcelain and paraffin wax.

Growth in a gel. 1 ml of saturated aqueous MTX solution was placed on top of a layer of congealed gelatin in a specimen tube and allowed to crystallize.

3. Results and discussion

The original material as supplied from the manufacturers is a fine powder consisting of anhydral particles and showing a low degree of crystallinity [3,4]. Our first task, therefore, was to prepare crystals of MTX. In the preliminary experiments, it was found that MTX was practically insoluble in solvents of low polarity (butanol, octanol, cyclohexane, carbon tetrachloride and benzene). Water, methanol and ethanol gave anhydral particles when the solutions were allowed to evaporate at room temperature. Dimethylsulphoxide (DMSO) is an extremely good solvent for MTX; however, it evaporates only very slowly at room temperature. When water was added to the MTX solution in DMSO, small solid aggregates were formed. The same solid was observed on cooling a supersaturated MTX solution. Attempts to form MTX by sublimation and condensation of MTX vapour failed: no observable sublimation took place at temperature below 160°C. On melting, white solid product appeared on condensation but UV spectra of this material showed that it was different from MTX, presumably a decomposition product. Well-shaped tetragonal crystals were formed when MTX was crystallized in the presence of thymidine (fig. 2); some of these were found to be twins, or aggregates, under stereo zoom microscope. As the thymidine concentration increased, the crystals were becoming to be round with irregular surface. Extensive investigations by spectroscopic and chromatographic techniques showed that MTX was not forming a complex with thymidine. The new crystalline material was in fact found to be a polymorph of MTX distinct from the original powder which also retained its optically active configuration [3,4]. In order to deduce why thymidine induced the formation of

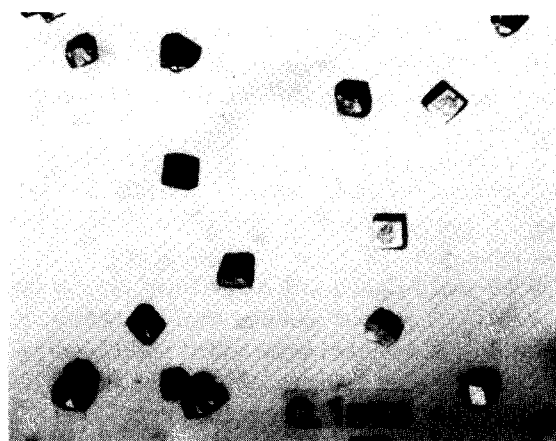


Fig. 2. Micrograph of tetragonal crystals of MTX.

tetragonal crystals of MTX, we tested the effects of (i) sugars (cf. the sugar moiety, ribose, in the thymidine) by adding either lactose or glucose (1 mg/ml) to MTX solutions (ii) pH (adjusted to 4.7 or 5.7 by HCl) and we also carried out controls by recrystallizing MTX from double distilled water. To our surprise, all the above systems including the controls now gave even better formed tetragonal crystals of MTX than in the original experiment with thymidine. Interestingly, Sutton et al. [49] reported totally independently from us on the preparation and single crystal X-ray diffraction analysis of the same solid form of MTX at almost the same time when we did [2].

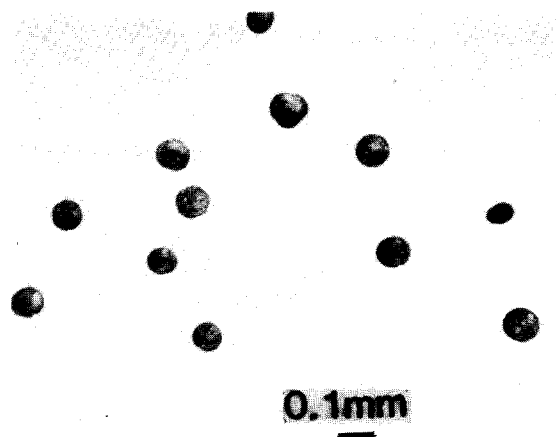


Fig. 3. Micrograph of spherically shaped crystals of MTX.

Table 3
Effect of solvents on solid form of methotrexate

Solvents	Habits
Dimethylsulphoxide (DMSO)	Anhedral particles
Dimethylformamide (DMF)	
DMF/H ₂ O	No crystals formed
DMSO/H ₂ O	
Glycerol	Very small particles with size smaller than the original powder
Glycerol : H ₂ O (%v/v)	
50 : 50	Very fine particles, no crystals
20 : 80	Very fine particles, plus some stepped spheres
10 : 90	Well-formed smooth and stepped spheres
Glycerol 3.5×10^{-2} g/dm ³	Single and twin tetragons
Propylene glycol	Tiny spherical particles
Acetone	Particles are almost identical to the original powder as MTX is insoluble
Acetonitrile	
Ethanol	Anhedral particles
Methanol	
Methanol : H ₂ O (v/v)	
6.0 : 6.0	No crystals formed at first, but spheres appeared 2 days later
5.5 : 6.0	Spheres
5.0 : 6.0	Rosettes of tetragons and spheres
4.5 : 6.0	
4.0 : 6.0	
3.5 : 6.0	Tetragon twins

The effects of various solvents on the crystal habit of MTX are shown in table 3. Water seems to be essential to obtain well formed crystals of MTX. This is somewhat surprising in view of the relatively loose crystal structure of the tetragonal form of MTX [2]. The presence of glycerol, or methanol, in water tends to promote spherical aggregates of MTX (fig. 3).

The influence of surfactants is also concentration dependent (table 4). Generally, at low concentrations (10^{-5} – 10^{-2} g/dm³), surfactants do not affect the habit of MTX. At intermediate

values (0.2–0.5 g/dm³), rounding off of the tetragons occurs and at high concentrations, crystal growth is inhibited and only very small particles are formed. Similar effects are exhibited by dyes (table 5) and “tailor-made” additives (table 6) as well as a variety of other compounds (table 7). The latter group contains a hydrogen bond effector urea, a clathrate forming agent β -cyclodextrin,

Table 4
Effect of surface active agents on solid form of methotrexate

Surfactant	Concentration (g/dm ³)	Habits
Cetrimide	2–3.5	Spherical particles (no crystals)
	0.2	Smooth spheres
	1.34×10^{-2}	Single and twin tetragons
	1.34×10^{-3}	
	1.34×10^{-4}	
CPC	2.5	Spherical particles
	0.5	Irregularly shaped spheres
	0.11	Single and twin tetragons
	1.25×10^{-2}	
	1.25×10^{-3}	
1.25×10^{-4}		
DPC	2.0	Tetragons and spheres
	1.10×10^{-2}	Single and twin tetragons
	1.10×10^{-3}	
	1.10×10^{-4}	
	1.10×10^{-5}	
SLS	1.0–2.0	Very tiny particles
	0.5	Mostly spheres with some tetragons
	1.21×10^{-2}	Single and twin tetragons
	1.21×10^{-3}	
	1.21×10^{-4}	
1.21×10^{-5}		
Tween 80	2.0	Smooth spheres
	0.2	Stepped spheres
	1.0×10^{-2}	Single and twin tetragons
	1.0×10^{-3}	
	1.0×10^{-4}	
1.0×10^{-5}		

aminoacids which could interfere with the incorporation of the glutamate moiety into the crystal lattice, and gelatin and trivalent cations reported to have affected the crystal growth of other compounds [48]. Five-fold variation of supersaturation had no effect on the crystal habit of MTX. Rapid cooling, on the other hand, prevented the formation of crystals of MTX. Instead, anhydral practices as in the original powder, were obtained. This was possibly due to excessive nucleation which depleted the mother liquor of MTX. Stirring only affected the size of the crystals but not the habit for time < 2 min. Further stirring gave very fine particles. The explanation of these observations is again probably in the depletion of the mother liquor by nucleation. Elevated temperature (50 versus 70 °C) led to less well formed tetragonal crystals but not to a major change in the habit. The presence of solid substrates in the crystallization medium did not affect the habit: tetragons were formed either away from the substrates, or on them (e.g. on the sintered glass surface). MTX

Table 5
Effect of dyes on solid form of methotrexate

Dye	Concentration (g/dm ³)	Habits
Methyl orange	2.0	Coloured, amorphous particles similar to the original powder
	2.2×10^{-2}	Orange-coloured, mainly imperfectly shaped spheres and some aggregates
	2.2×10^{-4}	Single and twin tetragons
Methylene blue	3.89×10^{-2}	Green coloured, mainly spheres some still retaining the tetragonal shape
	3.89×10^{-4}	Single and twin tetragons
Amaranth	2.0	Irregularly-shaped aggregates deep-red coloured
	6.7×10^{-2}	Red-coloured spheres
	6.7×10^{-4}	Single and twin tetragons

Table 6
Effect of "tailor-made" additives on solid form of methotrexate

Additives	Concentration (g/dm ³)	Habits
Folic acid	0.2	Spheres and aggregates of spheres
	2×10^{-2}	Single and twin tetragons
	2×10^{-3}	
NSC-350386	2×10^{-4}	
	2.0	Rounding increased with additive concentration, spheres with smooth instead of stepped surface
	0.2	
	2×10^{-2}	
	2×10^{-3}	Tetragons with stepped surface and rounding
NSC-181928 ^{a)}	0.5	No crystals formed
	0.1	Rosettes
	0.01	Spheres
	Saturated aqueous solution	Spherical aggregates
MTX-ethyl acetate	1-2	Very fine particles, no crystals
	0.4	Spheres of various sizes with aggregates
	0.2	Single spheres
Trimetrexate	1.0	Smooth spheres
	0.2	Stepped spheres
	0.01	Single and twin tetragons

^{a)} Prepared from DMF solutions.

grown in gelatin formed spheres but the purity of this new substance was not checked. It is, in fact, possible, that gelatin reacted chemically with MTX.

4. Conclusions

Methotrexate was shown recently to exist in the form of at least three different polymorphs. The

Table 7
Effect of miscellaneous compounds on solid form of methotrexate

Additives	Concentration (g/dm ³)	Habit
Urea	1.5	Stepped spheres
	5.0×10^{-3}	Single and twin tetragons
	5.0×10^{-5}	Single tetragons highly favoured
β -cyclodextrin	50	Solubilization
	5.0	Spheres
	0.5	Single and twin tetragons
	5×10^{-2}	
	5×10^{-3}	
	5×10^{-4}	
dl-Histidine	20 (5 mg MTX used)	Solubilization
	2.5 (5 mg MTX used)	Few smooth spheres
	1.0	Spheres and tetragons
	0.2	
L-Ornithine	20	Very fine particles
	2-0.5	Stepped spheres
	0.2	Single tetragons with a few spheres
L-Lysine	2.0	Single and twin tetragons
D-Glutamic acid	6.5	Very fine particles, no crystals
	2.0	Stepped spheres
	1.0	
	0.65	Single and twin tetragons
	6.5×10^{-2}	Single tetragons
	6.6×10^{-3}	
L-Glutamic acid	10	Very fine particles
	5.5	Both stepped and smooth spheres
	2.5	
	2.0	Tetragons, some round and some twins
	0.67	
	6.7×10^{-2}	
Gelatin	1.2	Spheres and ellipses
	0.5	Mostly ellipses with a few spheres
	0.2	
	0.1	Spheres
AlCl ₃ CrCl ₃ FeCl ₃	0.1	Fine particles only, no crystals (MTX not fully dissolved as solubility in solutions is low)
	1×10^{-3}	Single and twin tetragons in AlCl ₃ ; twin tetragons and stepped spheres in CrCl ₃ ; stepped and smooth spheres in FeCl ₃
	1×10^{-4}	

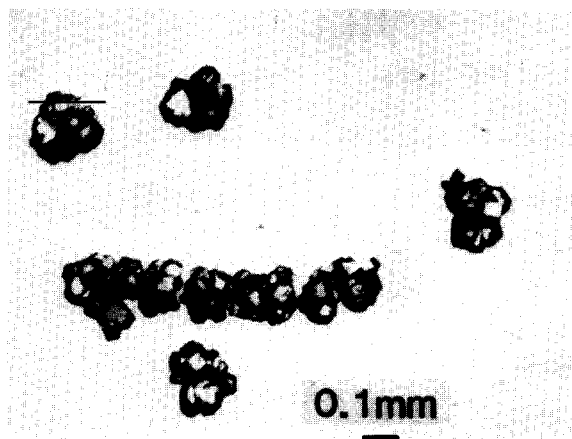


Fig. 4. Micrograph of spherically shaped "rosettes" of MTX.

form which exhibits the greatest tendency to exist as well-formed crystals, appears to be the most stable polymorph of MTX [1-4]. It is interesting that this solid was prepared in an attempt to make a crystalline complex of MTX with thymidine because it was thought at the time that well formed crystals of pure MTX could not be obtained. Following the success with preparation of the distinct tetragonal crystals of MTX, various methods to modify its crystal habit were employed. Water seems to be essential for the formation of crystals of MTX of observable size; this is probably the consequence of the role of water in the crystal lattice of MTX [2]. Mixtures of polar solvents with water and other additives all showed a similar concentration-dependent effect on the habit of MTX: at low concentrations, formation of spheres was observed; these were shown to be the same polymorph as the tetragonal crystals of MTX [3,4] and were probably mainly aggregates of microscopic tetragons of MTX (fig. 4). At higher concentration, the effect of the additives was disruption, or prevention of formation, of the crystalline structure.

Rounding of crystals is an interesting phenomenon which is often desirable to improve the flow and compaction properties of solid materials [50-56]. Although it is not practically relevant for MTX which is a drug normally used in an injectable form, nevertheless the present results suggest that a relatively simple manipulation may cause

formation of round solid particles during crystallization. We could suggest two explanations for this phenomenon. Firstly, inhibition of rate of crystal growth can give rise to the formation of a large number of small crystals which will agglomerate to minimize the total surface energy of the system. At high concentrations of additives, the inhibition of crystal growth is complete and, instead, amorphous MTX is formed. Secondly, we may consider the non-specificity of the rounding effect to be due to intermolecular interactions at the growing crystal faces but we have to postulate that all faces of MTX are approximately equally affected by all the additives tested. This could be true in the general sense for MTX since all the faces contain both hydrophilic (e.g. $-NH_2$ in pteridine and $-COOH$ in glutamate) and hydrophobic (e.g. the benzoyl and pteridine rings) groups [2]. Therefore, the additives, through their hydrophilic and hydrophobic parts could affect the growth of the nuclei, or small crystals, in an isotropic fashion. It may seem surprising that even "tailor-made" additives could behave in this manner. However, if one considers a typical representative of this class of compounds (fig. 1) folic acid, the isotropic influence can be explained. Folic acid has a chemical structure similar to MTX and would be therefore expected to form hydrogen bonds along the *c*-axis [2] and thus inhibit growth in that direction by breaking the chain of MTX molecules. However, since hydrogen bonding is possible along other crystal directions as well [2], the overall effect of folic acid could be quite isotropic. The fact that we encountered initially a very strong resistance of MTX to form well formed crystals is, perhaps, an indication that the crystal growth of this substance can be relatively easily inhibited, or disrupted, leading to the appearance of particles formed under the influence of non-specific, isotropic interactions.

Acknowledgements

We are grateful to Ms. Sandy Butler for typing the manuscript, to Dr. Ian Threadgold from the Department of Geology for encouragement and help with crystallography and to Dr. Andrew

Cheung for general advice on methotrexate. The following chemicals were supplied to us kindly free of charge: methotrexate (American Cyanamid Company, Pearl River, USA), 1-deaza-7,8-dihydropteridines (NSC 181928 and 350386, Pharmaceutical Chemistry Division, Southern Research Institute, USA), trimetrexate isethionic acid (Pharmaceutical Research Division, Warner-Lambert Company, USA) and β -cyclodextrin (Nikon Shokuhin Kako Co. Ltd., Japan). H.-K.C. was supported by a scholarship as a part of a University of Sydney Special Project Grant to I.G.

References

- [1] H.-K. Chan, T.W. Hambley and I. Gonda, *Australian J. Hosp. Pharm.* 16 (1986) 66.
- [2] T.W. Hambley H.-K. Chan and I. Gonda, *J. Am. Chem. Soc.* 108 (1986) 2103.
- [3] H.-K. Chan, *Crystal Growth and Aerodynamics of Drug Particles*, PhD Thesis, University of Sydney (1988).
- [4] H.-K. Chan and I. Gonda, *Intern. J. Pharmaceut.*, submitted.
- [5] I. Gonda, in: *Pharmaceutics - The Science of Dosage Form Design*, Ed. M.E. Aulton (Churchill, Livingstone, Edinburgh, 1988) pp. 341-358.
- [6] I. Gonda, *J. Pharm. Sci.* 77 (1988) 340.
- [7] A.F. Wells, *Phil. Mag.* 37 (1946) 184.
- [8] A.F. Wells, *Disc. Faraday Soc.* 5 (1949) 197.
- [9] R.J. Davey, in: *Current Topics in Materials Science*, Vol. 8, Ed. E. Kaldis (North-Holland, Amsterdam, 1982) pp. 431-479.
- [10] Z. Berkovitch-Yellin, *J. Am. Chem. Soc.* 107 (1985) 8239.
- [11] L. Addadi, Z. Berkovitch-Yellin, I. Weissbuch, J. Van Mil, L.J.W. Shimon, M. Lahav and L. Leiserowitz, *Angew. Chem. Intern. Ed. Engl.* 24 (1985) 466.
- [12] D.H. Watson, in: *Proc. 3rd Intern. Conf. on Electron Microscopy*, London, 1954, p. 497.
- [13] T.A. Ananikyan, A.G. Nalbandyan and H.G. Nalbandyan, *J. Crystal Growth* 73 (1985) 505.
- [14] A.S. Michaels and A.R. Colville, *J. Phys. Chem* 64 (1964) 13.
- [15] W. Kleber and H. Raidt, *Z. Physik, Chem.* 222 (1963) 1.
- [16] J.L. Torgesen and J. Strassburger, *Science* 146 (1964) 53.
- [17] N. Garti and F. Tibika, *Drug Dev. Ind. Pharmacy* 6 (1980) 379.
- [18] R.J. Davey, J.W. Mullin and M.J.I. Whiting, *J. Crystal Growth* 58 (1982) 304.
- [19] M.C. Etter, D.A. Jahn and B.S. Donahue, *J. Crystal Growth* 76 (1986) 645.
- [20] J.W. Mullin, *Crystallization* (Butterworths, London, 1972).
- [21] J.W. Mullin, Ed., *Industrial Crystallization* (Plenum, New York, 1976).
- [22] R.J. Davey, in: *Industrial Crystallization* 78, Eds. E.J. de Jong and S.J. Jančić (North-Holland, Amsterdam, 1979) pp. 169-183.
- [23] E.J. de Jong and S.J. Jančić, Eds., *Industrial Crystallization* 78 (North-Holland, Amsterdam, 1979).
- [24] S.J. Jančić and E.J. de Jong, Eds., *Industrial Crystallization* 81 (North-Holland, Amsterdam, 1982).
- [25] S.J. Jančić and E.J. de Jong, Eds., *Industrial Crystallization* 84 (Elsevier, Amsterdam, 1984).
- [26] J. Go and D.J.W. Grant, *Intern. J. Pharmaceut.* 36 (1987) 17.
- [27] Z. Berkovitch-Yellin, L. Addadi, M. Idelson, L. Leiserowitz and M. Lahav, *Nature* 296 (1982) 27.
- [28] Z. Berkovitch-Yellin, J. van Mil, M. Idelson, M. Lahav and L. Leiserowitz, *J. Am. Chem. Soc.* 107 (1985) 3111.
- [29] L. Addadi, Z. Berkovitch-Yellin, N. Domb, E. Gati, M. Lahav and L. Leiserowitz, *Nature* 296 (1982) 21.
- [30] L. Addadi and S. Weiner, *Mol. Crystals Liquid Crystals* 134 (1986) 305.
- [31] I. Weissbuch, L.J.W. Shimon, L. Addadi, Z. Berkovitch-Yellin, S. Weinstein, M. Lahav and L. Leiserowitz, *Israel J. Chem.* 25 (1985) 353.
- [32] I. Weissbuch, Z. Berkovitch-Yellin, L. Leiserowitz and M. Lahav, *Israel J. Chem.* 25 (1985) 362.
- [33] I. Weissbuch, D. Zbaida, L. Addadi, L. Leiserowitz and M. Lahav, *J. Am. Chem. Soc.* 109 (1987) 1869.
- [34] C. Frondel, *Am. Mineralogist* 25 (1940) 91.
- [35] W.G. France, in: *Colloid Chemistry*, Ed. J. Alexander (Reinhold, New York, 1944) pp. 443-457.
- [36] H.E. Buckley, *Crystal Growth* (Chapman and Hall, London, 1952).
- [37] J. Schroeder, W. Skudlarska, A. Szczepanik, E. Sikorska and S. Zielinski, in: *Industrial Crystallization*, Ed. J.W. Mullin (Plenum, New York, 1976) pp. 263-268.
- [38] N. Garti, E. Wellner and S. Sarig, *J. Crystal Growth* 57 (1982) 577.
- [39] J.K. Haleblan, *J. Pharm. Sci.* 64, (1975) 1269.
- [40] R. Boistelle, in: *Industrial Crystallization*, Ed. J.W. Mullin (Plenum, New York, 1976) pp. 203-214.
- [41] E.V. Khamskii, in: *Industrial Crystallization*, Ed. J.W. Mullin (Plenum, New York, 1976) pp. 215-221.
- [42] R.J. Davey, *J. Crystal Growth* 76 (1986) 637.
- [43] I. Tarjan and M. Matrai, *Laboratory Manual on Crystal Growth* (Akademiai Kiado, Budapest, 1972).
- [44] N. Garti, E. Wellner and S. Sarig, *Kristall Tech.* 15 (1980) 1303.
- [45] S.R. Byrn, *Solid State Chemistry of Drugs* (Academic Press, New York, 1982) pp. 79-148.
- [46] H. Wollmann and V. Braun, *Pharmazie* 38 (1983) H.1, 5.
- [47] K. Sato and R. Boistelle, *J. Crystal Growth* 66 (1984) 441.
- [48] S.V. Verdager and R.R. Clements, *J. Crystal Growth* 79 (1986) 198.
- [49] P.A. Sutton, V. Cody and G.D. Smith, *J. Am. Chem. Soc.* 108 (1986) 4155.
- [50] Y. Kawashima T. Handa, H. Takeuchi, M. Okumura, H. Katou and O. Nagata, *Chem. Pharm. Bull.* 34 (1986) 3376.

- [51] Y. Kawashima, T. Handa, H. Takeuchi and M. Okumura, Chem. Pharm. Bull. 34 (1986) 3403.
- [52] Y. Kawashima, M. Okumura and H. Takenaka, Science 216 (1982) 1127.
- [53] Y. Kawashima, M. Okumura, H. Takenaka and A. Kojima, J. Pharm. Sci. 73 (1984) 1535.
- [54] C.J. McCarthy, G.S. Riley and J.E. Rees, J. Pharm. Pharmacol. 35 (1983) (suppl.) 2P.
- [55] J.N. Staniforth, Intern. J. Pharm. Tech. Proc. Mfr. 5 (1984) 1.
- [56] A. Sano, T. Kuriki, T. Handa, H. Takeuchi and Y. Kawashima, J. Pharm. Sci. 76 (1987) 471.

ATTACHMENT 4

About the size of Google Scholar: playing the numbers

Enrique Orduña-Malea¹, Juan Manuel Ayllón², Alberto Martín-Martín²,
Emilio Delgado López-Cózar²

¹ EC3: Evaluación de la Ciencia y de la Comunicación Científica, Universidad Politécnica de Valencia (Spain)


² EC3: Evaluación de la Ciencia y de la Comunicación Científica, Universidad de Granada (Spain)

ABSTRACT

The emergence of academic search engines (Google Scholar and Microsoft Academic Search essentially) has revived and increased the interest in the size of the academic web, since their aspiration is to index the entirety of current academic knowledge. The search engine functionality and human search patterns lead us to believe, sometimes, that what you see in the search engine's results page is all that really exists. And, even when this is not true, we wonder which information is missing and why. The main objective of this working paper is to calculate the size of Google Scholar at present (May 2014). To do this, we present, apply and discuss up to 4 empirical methods: Khabsa & Giles's method, an estimate based on empirical data, and estimates based on direct queries and absurd queries. The results, despite providing disparate values, place the estimated size of Google Scholar in about 160 million documents. However, the fact that all methods show great inconsistencies, limitations and uncertainties, makes us wonder why Google does not simply provide this information to the scientific community if the company really knows this figure.

KEYWORDS

Google Scholar / Academic Search Engines / Size Estimation methods.

 <p>Grupo de Investigación EC3 Evaluación de la Ciencia y de la Comunicación Científica</p>	<p>EC3's Document Serie: EC3 Working Papers N° 18</p> <p>Document History Version 1.0, Published on 23 July 2014, Granada</p>
<p>Cited as Orduña-Malea, E.; Ayllón, J.M.; Martín-Martín, A.; Delgado López-Cózar, E. (2014). <i>About the size of Google Scholar: playing the numbers</i>. Granada: EC3 Working Papers, 18: 23 July 2014</p>	
<p>Corresponding author Emilio Delgado López-Cózar. edelgado@ugr.es Enrique Orduña-Malea. enorma@upv.es</p>	

1. INTRODUCTION

The calculation of the size of the Web in general (Lawrence & Giles, 1998; 1999; Dobra & Fienberg, 2004, among others) and the academic web in particular (Khabsa & Giles, 2014) has generated a debate in the scientific arena over the last two decades at different levels, among which we can highlight the following: a) from an information perspective (the extent to which all the knowledge produced is actually indexed, searchable, retrievable and accessible from a catalogue, index or database); b) from a methodological level (how to calculate it as accurately as possible); and c) from a socioeconomic perspective (how the composition and evolution of these contents affect their consumption in different countries according to different social, economic and political issues).

The emergence of academic search engines (Ortega, 2014), Google Scholar and Microsoft Academic Search essentially, has revived and increased the interest in the size of the academic web, changing the focus of the question: their aspiration to index the entirety of current academic knowledge leads us to believe, sometimes, that what you see in the search engine result page is all that really exists. And, even when this is not true, we wonder which information is missing.

In traditional bibliographic databases (WoS, Scopus), finding out the size (measured by the number of records at a given time) is a fairly trivial matter (it's only necessary to perform a query in the search interface), because the entire universe is catalogued and under control (always accounting for a low error rate due to a lack of absolute normalization in the catalogue). Moreover, the evolution of these databases is cumulative: the number of records always grows and never decreases, except for the exceptional elimination of records due to technical or legal issues. However, in the case of academic search engines, these assertions not always apply, making both the calculation of their size, and tracking the evolution of their data, extremely complicated tasks.

Despite the high dynamism of the Web (contents are continually added, changed and/or deleted worldwide) and the inherent technical difficulties to catalogue and update such a vast and diverse universe (Koehler, 2004), the main problem is the opaque information policies followed by those responsible for such databases, particularly Google Scholar (GS). Its policies differ from those of other discovery services like Microsoft Academic Search (MAS) or Web of Science (WoS), where just running a query that aims to get the number of documents indexed in a specific date will get you the answer instantly.

The recent work of Khabsa and Giles (2014) has estimated the number of circulating documents written in English in the academic Web on 114 million (and also that of those, GS has around 99.8 million), employing a procedure based on the use of the Lincoln-Petersen (capture-recapture) method, from the citations to a sample of articles written in English, included both in GS and MAS. However, this procedure (discussed in greater detail in the methods and results sections) leads us to formulate a number of questions, namely: is it possible to calculate the size of the academic web in general (and Google

Scholar in particular)? And, does Google Scholar really cover 87% of the global academic Web?

2. OBJECTIVES

The main objective of this working paper is to calculate the size of Google Scholar at present (May 2014). To do this, the specific objectives are the following:

- Explain and apply various empirical methods to estimate the size of Google Scholar.
- Point up the strengths and weaknesses of the different methods for estimating the size of Google Scholar.

3. METHODS

To estimate the size of Google Scholar we propose and test 4 different procedures, explained in further detail below:

a) Khabsa & Giles's method

This procedure is taken directly from the research carried out by Khabsa and Giles (2014), and recently "digested" by Orduña-Malea et al (2014). The method is as follows:

First, 150 English academic papers (journal and conference papers, dissertations and masters theses, books, technical reports and working papers) are selected from Microsoft Academic Search (MAS). These articles are randomly sampled from the most cited documents in each of the 15 fields covered by MAS (10 documents per field), considering only documents with less than 1,000 citations. These 150 articles are verified to be included in Google Scholar as well.

After this, the number of incoming citations to the 150 selected documents are obtained both from MAS (41,778 citations) and Google Scholar (86,870). The overlap between GS and MAS (citing documents contained in both search engines) is computed by means of the Jaccard similarity index (0.418). These data is collected in January 2013.

The number of scholarly documents available on the web is estimated then using the Lincoln-Petersen method (capture/recapture):

$$\frac{R}{M} = \frac{C}{N}$$

Where:

- N** (population) = size of GS + size of MAS;
- M** (elements captured in the first sample) = size of GS;
- C** (elements captured in the second sample) = size of MAS;

R (elements recaptured in the second sample) = overlap between GS and MAS, measured as the number of citations shared).

Note that the expression on the left side of the equal symbol is the original meaning of the Lincolnm-Petersen indicators, and the one on the right side is the analogy used by Khabisa and Giles. It is also noteworthy that whereas M and C correspond to the number of documents indexed on GS and MAS respectively, the overlap between them (R) is measured through the citing documents to the cited documents of the sample.

Since C is taken from the data provided by Microsoft (48,774,764 million documents, which is reduced to 47,799,627 after applying a correction factor for English of 0.98), and R is calculated directly by the authors, then N can be directly isolated, and subsequently M (the size of Google Scholar).

b) Estimates from empirical data

The second method consists of making estimates from empirical studies that have previously worked with samples and have compared GS with other databases. From these comparisons and the differences in coverage, a correction factor could be obtained, and consequently a hypothetical projection may be proposed.

To this end, an extensive collection of empirical studies dealing with the calculation of the sizes of academic databases has been gathered in Appendix I. It shows, in table format, each collected work, the database analysed (GS, WoS, MAS, Scopus, Pubmed, etc.), the unit of analysis (citations, documents, etc.) and the sample considered.

These studies use different units of analysis (journals, articles, books, etc.) and metrics (citation count, h-index, impact factor, etc.), but for our purpose, the synthesis of the results can only be applied to samples that are comparable to each other in the following levels:

- Studies examining the same databases used as data source.
- Studies working with documents or with unique citation documents.
- Studies that make comparisons between documents written in the same language (or do not make a distinction by language).

Hence, we have categorized the data offered in Appendix I according to the unit of study: journals, books, etc.; the indicator measured: citations, documents, citations per document (Appendix II); and according to the language of the documents (Appendix III).

Finally, only those studies comparing GS and WoS have been considered, since only just two studies provided information about empirical comparisons between GS and Scopus, and the remaining databases were even less well represented.

Then, for each case study we obtained the proportion between both databases dividing the number of documents retrieved for GS by the number of documents gathered for WoS. Finally, the geometric mean and the median of all studies are carried out to get a crude, but indicative, correction factor.

This same procedure has also been applied to the comparison of unique citing documents as unit of analysis (citing documents indexed in GS and non-indexed in WoS, and vice versa).

c) Direct query

The third method is based on interrogating the database itself, at least to the extent that this is possible. This can be done by two procedures: a) using the custom date range for the complete period of time; b) using the custom date range year by year and adding the results together at the end.

To this end, first we directly queried (by means of an empty query search) Google Scholar (the <google.scholar.com> version) filtering by single years¹ and gathering the estimated number of results, also called Hit Count Estimates (HCE). We have processed the data from 1700 to 2013 (data prior to 1700 are practically non-existent; namely, 49 records are found in the range 1000-1700). After this, the number of records obtained for each year are added together.

At the same time, we set the custom range from 1700 to 2013, to gather all documents in the period in the same query. This data collection process was carried out in May 2014.

The raw hit count estimates is comprised of three different types of results (some of which may be included or excluded in a query): ordinary records (documents indexed on Google Scholar, providing a link to the full text or to a paid gateway), citations (references to documents not indexed on Google Scholar), and patents (documents extracted from Google Patents).

To test the potential influence of citations and patents in the size of Google Scholar, we retrieved the following data for each year:

- All documents (records + citations + patents);
- Records + citations;
- Records + patents;
- Only records, excluding citations and patents.

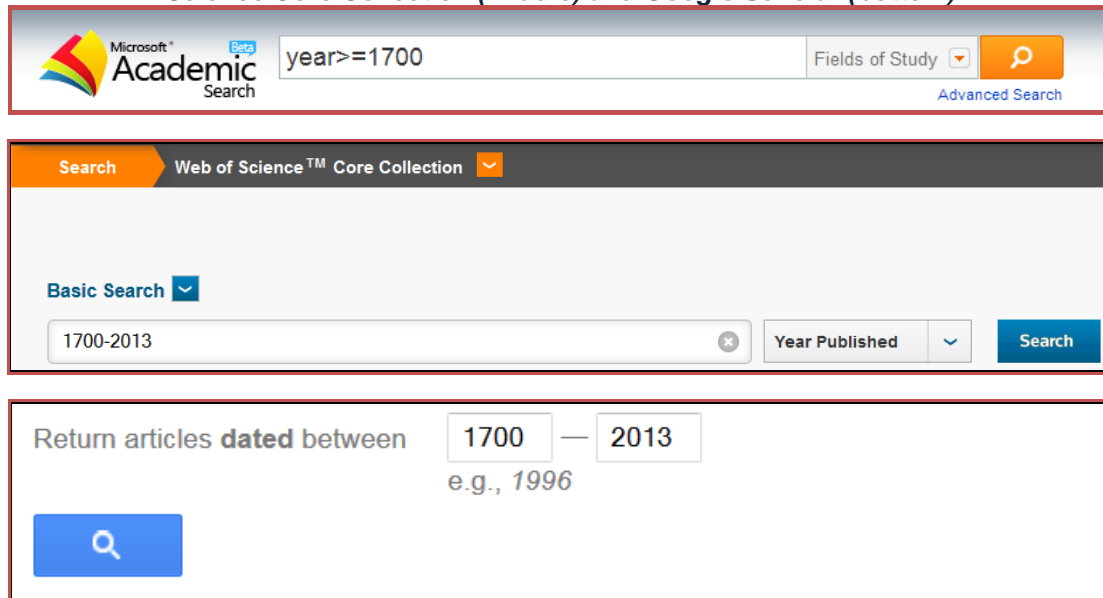
Finally, direct queries were performed on Microsoft Academic Search as well as on Web of Science Core Collection , with the aim of gathering both sectional and longitudinal data about the size of these databases at present (data were collected in May 2014) suitable to compare with those obtained previously for Google Scholar:

- Microsoft Academic Search: from <academic.research.microsoft.com>, direct query via the "year" command is performed.

- Web of Science Core Collection: the size is obtained from the basic search interface, specifying the year in the "Year published" field. Data concerning language and type of document were gathered as well.

For each database, a query per year (from 1700 to 2013) and a global query from 1700 were performed. In Figure 1 we show the interface of each of the 3 databases queried.

Figure 1. Direct query performed against Microsoft Academic Search (top), Web of Science Core Collection (middle) and Google Scholar (bottom)



d) Absurd query

The last method proposed is based on the use of some characteristics of the Boolean logic that are supported in Google Scholar's search box. In this case, the goal is to compose a query that somehow requires Google Scholar to return all its records. Although nowhere in the official documentation is it stated that such a query exists, we have run test queries using the following syntax: <common_term -site:unexistent_site>

The idea behind this is to query the occurrences of a very common term (likely to appear in almost all written records), and to filter out its appearances in a nonexistent site, which means that we are implicitly selecting every existing site in our query.

For example: <a -site:ssstfsffsdasdfs.com>, or <1 -site:ssstfsffsdasdfs.com>.

The reason for including a term before the "-site" command is that this command does not work on its own. These queries were run on Google Scholar (including and excluding citations and patents) in June 2014.

As in the case of the direct query method, the queries were performed in two different ways: a) setting the custom range from 1700 to 2013; b) running a query for each year and adding the results together.

4. RESULTS

Various estimates of the size of Google Scholar, as calculated from each of the 4 procedures outlined above, are offered, additionally providing a discussion about their advantages, disadvantages and shortcomings.

4.1. Khabsa & Giles's method

Khabsa and Giles (2014) estimate on 114 million the number of circulating documents written in English in the academic Web, and on 99.3 million the number of English documents in Google Scholar.

Nonetheless, these figures may be biased due to the following considerations: applying the Lincoln-Petersen method; considering GS and MAS as the whole academic web universe; collecting biased data; and considering a biased database for the estimation. Let's discuss each of these issues in greater detail.

a) Lincoln-Petersen estimate

The problem is not only related to the possible growth of the population among samples, or the condition on the equal probability of each element to be recaptured (conditions in the application of this estimate method), but also to the assumption that each sample is applied to different universes (Google Scholar and Microsoft Academic Search), when the original method consists of 2 (or more) captures in the same universe.

The authors use a complementary method obtaining similar results, and this reinforces the results. In any case, a reasonable uncertainty exists that should at least be noted.

b) Size of the Academic Web (N)

The estimate of the total size of the academic web was probably undersized: the summation of Google Scholar and Microsoft Academic Search is still far from representing the total academic web space, though it undoubtedly makes up a very high percentage.

N (i.e., the scholarly academic public web) is considered to be the summation of Google Scholar and Microsoft Academic Search. This issue keeps out other databases, such as Google Books (it is well-known that Google Scholar and Google Books databases do not entirely match), among others, although we are aware that these missing results are probably low and statistically insignificant.

Besides this, there is a more fundamental concern: the low indexation of institutional repositories on Google Scholar (Orduña-Malea and Delgado López-Cózar, in press) as well as some of GS's indexing policies (for example, files over 5MB are not indexed, a procedure which is especially critical for doctoral theses).

For these reasons, the assumption that Google Scholar covers 87% of the global academic Web (even considering only English documents) may constitute an over-representation.

Moreover, the method considers Microsoft Academic Search as a valid universe (48.7 million of documents, as of January 2013). However, the information about the total size of MAS is confusing at present. Microsoft Azure Marketplace shows (as of May 2014) 39.85 million documents, which does not match the data used in Khabsa & Giles's research; from the information collected on the Web, we can estimate 45.3 million documents, and 45.9 million documents if a query is performed manually in the website platform (as of May 2014). How can this disparate information affect the calculation of "N"?

c) Biased sample

On one hand, the sample of cited documents is not absolutely random because only documents in English with less than 1,000 citations are considered. The authors acknowledge this limitation: search engines impose a restriction on the number of retrievable results for all type of queries, unless an Application Programmable Interface (API) is provided (and Google Scholar does not provide an API at the moment).

Accessing to only the first 1,000 documents may bias the sample in an unknown way (and maybe differently for each field), although we can assume (though not demonstrate) that these records contain the more formalized, visible, and more circulating and cited documents. Moreover, this statistical error is equally distributed to the 15 samples, thus reducing its effect.

On the other hand, the sample is uniform for each field (10 articles) despite the fact that the output size of each field is quite different. This may introduce an important bias in the estimation.

d) Biased database

The data sample is taken from MAS, and this database is biased, among other reasons, due to the diversity of language and document types:

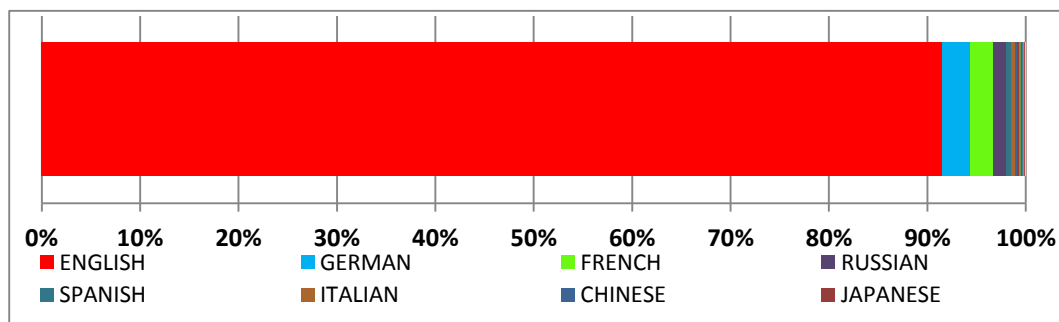
- It is oriented to collect English written literature, and specifically that produced in English-speaking countries.
- It is oriented to very specific types of publication: articles and conference papers. Although it has recently incorporated monographs, they are still a genuine minority. In contrast, document types in GS are not as skewed towards journal articles and conference as they are in MAS.

This concern applies not only to MAS but also to Scopus and WoS. Since it is not possible to take data from MAS according to language and document type, let us discuss these topics considering WoS as a basis for our examples.

Biases in the sample by language

The WoS database is clearly biased towards English, as is well known. In Figure 2 you can empirically test the proportion of languages for documents indexed in the Web of Science for the period 1900 to 2014, where English amounts to slightly over 90% of the records.

Figure 2. Language of documents indexed in the Web of Science (1900-2014)



Source: self-elaborated

The proportion of English documents on MAS is high as well. Khabsa and Giles estimate in their sample a 98% of documents in English. However, the proportion of English documents on Google Scholar, though high, is lower than the one obtained in WoS and MAS. This different proportion may influence the estimate.

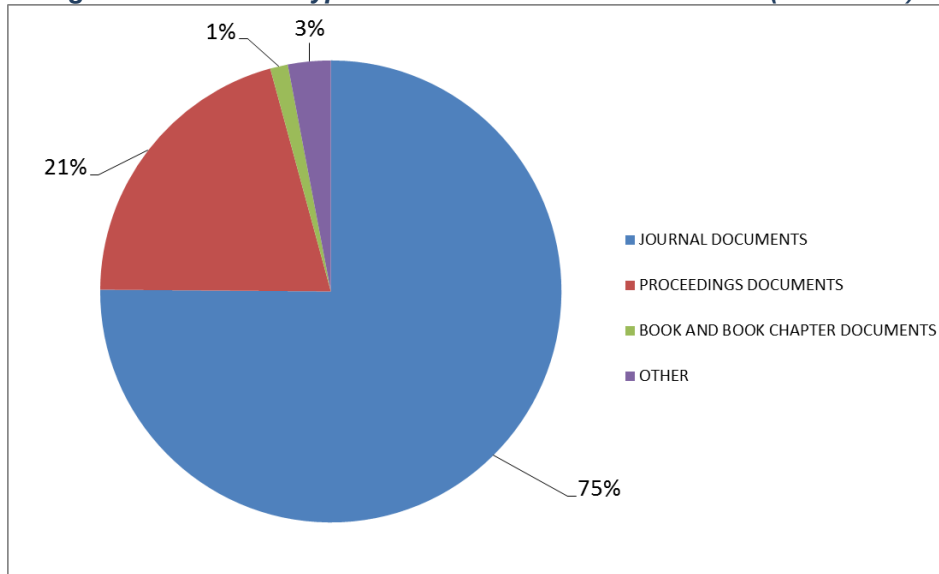
Otherwise, the estimated size of Google Scholar in English offered by Khabsa and Giles was probably oversized because, although the sample items (cited documents) were written in English, the citing documents could potentially be articles written in other languages. This means that it is highly probable that the measurement was not entirely limited to the English universe.

Biases in the sample by document type

On the other hand, there is a bias towards articles. In Figure 3 we show the percentages of documents by type, collected in the Web of Science, for the period 1900 to 2014, where the "Journal document type" (composed by articles, meeting abstracts, editorial material and letters) represents 75% of all documents, whereas "Book and Book chapters" only 1%.

This bias creates a clear infra-representation of disciplines using other communication vehicles different than the "Journal article" format. Therefore, and as happened in the case of languages, the distribution of document types is not the same in WoS than in Google Scholar or Microsoft Academic Search.

Figure 3. Document types indexed in the Web of Science (1900-2014)



Source: self-elaborated

Unfortunately, you cannot perform this typological analysis directly on Google Scholar, because searching by type of document is not supported. Taking a look at the empirical data available in Appendix II, and considering that it is difficult to summarize all this information accurately because each study has its own distinct nature and deals with a different field, we can surmise that journal articles make up, on average, 65% of the total number of records in GS. In the case of Microsoft Academic Search, you can only manually check the number of document in articles and conferences.

If we carry out a global search in WorldCat² (the largest bibliographic information system in the world), approximately 2 billion items in more than 470 languages are obtained, but obviously this catalog covers both scientific and non-scientific (or digitized) documents. In any case, this procedure can itself help us to determine the proportion of books and other documents that are not included in traditional bibliographic databases, and which may not be indexed on Google Scholar as well. For example, a search for “thesis” (doctoral, masters or degree) gives a current figure of 16.3 million documents (Figure 4).

Figure 4. Searching for Theses in Worldcat

WorldCat[®] search results for 'ti:a' (16,347,205 results). The first result is 'El habla de Sistema' by Joseph A. Fernández, a thesis/dissertation in Spanish published by the Consejo Superior de Investigaciones Científicas.

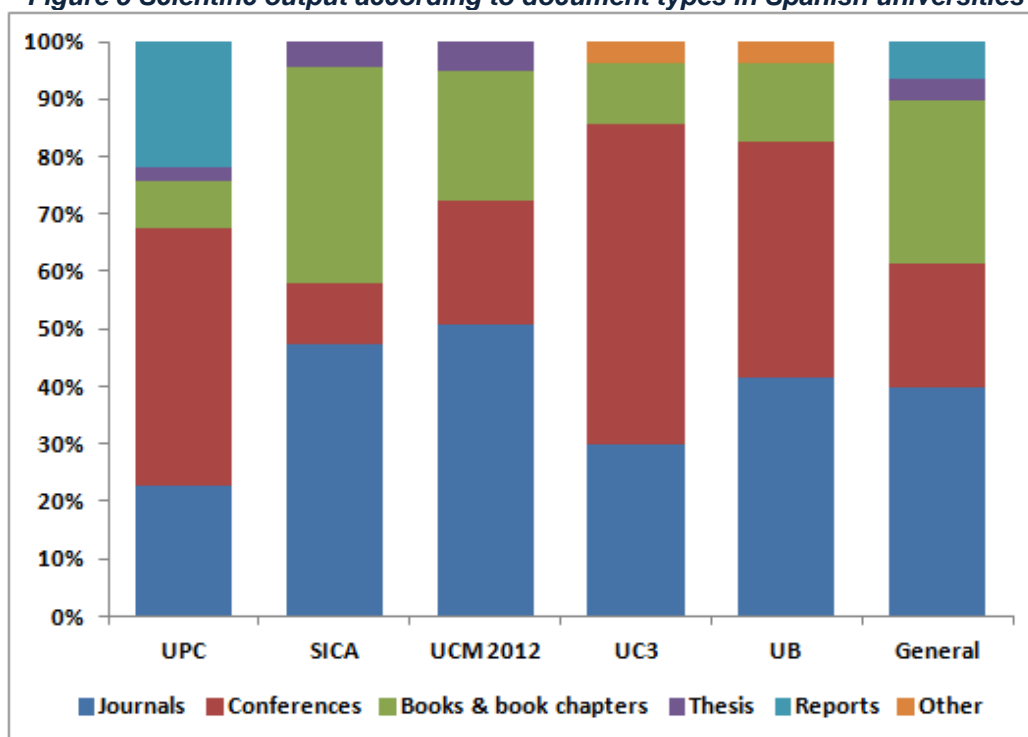
Source: Worldcat

Although Google Scholar indexes doctoral theses (since university institutional repositories are indexed, as well as some library services), this issue raises the following question: how many of these doctoral theses are indexed on Google Scholar, keeping in mind the limitation of a maximum of 5MB per file?, and how many are indexed on Microsoft Academic Search?

Moreover, if we analyze the production of scientific institutions, especially universities (the main producers of academic output), we can observe this production greatly varies from one university to another. Figure 5 offers the percentage of documents according to document type for various Spanish universities, such as the Polytechnic University of Catalonia (UPC), University of Barcelona (UB), Carlos III University of Madrid (UC3), and the aggregate value for all the universities in the region of Andalusia (collected from the Scientific information Service of Andalusia: SICA), as well as the overall average values. These empirical data can be extended by consulting the studies of Solis Cabrera (2008), Filippo et al (2011), the Annual Academic Report of the Complutense University of Madrid (*Vicerrectorado...*, 2012), and the FUTUR³ web portal.

The data shown in Figure 5 indicates that journal articles, in general, are the most abundant type of publication, but that it still does not exceed 40% of the total production; books and book chapters amount to 30%, and conference communications come to around 20%.

Figure 5 Scientific output according to document types in Spanish universities



Source: re-elaborated

Clearly, databases such as WoS (or MAS and Scopus), heavily skewed towards the journal article format, cannot be used exclusively to estimate the size of the academic Web by means of the Lincoln-Petersen method, since they don't

cover even 30% of the total scientific output of an institution, at least in the Spanish case (note that most of the journals or conferences where Spanish authors publish are not included in any of the aforementioned databases). Surely this same phenomenon occurs in other countries in a similar manner (always accounting for differences between university profiles).

Thus, the inferences from these databases, which are not representative of the entire collection of scientific literature (they are biased both by language and by document types), induce to not adequately estimate neither the size of the current academic literature in general nor the size of Google Scholar in particular.

4.2. Estimates from empirical data

In Table 1 we can observe the median, the geometric mean, and the number of studies that conforms the empirical set for each unit of study (number of documents and unique citing documents). The complete results obtained from the empirical data are available in Appendix I.

Table 1. Correction factor obtained from empirical studies

UNIT OF STUDY	MEDIAN	GEOMETRIC MEAN	N*
Number of documents	3	2.8	8
Unique citing documents	2.4	2.9	9

* The studies with less than 10 documents in the sample of WoS have not been finally considered since they are not representative enough.

We might assume (based on both empirical studies referenced in Appendix I and data provided in Table1) than at a general level (a round correction factor of 3), GS could triple the contents of WoS, although both databases would have a different English content distribution, and a different document type distribution. This is of particular importance as it is implying that the size comparison is not influenced by the biases of the database shown in the previous section.

Precisely, from the empirical studies, we test that the proportion of English documents in GS is around 65% (Appendix III). This would mean that, for documents in English, GS does not triple the number of WoS documents, but it probably does for documents in other languages. Knowing the general size correction factor, we don't need to worry about languages distribution for calculating estimates.

Therefore, the process of making inferences from samples of empirical studies previously conducted leads us in a simple way to multiply the size of WoS three times. As WoS has currently about 57 million records (see Figure 6), we may consider a quantity around 171 million records for GS.

These data and relationships can be expressed formally as follows (Table 2):

Table 2. Size relationships between Web of Science (WoS) and Google Scholar (GS)

EQUATION	OBSERVATIONS
$3 * WoS = GS$ [1]	We apply a correction factor of 3: GS triples WoS.
$WoS = WoSe*0.9 + WoSo*0.1$ [2]	WoSe: English contents in WoS WoSo: WoS content from other languages
$GS = GSe*0.65 + GSo*0.35$ [3]	GSe: English contents in GS (approx. 65%) GSo: GS content from other languages
$3 * (WoSe*0.9+WoSo*0.1) = GSe*0.65 + GSo*0.35$	Substituting [2] and [3] in [1]
WoS = 57 million documents; WoSe = 51.3 million documents;	We assume that WoS currently contains approximately 57 million records

Source: self-elaborated

With these data in hand, we obtain about 111.15 million documents in English, and applying the 65% of English documents in the estimate by Khabsa & Giles (99.8 million), we get a total of 153.5 million documents. This is an unexpected result because, as we discussed previously, we previously thought that Khabsa & Giles's method was overestimating the number of English documents in Google Scholar.

The estimates from empirical data, however, present some important shortcomings as well, since it is difficult to synthesize empirical results from:

- Different methods of sample selection and various sample sizes.
- Different topics: disciplines or specialties under study are varied. We must remember that communication patterns and dynamic publishing are very different among disciplines and this can seriously affect the results.
- Different periods in the samples: this is very important given the dynamic nature of the Web, and the changes to which it is subjected (uncontrolled creation, change, and deletion of documents).

4.3. Direct query to the database

The third strategy proposed in this working paper consists of asking directly the databases through their search interface, both globally (sectional query) and year by year (longitudinal query).

4.3.1. Sectional query (May 2014)

For a bibliographic database (such as WoS or Scopus), the query is simple and their results easily interpretable; however, in the case of academic search engines (such as MAS or especially GS), this procedure raises a number of unavoidable questions, for example:

- What do the search results obtained in a search engine like Google Scholar refer to? Should all records indexed in the database considered as unique documents?
- Can we trust, to some extent, the results it presents?

Regarding the first question:

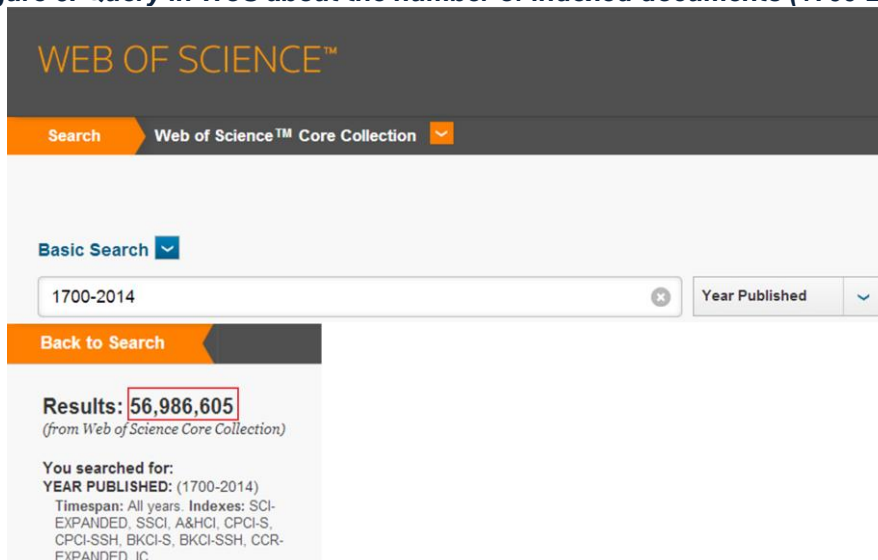
In the case of Google, the fact that there is no API for Google Scholar, and that Google only displays the first 1,000 results, prevents us from performing large scale empirical studies about the reliability and accuracy of the query commands (especially the "site" command). In the case of Microsoft, Bing Search also retrieves only the first 1,000 results (despite having an API). In the case of MAS (which also offers an API), its functionality is controlled and closer in nature to bibliographic databases, standing midway between a pure academic search engine (Google Scholar) and a pure bibliographic database (WoS).

Regarding the second question:

If we were still in 2005-2007, according to Jacsó (2005a; 2008; 2011) we should not trust these results, because at that moment there were paramount mistakes (mainly related to documents with wrong dates of publication and authorship, and duplicates due to not having correctly linked different versions). Today, the answer is probably yes, assuming an error rate that in GS could affect up to 10% of the results. The few empirical studies that have examined these errors have set this error rate below 10%, so our estimate of up to a 10% error rate is likely to be exaggerated. These errors are minimal in traditional bibliographic databases.

To illustrate these differences (and also obtain empirical data to work with), the queries that return the global coverage of both WoS and the academic search engines MAS and GS are offered below. In Figure 6, the query to WoS about the number of registered items from 1700-2014 is shown, obtaining a total of 56,980,000 records.

Figure 6. Query in WoS about the number of indexed documents (1700-2014)



Source: Web of Science v. 5.13.3

Nevertheless, if the query is performed for each type of document considered separately, the total differs slightly (Table 3), obtaining 59.5 million records (this is because a document may be classified in more than one type of document).

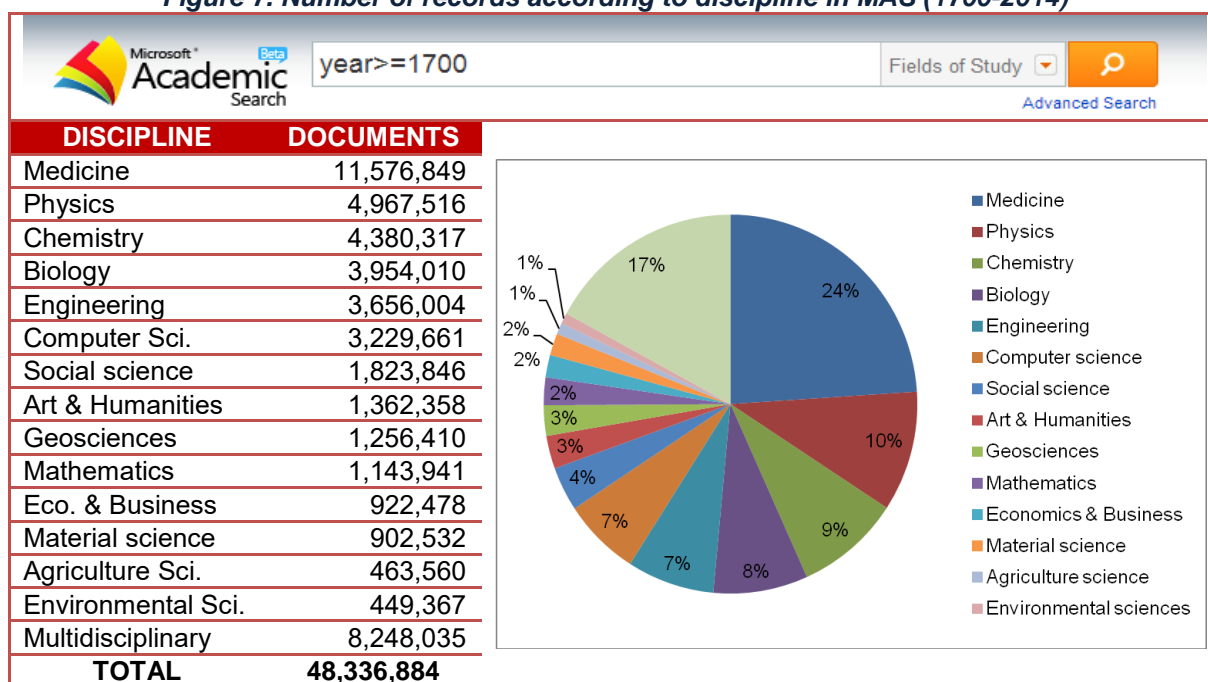
Table 3. Number of records according to document type (WoS, 1700-2014)

DOCUMENT TYPE	DOCUMENTS	DOCUMENT TYPE	DOCUMENTS
Article	33,876,866	Book	52,395
Meeting abstract	6,197,232	Fiction creative prose	45,357
Proceedings paper	6,089,411	Theater review	31,653
Book review	3,829,585	Dance performance review	21,911
Editorial material	2,192,341	Music score review	18,077
Letter	2,102,355	Reprint	16,093
Note	1,471,669	Software review	15,563
Review	1,219,612	Abstract of published item	13,434
Book chapter	684,105	Bibliography	11,952
News item	448,039	Excerpt	7,396
Poetry	247,393	Tv review radio review	6,881
Correction	163,682	Tv review radio review video	4,791
Correction addition	157,854	Script	2,686
Art exhibit review	104,853	Hardware review	2,540
Biographical Item	97,023	Database review	1,387
Item about an individual	92,399	Music score	1,240
Discussion	80,425	Chronology	1,210
Record REview	67,044	Main cite	13
Music performance review	61,449	Meeting summary	7
Film review	57,896	TOTAL	59,495,819

Source: Web of Science v. 5.13.3

Next, the direct query to MAS is performed (Figure 7) filtering for coverage from 1700 to the present (May 2014).

Figure 7. Number of records according to discipline in MAS (1700-2014)



Source: Microsoft Academic Search

There is a difference between the direct result (45,970,537 million documents) and that obtained from the summation of the different disciplines (48,336,884) due to the same reason as before: a document can be classified in several different disciplines at the same time.

In the case of Google Scholar, the database allows a temporal query (via custom range option). As with WoS and MAS, we perform a query from 1700 to 2013. Unfortunately, this procedure fails, returning only 596,000 documents. In Table 4, we show some other queries to demonstrate this malfunction.

Table 4. Malfunction of custom range option in Google Scholar

PERIOD	HCE
1700-2013	596,000
1750-2013	567,000
1800-2013	552,000
1850-2013	566,000
1900-2013	541,000
1950-2013	617,000
2000-2013	693,000

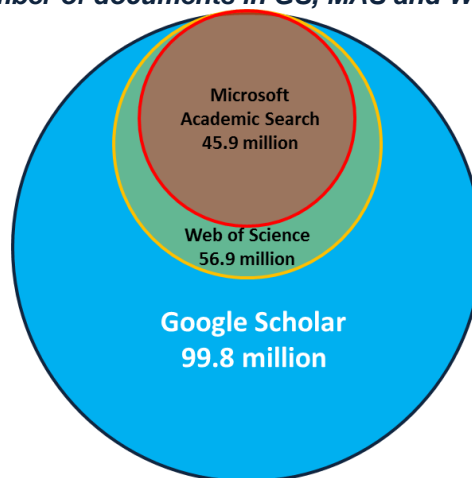
Source: self-elaborated

We can observe that the results displayed on Table 3 not only show a low number of results for such a wide timeframe, but also serious inconsistencies. In the time span “2000-2013” the system is retrieving more documents than in longer periods. However, if we execute the query introducing only 1 year in the custom range, the results seem to be more accurate. For example, for the year “1900”, we obtain 141,000 results and for the year “2000”, 2,410,000 results. Therefore, in order to solve this problem, a longitudinal analysis is required.

4.3.2. Longitudinal query

The sum of article records from 1700 to 2013 returns 99.8 million records in Google Scholar (59.8 million documents written in English). Comparative data of the three databases (WoS, MAS and GS) are shown in Figure 8.

Figure 8. Number of documents in GS, MAS and WoS (1700-2014)



Source: self-elaborated

Circles only represent relative size; intersections do not represent shared coverage

Next, longitudinal data of the 3 databases (GS, MAS and WoS) from 1800 to 2013 (Figures 8, 9 and 10) is offered⁴.

Figure 8. Google Scholar, Microsoft Academic Search and Web of Science (1800–1899)

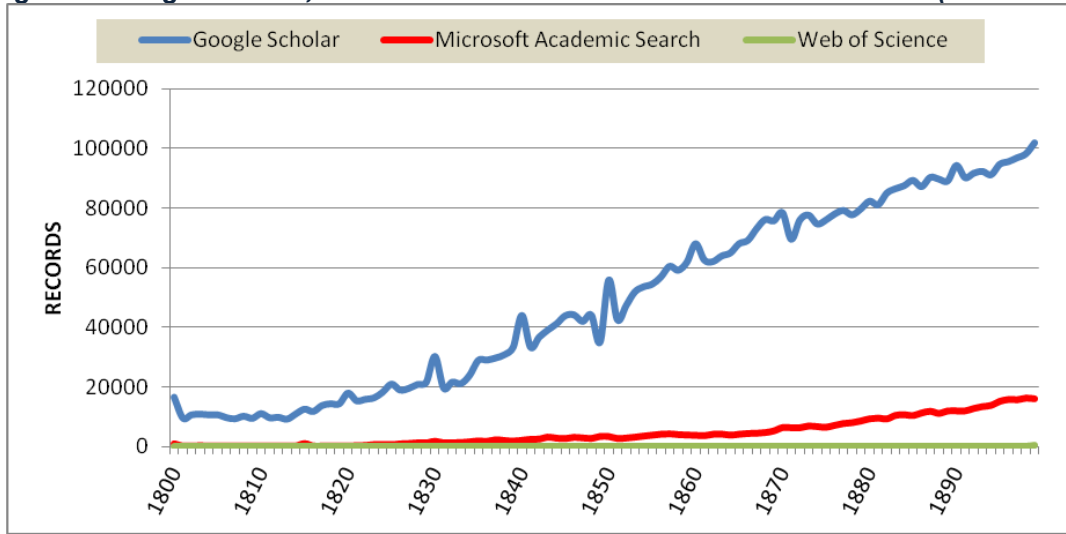


Figure 9. Google Scholar, Microsoft academic Search and Web of Science (1900–1949)

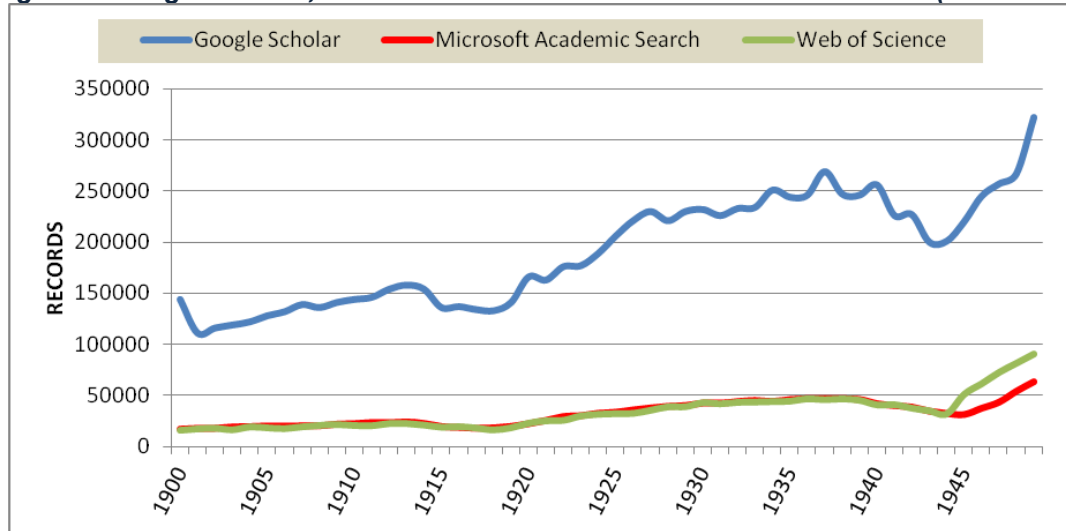
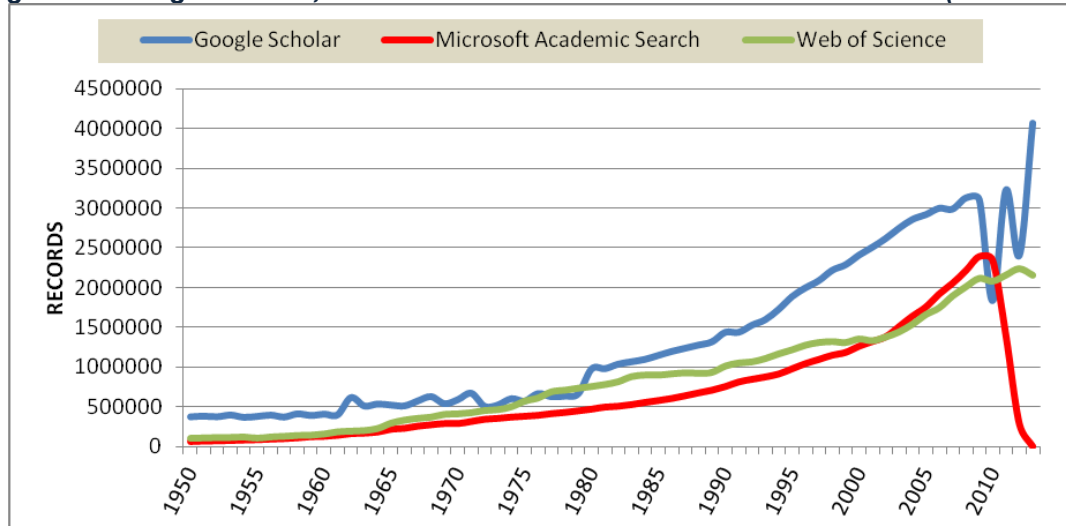


Figure 10. Google Scholar, Microsoft Academic Search and Web of Science (1950–2013)



On the one hand, these figures emphasize the primacy of Google Scholar practically during all this period (over 200 years), except in the 1970s, where its performance is similar to that provided by WoS.

The prevalence appears to accelerate again in the last decade of the twentieth century and the first years of the twenty-first, except for the problems identified in 2010 and 2011 (figure 10), probably due to internal changes within the search engine.

Indeed, the problems identified in recent years, very significant on the other hand (a fall of more than 1 million documents from 2009 to 2010, when world output actually accelerates) gives a good account of the dangerous instability of Google Scholar (Aguillo, 2011; Orduña-Malea & Delgado López-Cózar, 2014) and search engine hit count estimates (Jacsó, 2006).

This behavior does not occur (and it would not make sense if it did) on traditional bibliographic databases. The fact that the number of records decreases from year to year obviously does not mean less production in those years, but that the search engine has made internal adjustments, deleting duplicates, fixing bugs, among other technical issues.

On the other hand, these data highlight the similar sizes of WoS and MAS, exemplified in Figure 8, and are consistent with the data offered by Khabsa and Giles (2014). However, between these two products three important differences are emerging:

- MAS collects documents before 1900 unlike WoS (as this database states in its coverage policy).
- MAS drops down since 2010 (Orduña et al, 2014b).
- WoS is growing steadily; in the 1970 it even catches up with Google Scholar.

Finally, Table 5 offers the total count of records grouped by decades since 1950, for each database. Additionally, the relative size of MAS and WoS in relation to Google Scholar (in terms of global size and not on shared records) is offered as well.

For example, in the decade of 2001-2010, the size of WoS was almost two-thirds (62%) the size of Google Scholar, but on the 1970s WoS almost matched the size of Google Scholar (91%). In the case of MAS, 2001-2010 was when it got closest to the size of GS, even surpassing WoS, only to drop in 2010.

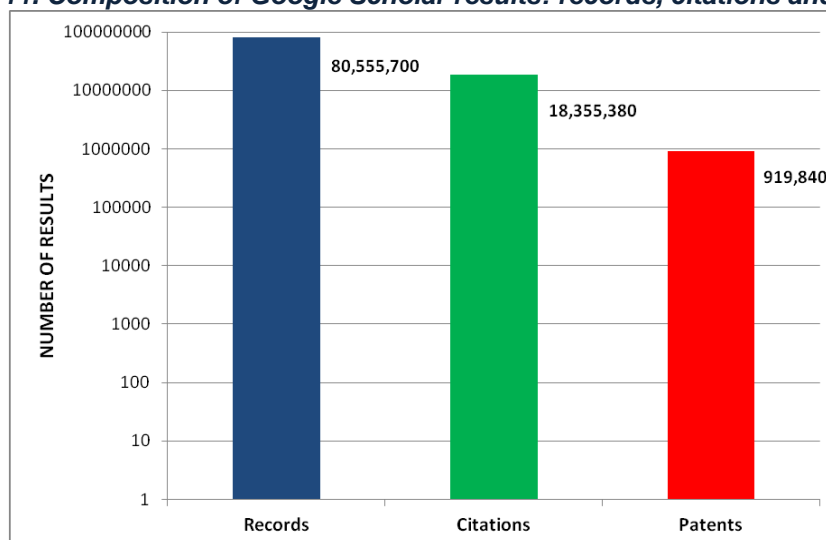
Table 5. Number of records by decade in each database (WoS, 1700-2014)

DECADE	GS	MAS	WoS	MAS	WoS
1951-1960	3,906,000	1,006,036	1,193,795	0.26	0.31
1961-1970	5,455,000	2,275,739	2,919,761	0.42	0.54
1971-1980	6,467,000	3,981,727	5,861,577	0.62	0.91
1981-1990	11,823,000	6,107,296	8,931,596	0.52	0.76
1991-2000	19,200,000	10,211,009	12,119,377	0.53	0.63
2001-2010	27,730,000	18,562,550	17,141,610	0.67	0.62
2011-2013	9,710,000	1,692,617	6,534,206	0.17	0.67

Source: self-elaborated

At this point it should be noted that the sum returning 99.8 million documents in Google Scholar includes both patents and citations. If we exclude these two types of documents from the query, the results fall dramatically to 80.5 million. In Figure 11, we present the results disaggregated by records (80.69%), citations (18.38%) and patents (0.92%), since 1700.

Figure 11. Composition of Google Scholar results: records, citations and patents



Source: self-elaborated

However, these results should be taken with caution, because the hit count estimates in Google Scholar for these queries are not accurate. For the 314 years calculated (from 1700 to 2013), we found inconsistencies up to 122 times between the complete query (records + citations + patents) and the query excluding patents (records + citations), finding more results with the latter than with the former. In short: excluding patents, the system sometimes retrieved more results than in the original complete query.

In Table 6 we show some years where these inconsistencies are present. As is apparent, the problem is accentuated in the last 15 years. Due to the order of magnitude of the results (millions of results since the end of the 20th century), error rates reach unsustainable values, especially if the aim is to make checksums per year, as showed previously in Figure 11.

Table 6. Inconsistencies in Google Scholar queries for patents and citations

YEAR	RECORDS+ CITATIONS+ PATENTS	RECORDS+ CITATIONS	DIFFERENCE
2013	4,070,000	4,150,000	-80,000
2010	1,840,000	2,020,000	-180,000
2009	3,110,000	3,230,000	-120,000
2007	2,990,000	3,110,000	-120,000
2006	3,000,000	3,050,000	-50,000
2005	2,920,000	2,950,000	-30,000
2004	2,860,000	2,930,000	-70,000
2002	2,620,000	2,720,000	-100,000
2000	2,410,000	2,550,000	-140,000

Source: self-elaborated

As regards citations, the total figure obtained (18,355,380 citations) is higher than we expected. The accuracy is better than that obtained for patents, presenting only 8 errors out of 314 years, and focused in a narrow time span of 20 years: 1969 (difference of -59,000 records), 1970 (-36,000), 1971 (-52,000), 1975 (-19,000), 1976 (-11,000), 1978 (-56,000), 1982 (-10,000) and 1988 (-40,000).

Citations present an additional problem: not all citations are really records Google Scholar hasn't been able to find on the web. In some cases, the same article appears as a record and a citation at the same time. For example, Figure 12 shows a query corresponding to the topic "H Index Scholar". Google Scholar retrieves the same article twice. The first result is considered as a citation, and the second as a regular record. Worst of all, both results count in the global Hit Count Estimate.

Figure 12. Duplicity in citations on Google Scholar

"H index Scholar" [Search]

4 results (0.07 sec)

[CITATION] **H Index Scholar**: the h-index for Spanish public universities' professors of humanities and social sciences
[E Delgado-Lopez-Cozar... - ..., 2014 - EPI APARTADO 32 280, ...](#)
Cite Save More

H Index Scholar: el índice h de los profesores de las universidades públicas españolas en [humanidades y ciencias sociales](#)
[E Delgado-López-Cózar... - El profesional ..., 2014 - elprofesionaldeinformacion. ...](#)
The **H-Index Scholar** is a bibliometric index that measures the productivity and scientific impact of the academic production in humanities and social sciences by professors and researchers at public Spanish universities. The methodology consisted of counting their ...
All 4 versions Cite Save More

Source: Google Scholar

Finally, Google Scholar includes one more type of document apart from the "Articles" category (which is composed of records, citations, and patents): these are the case laws from the Supreme Court of the United States of America, which also include citations (Figure 13).

Figure 13. Case laws and citations on Google Scholar

Scholar About 282,000 results (0.02 sec)

Articles Case law Federal courts California courts Select courts... My library

Any time Since 2014 Since 2013 Since 2010 Custom range... 2013 - 2013 Search

Sort by relevance Sort by date

include citations

[Alleyne v. US](#)
133 S. Ct. 2151, 570 US __, 186 L. Ed. 2d 314 - Supreme Court, 2013 - Google Scholar
Harris drew a distinction between facts that increase the statutory maximum and facts that increase only the mandatory minimum. We conclude that this distinction is inconsistent with our decision in *Apprendi v. New Jersey*, 530 US 466, 120 S.Ct. 2348, 147 L.Ed.2d 435 (2000), and ...
Cited by 1277 How cited Related articles Cite Save

[UNIV. OF TEX. SOUTHWESTERN MED. v. Nassar](#)
133 S. Ct. 2517, 570 US __, 186 L. Ed. 2d 503 - Supreme Court, 2013 - Google Scholar
When the law grants persons the right to compensation for injury from wrongful conduct, there must be some demonstrated connection, some link, between the injury sustained and the wrong alleged. The requisite relation between prohibited conduct and compensable injury is ...
Cited by 575 How cited Related articles All 2 versions Cite Save

[US v. Windsor](#)
133 S. Ct. 2675, 570 US 12, 186 L. Ed. 2d 808 - Supreme Court, 2013 - Google Scholar
Two women then resident in New York were married in a lawful ceremony in Ontario, Canada, in 2007. Edith Windsor and Thea Spyer returned to their home in New York City. When Spyer died in 2009, she left her entire estate to Windsor. Windsor sought to claim the ...
Cited by 570 How cited Related articles Cite Save

[McQuiggin v. Perkins](#)
133 S. Ct. 1924, 569 US __, 185 L. Ed. 2d 1019 - Supreme Court, 2013 - Google Scholar
This case concerns the "actual innocence" gateway to federal habeas review applied in *Schlup v. Delo*, 513 US 298, 115 S.Ct. 851, 130 L.Ed.2d 808 (1995), and further explained in *House v. Bell*, 547 US 518, 126 S.Ct. 2064, 165 L.Ed.2d 1 (2006). In those cases, a convincing ...
Cited by 574 How cited Related articles Cite Save

[Chaldez v. US](#)
133 S. Ct. 1103, 185 L. Ed. 2d 149, 568 US __ - Supreme Court, 2013 - Google Scholar

Source: Google Scholar

A longitudinal analysis of the number of case laws from 1700 has been performed, in a similar way as for the Articles, both including and excluding citations in the search results. In Figure 14 we can observe the evolution since 1800 (from 1700 to 1799 Google Scholar retrieves only 408 documents).

Figure 14. Number of Case laws per year (1800-2013)



Source: Google Scholar

A total of 26,510,689 case laws (31.3%) and citations to case laws (68.7%) have been obtained from 1700. We should highlight the great differences between the number of case laws, and the number of citations to case laws, during the second half of the eighteenth century and the first half of the nineteenth. After that, and until 2013, both types share a very similar behaviour.

If case laws and their citations are included in the calculation of Google Scholar's total size, the global figure obtained raises to 126,341,609 million documents, a figure about twice the size of WoS (56.9 million).

However, this method also has a number of limitations to consider:

- Errors in rounding results performed by the search engine (we cannot forget that the hit count estimates are, as their name suggests, an estimate).
- The influence of the number of versions (both for records and citations) in the queries.
- Although Articles and Case laws appear as separate categories, are the citations to each one independent? A record marked as citation to Article may be included as a citation to Case law (theoretically, one document can cite both articles and case laws in the reference section).

These shortcomings lead us to consider a priori that this method is probably providing inflated results due to duplicates, versions and upward estimates.

Nevertheless, the figures obtained are unexpectedly lower than those obtained by methods 1 and 2.

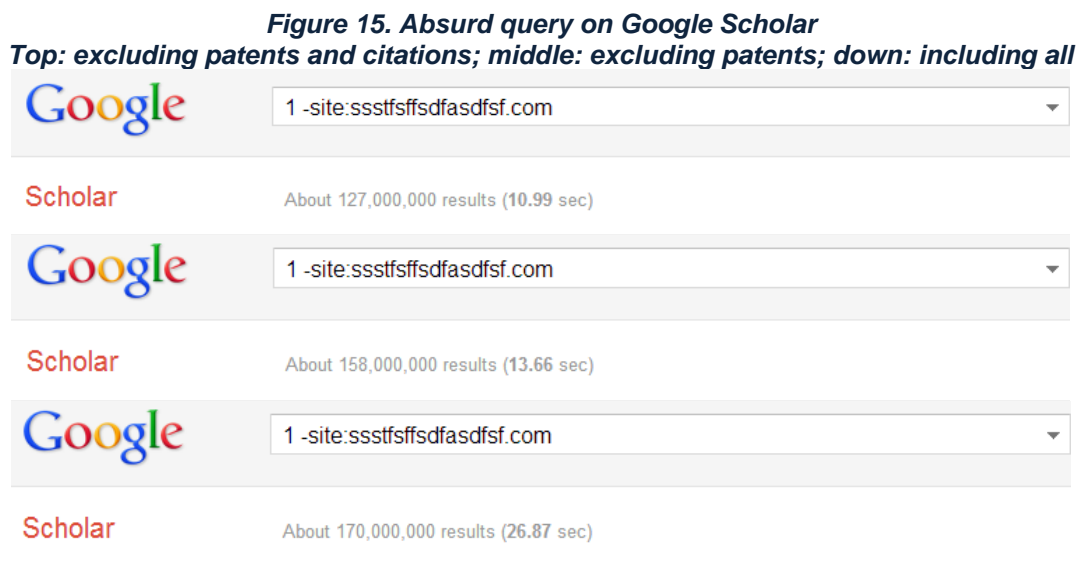
4.4. Absurd query

Lastly, the fourth method tested consists of applying an absurd query with the purpose of obtaining a hit count estimate about the total size of Google Scholar.

This method is applied under three different approaches: without temporal filter, using the custom range from 1700 to 2013, and finally by means of a longitudinal analysis (year by year).

a) Without temporal filter

The query <1 -site:ssstfsffsdfasdfs.com> is applied to Google Scholar's Articles category (Figure 15), excluding patents and citations (127,000,000 results), excluding only patents (158,000,000), and including patents and citations (170,000,000).

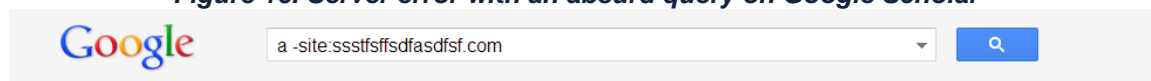


Source: Google Scholar

This same query, applied for Case laws (with citations) returns 4,550,000 results, far from the 26.5 million obtained through the longitudinal analysis discussed in the previous method.

The query <a -site:ssstfsffsdfasdfs.com> has been tested as well, obtaining 102,000,000 (excluding patents and citations) and 154,000,000 (excluding only patents). When trying to include patents and citations (theoretically the query with a higher count), an error message appears informing about technical problems to deliver results (Figure 16).

Figure 16. Server error with an absurd query on Google Scholar



Source: Google Scholar

b) custom range (1700 to 2013)

In this case, the query <1 -site:ssstfsffsdffasdfs.com> retrieves 176 million results for Articles and 4.3 million for Case laws. As regards the query <a -site:ssstfsffsdffasdfs.com>, this time it works returning 160 million articles and 6.8 million case laws..

c) Longitudinal analysis (1700 to 2013)

Finally, the absurd query has been performed for each year from 1700 to 2013. The query <1 -site:ssstfsffsdffasdfs.com> has been selected to do this since it retrieves more results with the procedure “b” than with another query.

The final summation gives an overall of 169.5 million articles and 3.4 million Case laws. These results, although they present different results from those obtained from the longitudinal analysis carried out with the null query used in section 4.3.2, they completely correlate. Pearson correlation (r) for Articles is r = .93 and for Case laws r = .71.

This confirms that Hit Count Estimates from Google Scholar are not useful to get an accurate performance for individual queries, but are useful to make performance comparison and relations.

In Table 7 we summarize the figures obtained from methods 3 and 4, both for Articles and Case laws, with the 3 procedures employed (total query, using the custom range for years, and querying year by year).

Table 7. Summary of data obtained by the methods of consultation (empty and absurd)

ARTICLES			
Absurd query		Empty query	
Procedure	HCE	Procedure	HCE
Total	170,000,000	Total	--
Longitudinal	169,526,760	Longitudinal	99,830,920
Time span	176,000,000	Time span	596,000

CASE LAWS			
Absurd query		Empty query	
Procedure	HCE	Procedure	HCE
Total	4,550,000	Total	--
Longitudinal	3,422,823	Longitudinal	26,510,689
Time span	4,340,000	Time span	629,000

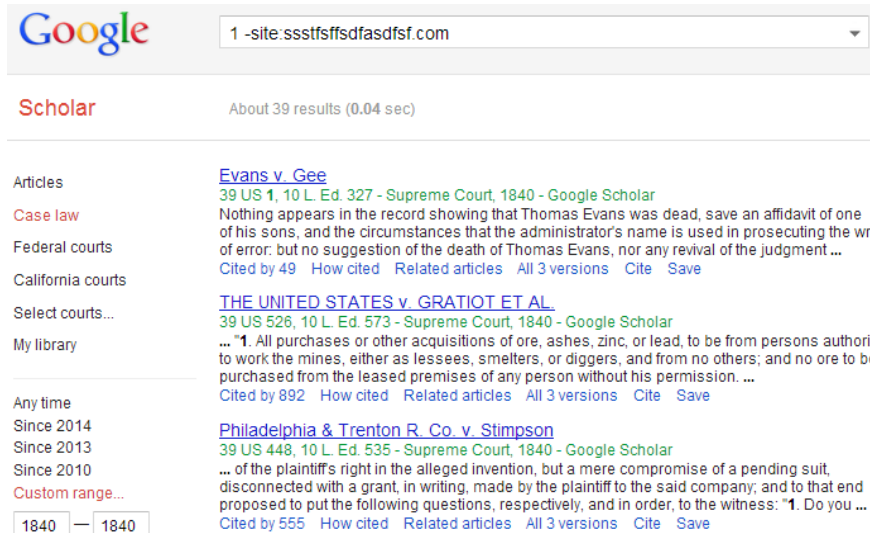
Source: self-elaborated

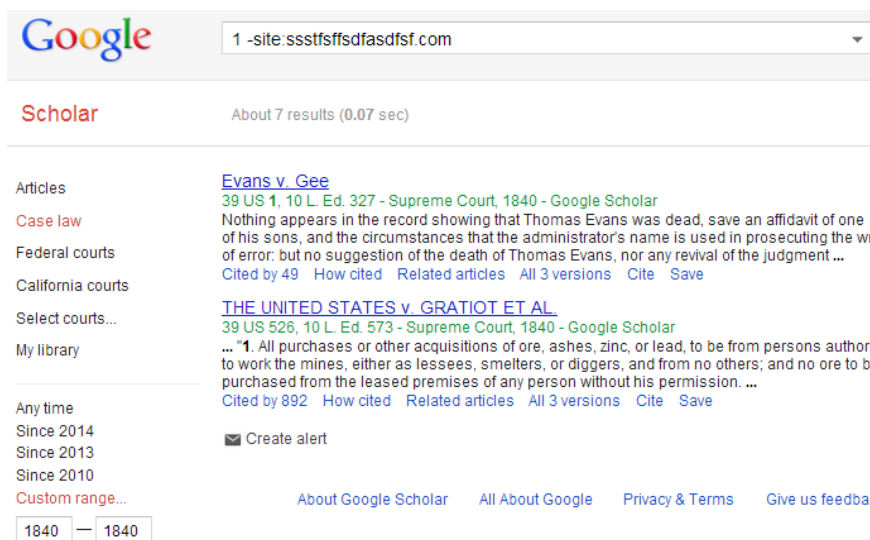
We can observe some similarities in the results obtained with the absurd query, both for Articles and Case laws, and independently of the procedure. However, the empty query generates results that are not consistent with those obtained in the previous method (direct query), especially in the case of the longitudinal approach for Articles.

In order to know the reason for the differences between these two methods (both based on querying directly the database), we proceeded to analyse the search results that the absurd query is generating more precisely. Thus we have identified the following weaknesses:

- a) The absurd query does not retrieve citations, independently of if this option is checked or not in the search options, both in the case of Articles and Case laws. Conversely, the empty query retrieves citations, as was noted in section 4.3.2. This may explain the differences between these queries in the longitudinal results for Case laws.
- b) Hit Count Estimates present serious inconsistencies in the activation / deactivation of the citation inclusion feature. In Figure 17 we display an example of this shortcoming. In the upper figure we display an example of the absurd query, filtered by the year 1840, with the option “Citations” deactivated, which obtains 39 results. At the same time, the bottom figure shows the same query but activating the “Citations” option. As we can see the results obtained in this case are only 9 (although theoretically we should have obtained at least the same as in the other query, or more). Moreover, the system only retrieves 2 documents even though the HCE says there are 9. Although this example deals with Case laws, it is true for Articles as well.

Figure 17. Hit count estimates inconsistencies in the activation/deactivation of citations

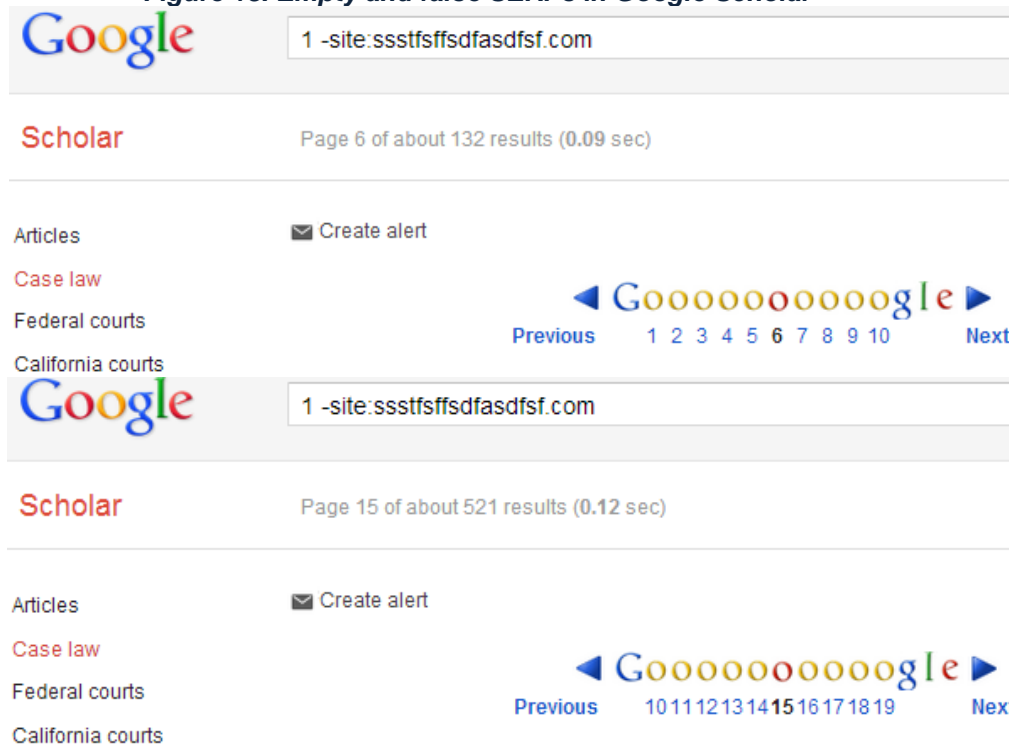




Source: Google Scholar

Lastly, we have verified the existence of empty and false SERPs (Search Engine Results Pages). In Figure 18 we display an example of an application of the absurd query, for any given year. In the upper figure we can observe that the system retrieves a Hit Count Estimate of 132, but the 6th SERP is empty. If we set the system to retrieve up to 20 results per SERP (the maximum allowed in GS), the 6th SERP should show results 101th to 120th, and never an empty result. Moreover, if we click on the 15th SERP (bottom figure), the system not only still retrieves an empty SERP but the HCE increases as well (521). It should be noted that the longitudinal analysis has been performed using the first SERP.

Figure 18. Empty and false SERPs in Google Scholar



Source: Google Scholar

Since this method of absurd query is more accurate than it seems at first because the search engine is forced to check the entire database to answer the query, as the time responses are suggesting (See Figure 15), these shortcomings are, as of yet, unexplained.

It is possible that the system goes into a loop when trying to answer a query of this type, but it is more surprising that the final figures provided seem logical and coherent, and close to those achieved by other methods, unlike what happens with the empty query method.

5. CONCLUSIONS

The main objective of this study was to measure the size of Google Scholar via four different techniques, discussing the advantages and disadvantages of each method and the disparity of results. Table 8 summarizes the results obtained for each method.

Table 8. Summary of Google Scholar size estimates

METHOD	GS SIZE ESTIMATE	COMMENTS
A. Khabsa & Giles	152.7 million	The authors estimate 99.3 million for English contents without patents. If we consider that English constitutes approximately 65% of contents we obtain this figure.
B. Empirical data	171 million	This method assumes GS triples the size of WoS
C.1. Empty query (custom range)	1.2 million	This method applies an empty query in setting the custom range from 1700 to 2013.
C.2. Empty query (longitudinal)	126.3 million	This method applies an empty query in a longitudinal analysis from 1700 to 2013 including Articles (99.8 millions) and Case laws (26.5 millions).
D.1. Absurd query (total)	174.5 million	170 million Articles and 4.5 million Case laws
D.2. Absurd query (custom range)	176.8 million	This method applies an absurd query setting the custom range from 1700 to 2014, both for Articles (170 millions) and Case laws (6.79 millions)
D.3. Absurd query (longitudinal)	172.9 million	This method applies an absurd query in a longitudinal analysis from 1700 to 2013, obtaining 169.5 million Articles and 3.4 million Case laws.

Source: self-elaborated

Method A (taking apart the identified problems about the use of the Lincoln-Petersen method, the academic size outside Google Scholar or the biased sample) poses a priori some methodological problems, because it parts from a false axiom: inferences should not be made from the comparison of databases with very different characteristics.

The database usually employed to make comparisons is the Web of Science, but it has, among others, two major flaws: the documents it indexes are mostly written in English, and most of them are journal articles.

The well-known problems of WoS can also be applied to MAS, from which the authors extract the sample. The main concerns in this sense are the following:

- a) The “Citing documents” can be in all languages. The authors identify that 98% of citing papers in MAS are written in English (a correction is

applied instead of eliminating non-English documents), but this percentage in Google Scholar must be lower, and it is not indicated in the estimation of the size of Google Scholar. Probably all documents written in languages other than English should have been avoided if the estimation of just the “English academic web” was the objective.

- b) The sample is composed by articles, whereas citing documents are diverse in their typology. If the target of this research had been calculating the number of “articles” included on Google Scholar, then this procedure would have been appropriate (after the elimination of citing documents other than articles). If the purpose was to calculate the size of the entire database, a sample composed uniquely by articles is not representative of Google Scholar.

For this reason, we believe that on the one hand, the calculation is oversized with respect to the language (the 99.3 millions obtained cover more than English Article contents), and undersized with respect to the document types. Moreover, patents and case laws (integrated in the Google Scholar database) should be added in the final count.

Nevertheless, the research design performed by the authors is novel and brilliant. It is based on the gathering of citing documents to a sample of cited articles. This procedure has several advantages (such as that the search engine is forced to query its entire database to find all documents that match with a citation to any of the documents of the sample). Nonetheless, there are types of documents on Google Scholar (such as syllabi, conferences, teaching material, etc.) that maybe do not cite any other document (and they will never be cited as well), that are not considered in this method.

Method B (making estimates based on the comparative differences between databases), provides an approximate size for Google Scholar of 171 million records. This method has the advantage of not being affected by comparisons with other biased databases, although the estimate is very rough and imprecise as it is synthesized from diverse empirical results. Despite this, the results obtained are close to those obtained in the method A.

A priori, method C seemed to be more accurate, being based on direct querying of Google Scholar, but the results are unexpectedly more distant from those of other methods (Case laws are overrepresented and Articles are underrepresented), especially method C.1, which is discarded completely.

In the case of method C.2 (longitudinal query), the problems stem from the following issues:

- Lack of precision: the extent to which the search engine returns data for our query that corresponds with the reality of its cataloged universe. We rely on hit count estimates (Google explicitly states “about xxx results”), which are affected by unknown rounding routines.

- Lack of reliability: the extent to which the search engine returns, from our query, what we really want to measure, i.e., the number of unique records indexed in Google Scholar.

In that sense, if we consider up to a 10% of errors (related with cataloging issues, duplicate records, etc.), and errors related with the accuracy and reliability, the actual size of Google Scholar may be even lower according to this method, around 114 million records.

These results are completely unexpected, especially because when using this method we are including patents and the Case law category (with their corresponding citations). Both types of documents are excluded from both the methods A and B. Therefore we expected higher results, not the opposite.

It is probable, however, that:

- Method A gives inflated results if we assume that the 99.3 million gathered are including not only English documents, and not only article-type documents.
- Method B gives inflated results for Google Scholar because these comparisons focus on periods where the supremacy of Google Scholar was higher. This is because GS may triple WoS if we consider the whole timeframe, but there are periods in which this is not true (Figures 8-10).

Finally, method D, though simple, produces similar results to those obtained by methods A and B. This procedure, though it is closer to method C, presents results which are significantly different to those of method C. The Hit Count Estimates, which skew both methods, may influence each method differently, or it is also possible that the absurd query does not work properly for some reason.

In this sense, we have checked that the absurd query does not retrieve citations (independently of if this option is checked or not in the search options), and it also creates empty and false SERPs. Moreover there is a dysfunction between the inclusion of Citations and Patents and the HCE obtained (in this case both for the Articles and Case law categories, and both for empty queries and absurd queries).

The underlying questions here are the following: what should we have to doubt more?, the reliability of Google Scholar's HCEs (assuming 10% error in the results)?, the estimate from MAS using the Lincoln-Petersen method?, or the estimate based on a correction factor?

Most surprising of all is that even though all methods seem invalid for various and diverse reasons, all return similar results (except Method C). Probably a sensible estimate, observing all the results obtained and taking into account a 10% of possible internal errors, would be a total of around 160 million unique records (without considering different versions for the same record).

The information policies of GS ("no comments" is the house brand) encourage speculation and force researchers to make conjectures about the real size of this database (and its entrails). A figure that otherwise surely would be easy to determine by their technical and manager staff simply by pressing a key on their computer at the office.

Logically this matter should be solved by simply asking this information to Google, and to the people directly responsible for Google Scholar. Their answer would avoid all our concerns, efforts and resources dedicated to finding this sort of "golden fleece" that this issue has become.

Although it seems impossible that Google will publish this information (at least on a short term), we wonder if anyone can "press the button" and tell us what the size of Google Scholar is. Perhaps even Google Scholar does not know this "number"... a number that approximately represents the online scientific heritage circulating at present.

Notes

1. The option Custom range appears after a query is submitted, in the search box of Google Scholar (not before). Moreover, we can execute this query directly on the browser via http as well. Once we obtain the first results via hit count estimates, we can generate new queries without introducing any keyword in the search box, and only selecting the time span required.
2. <https://www.worldcat.org>
3. <http://futur.upc.edu>
4. Web of Science does not provide data until 1898.

References

- Abdullah, A.; Thelwall, M. (2014). "Can the Impact of non-Western Academic Books be Measured? An investigation of Google Books and Google Scholar for Malaysia". *Journal of the Association for Information Science and Technology*.
- Adriaanse, L.S.; Rensleigh, C. (2011). "Comparing Web of Science, Scopus and Google Scholar from an Environmental Sciences perspective". *South African Journal of Library Information Science*, 77(2), 169–178.
- Aguillo, I.F. (2011). "Is Google Scholar useful for bibliometrics? A webometric analysis". *Scientometrics*, 91(2), 343–351.
- Amara, N.; Landry, R. (2012). "Counting citations in the field of business and management: why use Google Scholar rather than the Web of Science". *Scientometrics*, 93(3), 553–581.
- Bakkalbasi, N.; Bauer, K.; Glover, J.; Wang, L. (2006). "Three options for citation tracking: Google Scholar, Scopus and Web of Science". *Biomedical digital libraries*, 3(1), 7.
- Bar-Ilan, J. (2007). "Which h-index? — A comparison of WoS, Scopus and Google Scholar". *Scientometrics*, 74(2), 257–271.
- Bar-Ilan, J. (2010). "Citations to the "Introduction to informetrics" indexed by WOS, Scopus and Google Scholar". *Scientometrics*, 82(3), 495–506.
- Bornmann, L.; Marx, W.; Schier, H.; Rahm, E.; Thor, A.; Daniel, H. D. (2009). "Convergent validity of bibliometric Google Scholar data in the field of chemistry—Citation counts for papers that were accepted by *Angewandte Chemie International Edition* or rejected but published elsewhere, using Google Scholar, Science Citation Index, Sc.". *Journal of Informetrics*, 3(1), 27–35.

- Cabezas-Clavijo, A.; Delgado López-Cózar, E. (2013). "Google Scholar and the h-index in biomedicine: the popularization of bibliometric assessment". *Medicina intensiva*, 37(5), 343–354.
- Cardenas, J.; Udo, G.J. (2013). "Knowledge Management Literature Trends: an ISI Web of Science and Google Scholar comparison". In: *Proceedings of the 44th Annual Meeting of the Decision Sciences Institute*. Baltimore, November 16-19.
- De Groote, S.L.; Raszewski, R. (2012). "Coverage of Google Scholar, Scopus, and Web of Science: A case study of the h-index in nursing". *Nursing Outlook*, 60(6), 391-400.
- Delgado López-Cózar, E.; Repiso, R. (2013). "The Impact of Scientific Journals of Communication: Comparing Google Scholar Metrics, Web of Science and Scopus". *Comunicar*, 41, 45-52.
- Dobra A.; Fienberg, S.E. (2004) "How large is the world wide web". *Web Dynamics*, 23–44.
- Filippo, D. de; Sanz-Casado, E.; Urbano, C.; Ardanuy, J.; Gómez-Caridad, I. (2011). "El papel de las bases de datos institucionales en el análisis de la actividad científica de las universidades". *Revista española de documentación científica*, 34(2), 165-189.
- Franceschet, M. (2009). "A comparison of bibliometric indicators for computer science scholars and journals on Web of Science and Google Scholar". *Scientometrics*, 83(1), 243–258.
- Gil Roales-Nieto, J.; O'Neill, B. (2012). "A Comparative Study of Journals Quality based on Web of Science , Scopus and Google Scholar : A Case Study with IJP & PT". *International Journal of Psychology and Psychological Therapy*, 12(3), 453–479.
- Haley, M.R. (2014). "Ranking top economics and finance journals using Microsoft academic search versus Google scholar: How does the new publish or perish option compare?". *Journal of the Association for Information Science and Technology*, 65(5), 1079–1084.
- Harzing, A.W.; van der Wal, R. (2009). "A Google Scholar h-index for journals: An alternative metric to measure journal impact in economics and business". *Journal of the American Society for Information Science and Technology*, 60(1), 41-46.
- Huh, S. (2013). "Citation Analysis of the Korean Journal of Urology From Web of Science, Scopus, Korean Medical Citation Index, KoreaMed Synapse, and Google Scholar". *Korean journal of urology*, 54(4), 220–228.
- Jaćimović, J.; Petrović, R.; Živković, S. (2010). "A citation analysis of Serbian Dental Journal using Web of Science, Scopus and Google Scholar". *Stomatoloski glasnik Srbije*, 57(4), 201-211.
- Jacobs, J.A. (2009). "Where Credit Is Due: Assessing the Visibility of Articles Published in Gender & Society with Google Scholar". *Gender & Society*, 23(6), 817–832.
- Jacsó, P. (2005a). "Google Scholar: the pros and the cons". *Online information review*, 29(2), 208-214.
- Jacsó, P. (2005b). "Comparison and analysis of the citedness scores in Web of Science and Google Scholar". In *Digital libraries: Implementing strategies and sharing experiences*, 360-369.
- Jacsó, P. (2006). "Dubious hit counts and cuckoo's eggs". *Online Information Review*, 30(2), 188-193.
- Jacsó, P. (2008). "Google scholar revisited". *Online information review*, 32(1), 102-114.
- Jacsó, P. (2011). "The pros and cons of Microsoft Academic Search from a bibliometric perspective". *Online Information Review* 35(6), 983-997.
- Khabsa, M.; Giles, C.L. (2014). "The Number of Scholarly Documents on the Public Web". *Plos One*, 9(5).
- Koehler, Wallace (2004). "A longitudinal study of Web pages continued a consideration of document persistence". *Information research*, 9(2) .
<http://informationr.net/ir/9-2/paper174.html>

- Kousha, K.; Thelwall, M. (2008). "Sources of Google Scholar citations outside the Science Citation Index: A comparison between four science disciplines". *Scientometrics*, 74(2), 273–294.
- Kousha, K.; Thelwall, M.; Rezaie, S. (2011). "Assessing the citation impact of books: The role of Google Books, Google Scholar, and Scopus". *Journal of the American Society for Information Science*, 62(11), 2147–2164.
- Kulkarni, A.V.; Aziz, B.; Shams, I.; Busse, J.W. (2009). "Comparisons of citations in Web of Science, Scopus, and Google Scholar for articles published in general medical journals". *Jama*, 302(10), 1092-1096.
- Lasda Bergman, E.M., (2012). "Finding Citations to Social Work Literature: The Relative Benefits of Using Web of Science, Scopus, or Google Scholar". *Journal of Academic Librarianship*, 38(6), 370–379.
- Lawrence, S.; Giles, C. (1998). "Searching the world wide web". *Science*, 280, pp. 98–100.
- Lawrence, S.; Giles, C. (1999). "Accessibility of information on the web". *Nature*, 400, pp. 107–9.
- Martell, C. (2009). "A Citation Analysis of College & Research Libraries Comparing Yahoo, Google, Google Scholar, and ISI Web of Knowledge with Implications for Promotion and Tenure". *College Research Libraries*, 70(5), 460–472.
- Meho, L.I.; Yang, K. (2007). "Impact of data sources on citation counts and rankings of LIS faculty: Web of Science versus Scopus and Google Scholar". *Journal of the American Society for Information Science and Technology*, 58(13), 2105–2125.
- Mikki, S. (2009). "Comparing Google Scholar and ISI Web of Science for Earth Sciences". *Scientometrics*, 82(2), 321–331.
- Mingers, J.; Lipitakis, E.A.E.C.G. (2010). "Counting the citations: a comparison of Web of Science and Google Scholar in the field of business and management". *Scientometrics*, 85(2), 613–625.
- Miri, S.M.; Raoofi, A.; Heidari, Z. (2012). "Citation Analysis of Hepatitis Monthly by Journal Citation Report (ISI), Google Scholar, and Scopus". *Hepatitis Monthly*, 12(9).
- Moskovkin, V.M. (2009). "The potential of using the Google Scholar search engine for estimating the publication activities of universities". *Scientific and Technical Information Processing*, 36(4), 198–202.
- Noruzi, A. (2005). "Google Scholar: The New Generation of Citation Indexes". *Libri*, 55(4), 170-180.
- Onyancha, O.B.; Ocholla, D.N. (2009). "Assessing researchers' performance in developing countries: is Google scholar an alternative?". *Mousaion*, 27(1), 43–64.
- Orduña-Malea, E.; Delgado López-Cózar, E. (2014a). "Google Scholar Metrics evolution: an analysis according to languages". *Scientometrics*, 98(3), 2353–2367.
- Orduña-Malea, E.; Delgado López-Cózar, E. (2014b). The dark side of Open Access in GoogleScholar: the case of Latin-American repositories. *Scientometrics* (in press)
- Orduña-Malea, E.; Ayllón, J.M.; Martín-Martín, A.; Delgado López-Cózar, E. (2014a). "How many academic documents are visible and freely available on the Web?". *EC3 Google Scholar Digest Reviews*, n. 1.
- Orduña-Malea, E.; Ayllón, J.M.; Martín-Martín, A.; Delgado López-Cózar, E. (2014b). *Empirical Evidences in Citation-based search engines: is Microsoft Academic Search dead?* Granada: EC3 Reports, 16.
- Ortega, J.L. (2014). *Academic search engines: a quantitative outlook*. Netherlands: Elsevier [Chandos Information Professional Series].
- Ortega, J.L.; Aguillo, I.F. (2014). "Microsoft academic search and Google scholar citations: Comparative analysis of author profiles". *Journal of the Association for Information Journal of the Association for Information Science and Technology*, 65(6), 1149–1156.
- Salisbury, L.; Tekawade, A. (2006). "Where is agricultural economics and agribusiness research information published and indexed? A comparison of coverage in Web of

- Knowledge, CAB Abstracts, Econlit, and Google Scholar". *Journal of agricultural & food information*, 7(2-3), 125-143.
- Šember, M.; Utrobičić, A.; Petrak, J. (2010). "Croatian Medical Journal Citation Score in Web of Science, Scopus, and Google Scholar". *Croatian Medical Journal*, 51(2), 99–103.
- Solis Cabrera, F.M. (2008). "El sistema de información científica de Andalucía, una experiencia pionera en España". *Madrid+d*, 22 [Las comunidades autónomas frente a la I+D+i]. Junta de Andalucía, 13-18.
- Reina Leal, L.M.; Repiso, R.; Delgado López-Cózar, E. (2013). H Index of scientific Nursing journals according to Google Scholar Metrics (2007-2011). EC3 Reports, 5. *Vicerrectorado de investigación de la Universidad Complutense de Madrid. Memoria 2012*. Madrid: Servicio de investigación UCM. Available at: <https://www.ucm.es/data/cont/docs/3-2014-01-09-2.3.%20Vdo.%20Investigación.pdf>
- Winter, J.C.F.; Zadpoor, A.; Dodou, D. (2014). "The expansion of Google Scholar versus Web of Science: a longitudinal study". *Scientometrics*, 98(2), 1547–1565.
- Yang, K.; Meho, L.I. (2006). "Citation Analysis: A Comparison of Google Scholar, Scopus, and Web of Science". In *Proceedings of the American Society for Information Science and Technology*, 43(1), 1–15.
- Zarifm Mahmoudi, L.; Kianifar, H.R.; Sadeghi, R. (2013). "Citation analysis of Iranian Journal of Basic Medical Sciences in ISI web of knowledge, Scopus, and Google Scholar". *Iranian Journal of Basic Medical Sciences*, 16(10), 1027–1030.
- Zarifm Mahmoudi, L.; Sadeghi, R. (2012). "Citation analysis of Iranian journal of nuclear medicine: Comparison of SCOPUS and Google scholar". *Iranian Journal of Nuclear Medicine*, 20(2), 1–7.

APPENDIX I. CATALOGUE OF EMPIRICAL WORK RELATED TO THE SIZE OF GOOGLE SCHOLAR

AUTHORS	SAMPLE	UNIT	GS/GSM/GSC	WoS	MAS	SCO	PUB	PSY	ERIC	SSCI	SJR	JCR	JSC	ECO	CAB	SCI	CA	GB	KoMCI	KMS	CINAHL	
KHABSA & GILES (2014)	150 English written documents from MAS; 10 of the most cited documents in each of the 15 fields are randomly sampled.	Citations	86,870		41,778																	
KHABSA & GILES (2014)	1,500 documents from MAS; 100 documents belonging to each field, with at least 1 citation.(n=114 million)	Documents (million)	99.3	49	48																	
ABDULLAH & THELWALL (2014)	Books (n= 1,357) and citations (n=2,254) of Malaysian books in AHSS disciplines.	Documents	499															314				
		Documents (%)	37																23			
		Citations per book	1.67														1.61					
WINTER & ZADPOOR & DODOU (2014)	Number of citations on to Garfield for WoS and GS	Citations	1,231	607																		
		Unique citations	703	153																		
		(%)	90	48.2																		
HALEY (2014)	50 Top economics finance journals were selected and then scored using both GS and MAS using the PoP software (1993–2012)	Average h-index	154.40	78.98																		
		Average h-index	267.36	125.72																		
		Average AWCR	9834.00	2740.87																		
		Average e-index	186.21	81.00																		
ORDUÑA-MALEA & DELGADO LÓPEZ-CÓZAR (2014)	World weekly (average) size and monthly growth rate per source	Weekly size (average)	68,545,750	27,904,896	28,113,479																	
		Monthly growth rate (%)	11.15	0.37	0.41																	
ORTEGA & AGUILLO (2014)	Analysis of the 771 personal profiles appearing in both the MAS and the GSC.	Documents (%)	158.3		89.5																	
		Citations (%)	327.4		76.7																	
		%	155.8		72.1																	
CARDENAS & UDO (2013)	Knowledge management (KM) articles published between 1993 and 2012	KM papers	33,600	9,887																		
		KM papers (%)	77.26	28.59																		
		KM papers in USA	12,434	2,084																		
		KM papers in USA (%)	80.65	14.35																		
ADRIAANSE & RENSLEIGH (2013)	African scholarly environmental sciences journals the period 2004-2008 (n=3,199)	Citations	2,715	2,740		2,192																
		Overall coverage (%)	84.9	85.7		68.5																
		Inconsistencies	448	165		14																
		(%)	14	5.2		0.4																
CABEZAS-CLAVIJO & DELGADO LÓPEZ-CÓZAR (2013)	Most relevant journals and researchers in the field of intensive care medicine.	Average Journal H-index	36	28		32																
		Average Author H-index	29	23		25																
DELGADO LÓPEZ-CÓZAR & CABEZAS-CLAVIJO (2013)	N° Journals indexed GSM, JCR and SJR	Journals	40,000								19,708	10,677										
DELGADO LÓPEZ-CÓZAR & REPISO (2013)	Sample of journals from the field of communication studies indexed in three databases.	Number of communication journals covered	277	106		167																

JACIMOVIĆ & PETROVIĆ & ŽIVKOVIĆ (2010)	Number of cited articles from SDJ	Unique citations	144	4	9																
		(%)	57.84	1.61	3.61																
		GS and WoS= 2 (0.80 %); GS and SCO = 6 (2.41%); SCO and WoS = 37 (14.86%); GS and SCO and WoS = 37 (14.86%)																			
MINGERS & LIPITAKIS (2010)	Publications from 3 UK Business Schools. (n=4,600)	Publications	3,023	1,004																	
		(%)	65.72	21.83																	
ŠEMBER & UTROBIČIĆ & PETRAK (2010)	Croatian medical journal indexed articles (2005-2006)	Unique citations	86	12	39																
		(%)	22	3	10																
		GS and WoS= 9 (2 %); GS and SCO = 36 (9 %); SCO and WoS = 47 (12%); GS and SCO and WoS = 166 (42%)																			
BORNMANN et al (2009)	Papers published (n=1,837) by the journal Angewandte Chemie	Publications	1,747		1,827														1,837	1,837	
		(%)	95.1		99.5															100	100
		Citations	9,320		44,601															44,502	48,160
FRANCESCHET (2009)	Publications and cites of computer scientist group	Publications	1,776	324																	
		(%)	84.57	15.43																	
		Citations	10,690	1,378																	
JACOBS (2009)	Comparison of citation counts for 30 Top articles in Gender & Society	Citations	8,047	3,667																	
		Ratio of Google to ISI = 2.19																			
KULKARNI et al (2009)	Cohort study of 328 articles published	Citations	83,538	68,088	82,076																
MARTELL (2009)	Title search, citations and average citations per article (n=217)	Citations	1,394	680																	
		Average citations per article	6.4	3.1																	
MIKKI (2009)	GS is compared with WoS for earth science authors (n=29)	Publications	5,048	1,573																	
		(%)	76.25	23.76																	
		Citations	40,908	43,028																	
		Average h-index	16.0	16.7																	
MOSKOVKIN (2009)	Publications of the 10 largest universities (2008)	Publications	565,709	55,581																	
		(%)	91.06	8.94																	
ONYANCHA & OCHOLLA & (2009)	Comparison of 10 purposefully selected LIS researchers in South Africa	Publications	384	182	96																
		(%)	58	27.50	14.50																
		Citations	887	125	190																
		Average H-index	5	1.7	2.3																
HARZING & VAN DER WAL (2008)	Comparison between WoS and GS for the impact of books between 1991-2001	Citations	883	346																	
		GS reports 2.5 times as many citations as WoS.																			
KOUSHA & THELWALL (2008)	A sample of 882 articles from 39 open access ISI-indexed journals in 2001	Citations	5,589	4,184																	
		Unique citations	3,202	1,797																	
		ISI citations overlapping with Google Scholar = 2,387																			
BAR-ILAN (2007)	Compares the h-index of highly-cited Israeli researchers	Average H-index	17.55	17.3	17.1																
		Average citations	245.64	162.85	170.27																
MEHO & YANG (2007)	Citations to 25 library and information science (LIS) faculty members (n=5,285)	Distribution of unique and overlapping citations	GS =2,552 (48.3%); SCO AND WoS=1,104 (20.9%); GS, SCO AND WoS=1,629 (30.8%); GS identifies 1,448 (53.0%) more citations than WoS and Scopus together (4,181 citations for GS in comparison to 2,733 for the union of WoS and Scopus)																		
BAKKALBASI et al (2006)	11 journal titles from each discipline (oncology) using the JCR. All articles (n=614) published 1993- 2003	Unique Citations	78	41	74																
		(%)	13	7	12																
		GS AND WoS = 26 (4%); GS AND SCO = 31 (5%); WoS AND SCO = 175 (28%); GS, WoS AND SCO = 189 (31%)																			
BAKKALBASI et al (2006)	11 journal titles (condensed matter physics) using the JCR. All articles(n=296) published 1993-2003	Unique citations	50	63	25																
		(%)	17	20	8																
		GS AND WoS = 21 (9%); GS AND SCO = 9 (3%); WoS AND SCO = 65 (22%); GS, WoS AND SCO = 63 (21%)																			

SALISBURY & TEKAWADE (2006)	Journal coverage for Agricultural Economics and AgriBusiness (2004-2005)	No. of Titles	184									92	133						
		Average year	92										46	66.5					
		% (n=108)	85.19										42.59	61.57					
		Citations (%)	39										17	44					
YANG & MEHO (2006)	Items published by two Library and Information Science full-time faculty members.	Unique citations	38	89	25														
		(%)	9.9	23.1	6.5														
JACSO (2005b)	Citations count for the papers published in 22 volumes of APJAI (n=698)	Documents	680	675															
		(%)	97.42	96.60															
		Citations	595	1,355															
NORUZI (2005)	Citation counts from Google Scholar and Web of Science (WoS) for Almind & Ingwersen	Citations	98	81															
		Unique citations	64	47															
		Citations GS AND WoS = 34																	
NORUZI (2005)	Average cites of the most-cited 36 Authors in the field of Webometrics on GS and WoS	Citations	1,110	729															
		Average citations	30.84	20.25															

ACRONYMS:

GS/GSM/GSC: Google Scholar / Google Scholar Metrics/ Google Scholar Citations

WoS: Web of Science

MAS: Microsoft Academic Search

SCO: Scopus

PUB: Pubmed

PSY: PsycINFO

ERIC: Education Resources Information Center

SSCI: Social Sciences Citation Inde

SJR: SCImago Journal Rank

JCR: Journal Citation Reports

ECON: Econlit

CAB: CAB ABSTRACTS

SCI: Science Citation Index

CA: Chemical Abstracts

GB: Google Books

KoMCI: Korean Medical Citation Index

KMS: KoreaMed Synapse

CINAHL: Cumulative Index to Nursing and Allied Health Literature

APPENDIX II. EMPIRICAL STUDIES ABOUT GOOGLE SCHOLAR ACCORDING TO UNIT OF ANALYSIS AND DOCUMENT TYPE

AUTHORS	SAMPLE	TYPE ANALYSIS GS	UNIT	GS/GSM/GSC	%	WoS	%	SCO	%
WINTER, & ZADPOOR & DODOU (2014)	Number of citations on 5 April 2013 to Garfield (1955) for WoS and GS as a function of document type	Citations	Journals	805	69.6	546	90.1		
			Conferences	123	10.6	53	8.7		
			Books or book chapters	63	5.4	7	1.2		
			Theses	75	6.5	0	0		
			Reports	13	1.1	0	0		
			Other	43	3.7	0	0		
			Unknown	34	2.9	0	0		
			Duplicates	64	-	0	0		
			False positives	11	-	1	-		
			All types (incl. duplicates and false positives)	1,231		607			
MIRI & RAOOFI & HEIDARI (2012)	Comparison of Number of Citations in ISI, GS, and SC based on Article Types Published in Hepatitis Monthly (2008, 2009)	Citations	Original Article	39		30		40	
			Review Article	28		25		25	
			Brief Report	9		8		10	
			Editorial	8		4		7	
			Case Report	3		3		3	
			Letter to the Editor	-		1		1	
			Guidelines and Clinical Algorithm	-		-		-	
LASDA BERGMAN (2012)	Source types of citing references	Citations	Article	1,951	59.6	1,735	99.7	1,782	83.8
			Book	318	9.7	-	-	1	0.0
			Conference Paper	32	1.0	-	-	25	1.2
			Foreign Language	281	8.6	-	-	-	-
			Government Document	44	1.3	-	-	-	-
			Dissertation	329	10.1	-	-	-	-
			Master's Thesis	108	3.3	-	-	-	-
			Bachelor's Thesis	6	0.2	-	-	-	-
			Report	84	2.6	-	-	-	-
			Syllabus	5	0.2	-	-	-	-
			Unpublished Manuscript	44	1.3	-	-	-	-
			Working Paper	35	1.1	-	-	-	-
			Review	24	0.7	-	-	248	11.7
			Presentation Slides	3	0.1	-	-	-	-
			Blog	3	0.1	-	-	-	-
			Editorial	1	0.0	-	-	28	1.3
			Letters to the Editor	1	0.0	-	-	10	0.5
			Supplementary Material	1	0.0	-	-	-	-
			Web Page	1	0.0	-	-	-	-
			Guideline	1	0.0	-	-	-	-
			Series	-	-	6	0.3	-	-
Short Survey	-	-	-	-	5	0.2			
Note	-	-	-	-	27	1.3			
Total			3,272	100.0	1,741	100.0	2,126	100.0	
MEHO & YANG (2007)	Citations to the work of 25 LIS faculty members Citation count by document type (1996 –2005).	Citations	Journal articles	2,215	40.32	1,529	75.6	1,754	76.2
			Conference papers	1,849	33.66	229	11.3	359	15.6
			Review articles	86	1.57	172	8.5	147	6.4
			Editorial materials	25	0.46	63	3.1	36	1.6
			Book reviews	3	0.05	17	0.8	0	0.0
			Letters to the editor	2	0.04	9	0.4	2	0.1
			Biographical item	1	0.02	2	0.1	1	0.0
			Doctoral dissertations	261	4.75	-	-	-	-
			Master's theses	243	4.42	-	-	-	-
			Book chapters	199	3.62	-	-	-	-
Technical reports	129	2.35	-	-	-	-			

			Reports	110	2.00	-	-	-	-
			Books	102	1.86	-	-	-	-
			Conference presentations	72	1.31	-	-	-	-
			Unpublished papers	65	1.18	-	-	-	-
			Bachelor's theses	34	0.62	-	-	-	-
			Working papers	31	0.56	-	-	-	-
			Research reports	23	0.42	-	-	-	-
			Workshop papers	15	0.27	-	-	-	-
			Doctoral dissertation proposals	9	0.16	-	-	-	-
			Conference posters	9	0.16	-	-	-	-
			Book reviews	3	0.05	-	-	-	-
			Master's thesis proposals	3	0.05	-	-	-	-
			Preprints	3	0.05	-	-	-	-
			Conference paper proposals	2	0.04	-	-	-	-
			Government documents	2	0.04	-	-	-	-
			Total	5,493	100.00	2,023	100.0	2,301	100.0
			Total from journals	2,332	42.45	1,794	88.7	1,942	84.4
			Total from conference papers	1,849	33.66	229	11.3	359	15.6
			Total from journals and conferences	4,181	76.12	2,023	100.0	2,301	100.0
			Total from dissertations/theses	538	9.79	-	-	-	-
			Total from books	301	5.48	-	-	-	-
			Total from reports	262	4.77	-	-	-	-
			Total from other document types	211	3.84	-	-	-	-
YANG & MEHO (2006)	Breakdown of Citations Found in Google Scholay by Document Type by two Library and Information Science full-time faculty members. In	Citations	Journal Articles	169	48,4				
			Conference Papers	90	25,8				
			Research reports	39	11,2				
			Dissertations and Theses	15	4,3				
			Dead links	7	2,0				
			Editorial Materials No access	6	1,7				
			Workshops	5	1,4				
			No access	4	1,1				
			Technical reports	3	0,9				
			Websites	3	0,9				
			Other (chapters, bibliographies)	8	2,3				
			Total	349	100				

AUTHORS	SAMPLE	TYPE ANALYSIS GS	PUBLICATION TYPES	N	% OF OUTPUTS	NO. OF PUBS FOUND IN GS	NO. OF PUBS FOUND IN WOS	% GS	% WOK	NO. OF CITATIONS FOUND IN GS	NO. OF CITATIONS FOUND IN WOS	GS CITATION PER PAPER (CPP)	WOS CITATION PER PAPER (CPP)
MINGERS & LIPITAKIS (2010)	GS and WoS citations by publication type. No. Publications from 3 UK Business Schools. (n= 4,600)	Cites/Documents	Total books	95	2.1	70				2,257		32.24	
			Books A	45	2.3	38		84.4		1,285		33.8	
			Books B	31	2.1	21		67.7		567		27.0	
			Books C	19	1.6	11		57.9		405		36.8	
			Total edited books	76	1.7	58				1,763		30.40	
			Edited Books A	48	2.5	39		81.3		1,394		35.7	
			Edited Books B	16	1.1	11		68.8		56		5.1	
			Edited Books C	12	1.0	8		66.7		313		39.1	
			Total book chapters	619	13.4	287				1,946		6.78	
			Book Chapters A	326	16.9	149		45.7		1,178		7.9	
			Book Chapters B	184	12.6	74		40.2		289		3.9	
			Book Chapters C	109	9.0	64		58.7		479		7.5	
			Total journal articles	2,109	45.8	1,882	1,004			27,606	8,434	14.67	8.40
			Journal Articles A	801	41.4	705	403	88.0	50.3	15,167	4,554	21.5	11.3
			Journal Articles B	715	49.1	629	309	88.0	43.2	6,831	2,361	10.9	7.6
			Journal Articles C	593	48.9	548	292	92.4	49.2	5,608	1,519	10.2	5.2
			Total conference papers	1,013	22.0	340				848		2.49	
			Conference Papers A	298	15.4	73		24.5		151		2.1	
			Conference Papers B	356	24.5	99		27.8		240		2.4	
			Conference Papers C	359	29.6	168		46.8		457		2.7	
			Total working papers	417	8.8	286				1,535		5.37	
			Working Papers A	317	16.4	235		74.1		1340		5.7	
			Working Papers B	5	0.3	1		20.0		0		0.0	
			Working Papers C	85	7.0	50		58.8		195		3.9	
			Total reports	171	3.7	59				491		8.32	
			Reports A	79	4.1	32		40.5		306		9.6	
			Reports B	62	4.3	14		22.6		61		4.4	
			Reports C	30	2.5	13		43.3		124		9.5	
			Total others	110	2.4	41				133		3.24	
			Others A	19	1.0	10		52.6		77		7.7	
			Others B	86	5.9	27		31.4		37		1.4	
			Others C	5	0.4	4		80.0		19		4.8	
Total	4,600		3,023	1,004			36,579	8,434	12.1	8.4			
Total A	1,933	100.0	1,281	403	66.3	50.3	20,898	4,554	16.3	11.3			
Total B	1,455	100.0	876	309	60.2	43.2	8,081	2,361	9.2	7.6			
Total C	1,212	100.0	866	292	71.5	49.2	7,6	1,519	8.8	5.2			

AUTHORS	SAMPLE	TYPE ANALYSIS GS	UNIT	GS/GSM/GSC	WoS	SCO
JAĆIMOVIĆ & PETROVIĆ & ŽIVKOVIĆ (2010)	SDJ citation was collected in September 2010	Cites	Article	117	50	56
			Review	13	4	6
			Editorial	3	1	3
			Proceedings	3	2	3
			Miscellaneous	8	-	-
BAKKALBASI et al (2006)	11 journal titles from each discipline (1993-2003).	Cites		Oncology	CM Phys	
			Journal	31 (62 %)	18 (37%)	
			Archive	3 (6%)	12 (25%)	
			College or University	9 (18%)	6 (13%)	
			Government	3 (6%)	4 (8%)	
			Non-Governmental Organization	2 (4%)	8 (17 %)	
			Commercial	0	0	
			Other	2 (4%)	0	
			Total	50	48	

AUTHORS	SAMPLE	TYPE ANALYSIS GS	PUBLICATION TYPES	GS		WoS		SCOPUS		TOTAL	
				NUMBER OF CITED	NUMBER OF RECEIVED CITATIONS	NUMBER OF CITED	NUMBER OF RECEIVED CITATIONS	NUMBER OF CITED	NUMBER OF RECEIVED CITATIONS	NUMBER OF CITED	NUMBER OF RECEIVED CITATIONS
JAĆIMOVIĆ & PETROVIĆ & ŽIVKOVIĆ (2010)	Type of cited articles from SDJ was collected in September 2010	Cites	Informative article	23	43	13	14	18	22	32	55
			Original scientific article	57	86	39	50	38	50	76	119
			Case report	5	5	2	3	2	3	5	6
			Proceedings	20	31	7	7	5	5	26	37
			Review	7	16	4	6	4	6	8	17
			Professional article	3	6	4	5	5	7	8	12
			Preliminary communication	1	1	0	0	0	0	1	1
			Article from praxis	0	0	0	0	1	1	1	1
			Book review	1	1	0	0	0	0	1	1
			Total	117	189	69	85	73	94	158	249

AUTHORS	SAMPLE	TYPE ANALYSIS GS	UNIT	GS/GSM/GSC	%	WoS	%	SCO	%	CA	%
BAR-ILAN (2010)	Document types of the unique items retrieved by GS (=109) were collected 2008.	Documents	Journal	28	25.7%						
			Proceedings	25	22.9%						
			Thesis	15	13.8%						
			Book chapter	13	11.9%						
			Report	10	9.2%						
			Manuscript	7	6.4%						
			In Chinese	4	3.7%						
			Book	3	2.8%						
			Newsletter	2	1.8%						
			Encyclopedia entry	1	0.9%						
LEVINE-CLARK & KRAUS (2007)	Compare GS and CA for finding chemistry information in six different searches (n=702)	Documents	Journal Articles (n=564)	482	85.5					521	92.4
			Patent (n=54)	4	7.4					24	100
			Problem (n=26)	26	100					0	0.0
			Conference proceedings(n=23)	11	47.8					12	52.2
			Book(n=21)	21	100					7	33.3
			Dissertation(n=9)	9	100					5	55.6
Other (n= 5)	5	100					2	40			

APPENDIX III. EMPIRICAL STUDIES ABOUT GOOGLE SCHOLAR ACCORDING TO LANGUAGES

AUTHORS	SAMPLE	UNIT GS	LANGUAGE	GS/GSM/GSC	%	WoS	%	SCO	%
DELGADO LÓPEZ-CÓZAR & REPISO (2013)	Sample of journals from the field of communication studies indexed in three databases (n=277).	Journals	English	181	65.3	93	87.8	153	91.6
			Spanish	42	15.2	6	5.7	10	6
			Chinese	27	9.7	0	0.0	1	0.6
			Portuguese	24	8.7	0	0.0	3	1.8
			French	12	4.3	4	3.8	4	2.4
			German	7	2.5	0	0.0	0	0.0
			Italian	2	0.7	1	0.9	1	0.6
			Russian	1	0.4	0	0.0	0	0.0
			Danish	1	0.4	0	0.0	1	0.6
			Japanese	3	1.1	0	0.0	0	0.0
			Romanian	1	0.4	0	0.0	0	0.0
			Polish	0	0.0	0	0.0	0	0.0
			Croatian	1	0.4	1	0.9	0	0.0
			Dutch	0	0.0	1	0.9	0	0.0
Nowegian	0	0.0	0	0.0	0	0.0			
REINA-LEAL & REPISO & DELGADO LÓPEZ-CÓZAR (2013)	Nursing journals on Google Scholar Metrics	Journals	English	208	61.9				
			Chinese	24	7.1				
			Portuguese	12	3.6				
			Multilangage	12	3.6				
			German	4	1.2				
			Korean	15	4.5				
			Spanish	19	5.7				
			French	11	3.3				
			Hindi	1	0.3				
			Italian	1	0.3				
			Persian	2	0.6				
			Polish	1	0.3				
			Japanese	19	5.7				
			Dutch	3	0.9				
			Bulgarian	1	0.3				
			Catalan	1	0.3				
			Italian	1	0.3				
Tukish	1	0.3							
JAĆIMOVIĆ & PETROVIĆ & ŽIVKOVIĆ (2010)	Type of cited articles from SDJ was collected in September 2010	Citations	English	39	27	40	82.4	47	69.1
			Serbian	61	42.3	17	29.8	17	25
			Bilingual	36	25	0	0.0	2	2.9
			Other	8	5.5	0	0.0	2	2.9
KOUSHA & THELWALL (2008)	A sample of 882 articles from 39 open access ISI-indexed journals in 2001	Citations		BIOLOGY	CHEMISTRY	PHYSICS	COMPUTING		
				%	%	%	%		
			English	57	65.5	96	96		
			Chinese	36	23	2	1.5		
	Non-English	7	11.5	2	2.5				
MEIER & CONKLING (2008)	Records retrieved from Compendex were searched in Google Scholar, (1950-2007)	Documents		The range of non-English language content in Compendex varied between 10.8% and 28.8% for the disciplines. The average amount of non-English materials was 20.5% for those years. In this study, only 11.3% of the missed papers in Google Scholar were non-English.					
MEHO & YANG (2007)	Citations to the work of 25 LIS faculty members. Citation count distribution by language (1996 –2005)	Citations	English	3,891	93.06	2	98.86	2,285	99.30
			Portuguese	92	2.20				
			Spanish	63	1.51	4	0.20	3	0.13
			German	38	0.91	13	0.64	9	0.39

			Chinese	44	1.05				
			French	32	0.77				
			Italian	8	0.19	3	0.15	1	0.04
			Japanese	1	0.02				
			Swedish	3	0.07	3	0.15	3	0.13
			Czech	2	0.05				
			Dutch	2	0.05				
			Finnish	2	0.05				
			Croatian	1	0.02				
			Hungarian	1	0.02				
			Polish	1	0.02				
			Non-English	290	6.94	23	1.14	16	0.70
			Total	4,181	100	2,023	100	2,301	100
NEUHAUS & NEUHAUS & ASHER & WREDE (2006)	Contents of 47 different databases with that of Google Scholar, (April-July, 2005)	Documents (PsycINFO)	English		68				
			Non-English		12				

ATTACHMENT 5

GO

Frequently Asked Questions

[[The Internet Archive](#) | [Search Tips](#) | [Prelinger Movies](#) | [The Wayback Machine](#) | [Audio](#) | [MS-DOS Emulation](#) | [Archive BitTorrents](#) | [Accounts Information](#) | [Navigation](#) | [Live Music Archive](#) | [Movies](#) | [Collections](#) | [Downloading Content](#) | [Law Enforcement Requests](#) | [The Internet Arcade](#) | [Uploading Content](#) | [Books and Texts](#) | [Item page management](#) | [Rights](#) | [Borrow from Lending Library](#) | [The Grateful Dead Collection](#) | [Report Item](#) | [Forums](#) | [SFLan](#) | [Archive-It](#) | [Equipment](#) | [Errors](#)]

Questions

Does the Archive issue grants?

Can I donate BitCoins?

What is the nonprofit status of the Internet Archive? From where does its funding come?

How do I get assistance with research? How about research about a particular book?

What statistics are available about use of Archive.org?

What's the significance of the Archive's collections?

The Internet Archive

Does the Archive issue grants?

No; although we promote the development of other Internet libraries through [online discussion](#), [colloquia](#), and other means, the Archive is not a grant-making organization.

Can I donate BitCoins?

Yes, please do. Our BitCoin address is: 1Archive1n2C579dMsAu3iC6tWzuQJz8dN . Every bit helps.

What is the nonprofit status of the Internet Archive? From where does its funding come?

The Internet Archive is a 501(c)(3) nonprofit organization. It receives in-kind and financial donations from a variety of sources as well as [you](#).

How do I get assistance with research? How about research about a particular book?

The Internet Archive focuses on preservation and providing access to digital cultural artifacts. For assistance with research or appraisal, you are bound to find the information you seek elsewhere on the internet. You may wish to inquire about reference services provided by your local public library. Your area's college library may also support specialized reference librarian services. We encourage your support of your local library, and the essential services your library's professional staff can provide in person. Local libraries are still an irreplaceable resource!

What statistics are available about use of Archive.org?

What user stats do you keep and share?

The only users stats we track are the "views" of items on the site.

Where are they?

For collections they are viewable in a chart form in the "About" tab on a collection page. These numbers represent views in all the items in that collection. These are updated daily. For items they are shown on the right side of the details page. These are updated daily. Search results pages also show the "views" to the left of the page title. These numbers may differ from those on item and collection pages because they are updated monthly rather than daily.

What is a "view"?

A "view" used to be called a "download" on archive.org. How are "views" counted? archive.org calculates a view as: one action (read a book, download a file, watch a movie, etc.), per day, per IP Address. So, for each item page, using multiple files or accessing from multiple accounts in a single day will only count as one view.

How often are they counted?

Item pages are updated daily so the current number would reflect the count through the previous day.

Collection counts shown in the graph on the "About" page are updated monthly.

Other Internet Archive stats links

Aggregated operational stats are viewable at <https://archive.org/stats/>

What's the significance of the Archive's collections?

Societies have always placed importance on preserving their culture and heritage. But much early 20th-century media -- television and radio, for example -- was not saved. [The Library of Alexandria](#) -- an ancient center of learning containing a copy of every book in the world -- disappeared when it was burned to the ground.

Questions Search Tips

Where is advanced search?

Where is advanced search?

On archive.org there is an "Advanced Search" link just below the search input field. For searches done in the search top black nav bar the "Advanced Search" link will be present on the search results page just below the search input

What search APIs are available?

What search APIs are available

Information about how to use the various search APIs can be found at <https://archive.org/help/aboutsearch.htm>

Can I search by Creative Commons license?

Can I search by Creative Commons license?

Yes, you can. But it's a little complicated.

Here's how to break it down. See the license types at [creative commons](#). When you want to find all of the items assigned a certain license by an uploading party, you'll plug their abbreviation for it into this search query:

How do I sort search results?

`licenseurl:http*abbreviation*`

How do I search just within a collection?

So if you're looking for Attribution Non-commercial No Derivatives (by-nc-nd), you'd put this in the search box:

`licenseurl:http*by-nc-nd*`

If you want to use this in combination with other queries, like "I want by-nc-nd items about dogs" you'd do this:

`licenseurl:http*by-nc-nd* AND dog.`

How can I use list view instead of tile view?

The AND tells the search engine all the items returned should have that license AND they should contain the word dog to be in all caps.

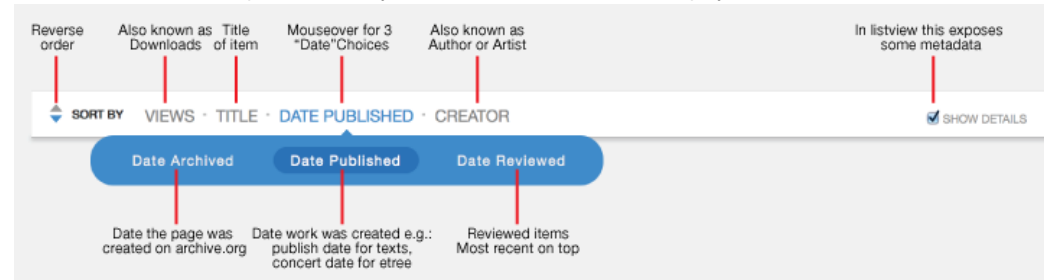
What is indexed in the search engine?

Just to make it easier, here are the basic searches:

- [Public Domain](#)
- [Attribution Non-commercial No Derivatives \(by-nc-nd\)](#)
- [Attribution Non-commercial Share Alike \(by-nc-sa\)](#)
- [Attribution Non-commercial \(by-nc\)](#)
- [Attribution No Derivatives \(by-nd\)](#)
- [Attribution Share Alike \(by-sa\)](#)
- [Attribution \(by\)](#)

How do I sort search results

The "SORT BY" bar has options to allow you to control which results are displayed, in what order and what "view":



How do I search just within a collection?

On a collection page there will be a "Search this Collection" input field on the right side of the page. Enter a term the you return/enter key. The results will be of items in that collection.

For advanced boolean search you can use "AND collection:[IDENTIFIER]" in your query.

How can I use list view instead of tile view?

For most search results pages you can choose the view in the "Sort by" bar; Tile view (the icon with three rectangles (the icon with multiple lines.)) Tile view is the default view.

What is indexed in the search engine?

Only the metadata in an item page is indexed. So the search engine does not have the text of books, individual file embedded metadata.

Questions

How did you digitize the films?

Do I need to inform the Internet Archive and/or Prelinger Archives when I reuse these movies?

How can I get access to stock footage from these films?

An article on re-coding Prelinger Archive films to SVCD so you can watch them on your DVD player.

Do I need to credit the Internet Archive and Prelinger Archives when I reuse these movies?

What parameters were used when making the Real Media files on the website?

Are there restrictions on the use of the Prelinger Films?

Can you point me to resources on the history of ephemeral films?

Why are there very few post-1964 movies in the Prelinger collection?

For more information...

Prelinger Movies

How did you digitize the films?

The [Prelinger Archives](#) films are held in original film form (35mm, 16mm, 8mm, Super 8mm, and various obsolete formats like 28mm and 9.5mm). Films were first transferred to Betacam SP videotape, a widely used analog broadcast video standard, on telecine machines manufactured by Rank Cintel or Bosch. The film-to-tape transfer process is not a real-time process: It requires inspection of the film, repair of any physical damage, and supervision by a skilled operator who manipulates color, contrast, speed, and video controls.

The videotape masters created in the film-to-tape transfer suite were digitized in 2001-2003 at Prelinger Archives in New York City using an encoding workstation built by [Rod Hewitt](#). The workstation is a 550 MHz PC with a [FutureTel](#) NS320 MPEG encoder card. Custom software, also written by Rod Hewitt, drove the Betacam SP playback deck and managed the encoding process. The files were uploaded to hard disk through the courtesy of [Flycode, Inc.](#)

More recently, Prelinger films have been digitized and uploaded by Skip Elsheimer at [AV Geeks](#). We are also digitizing home movies and other materials on Internet Archive's ScanStation scanner.

The files were encoded at constant bitrates ranging from 2.75 Mbps to 3.5 Mbps. Most were encoded at 480 x 480 pixels (2/3 D1) or 368 x 480 (roughly 1/2 D1). The encoder drops horizontal pixels during the digitizing process, which during decoding are interpolated by the decoder to produce a 720 x 480 picture. (Rod Hewitt's site [Coolstf](#) shows examples of an image [before](#) and [after](#) this process.) Picture quality is equal to or better than most direct broadcast satellite television. Audio was encoded at MPEG-1 Level 2, generally at 112 kbps. Both the MPEG-2 and MPEG-4 movies have mono audio tracks.

To convert the MPEG-2 video to MPEG-4, we used a program called Flak MPEG. This is an MPEG-1/2 to AVI conversion tool that reads the source MPEG-2 and outputs an AVI file containing the video in MPEG-4 format and audio in uncompressed PCM format. We then use a program called Virtual Dub that recompresses the audio using the MPEG-1 Level 3 (MP3) format. This process is automated by the software that runs the system.

Do I need to inform the Internet Archive and/or Prelinger Archives when I reuse these movies?

No. However, we would very much like to know how you have used this material, and we'd be thrilled to see what you've made with it. This may well help us improve this site. Please consider sending us a copy of your production (postal mail only), and let us know whether we can call attention to it on the site. Our address is:

Rick Prelinger
PO Box 590622
San Francisco, CA 94159
United States

How can I get access to stock footage from these films?

Access to the movies stored on this site in videotape or film form is available to commercial users through [Getty Images](#), representing Prelinger Archives for stock footage sales. Please contact Getty Images directly:

[Getty Images](#)

Please visit us at www.prelinger.com/prelarch.html for more information on access to these and similar films. Prelinger Archives regrets that it cannot generally provide access to movies stored on this Web site in other ways than through the site itself. We recognize that circumstances may arise when such access should be granted, and we welcome email requests. Please address them to [Rick Prelinger](#).

The Internet Archive does not provide access to these films other than through this site.

An article on re-coding Prelinger Archive films to SVCD so you can watch them on your DVD player.

See [archived version of www.moviebone.com/](#)

Do I need to credit the Internet Archive and Prelinger Archives when I reuse these movies?

We ask that you credit us as a source of archival material, in order to help make others aware of this site. We suggest the following forms of credit:

Archival footage supplied by Internet Archive (at archive.org) in association with Prelinger Archives

or

Archival footage supplied by Internet Archive (at archive.org)

or

"Archival footage supplied by archive.org"

What parameters were used when making the Real Media files on the website?

Rod Hewitt posted some very useful information [here](#)

Are there restrictions on the use of the Prelinger Films?

The films are available for reuse according to the Creative Commons licenses, if any, that appear with on each film's detail page. Pursuant to the Creative Commons license, you are warmly encouraged to download, use and reproduce these films in whole or in part, in any medium or market throughout the world. You are also warmly encouraged to share, exchange, redistribute, transfer and copy these films, and especially encouraged to do so for free.

Any derivative works that you produce using these films are yours to perform, publish, reproduce, sell, or distribute in any way you wish without any limitations.

Descriptions, synopses, shotlists and other metadata provided by Prelinger Archives to this site are copyrighted jointly by Prelinger Archives and Getty Images. They may be quoted, excerpted or reproduced for educational, scholarly, nonprofit or archival purposes, but may not be reproduced for commercial purposes of any kind without permission.

If you require a written license agreement or need access to stock footage in a physical format (such as videotape or a higher-quality digital file), please contact [Getty Images](#). The Internet Archive does not furnish written license agreements, nor does it comment on the rights status of a given film above and beyond the Creative Commons license.

We would appreciate attribution or credit whenever possible, but do not require it.

Can you point me to resources on the history of ephemeral films?

See the bibliography and links to other resources at [www.prelinger.com/ephemeral.html](#).

Why are there very few post-1964 movies in the Prelinger collection?

Largely because of copyright law. While a high percentage of ephemeral films were never originally copyrighted or (if initially copyrighted) never had their copyrights properly renewed, copyright laws still protect most moving image works produced in the United States from 1964 to the present. Since the Prelinger collection on this site exists to supply material to users without most rights restrictions, every title has been checked for copyright status. Those titles that either are copyrighted or whose status is in question have not been made available. For information on recent changes in copyright law, see the circular [Duration of Copyright](#) (in [PDF format](#)) published by the Library of Congress

For more information...

Check out our [Prelinger Archives Forum](#)

Questions

[Can I link to old pages on the Wayback Machine?](#)

[Who was involved in the creation of the Internet Archive Wayback Machine?](#)

[How was the Wayback Machine made?](#)

[How do you archive dynamic pages?](#)

[How can I use the Wayback Machine's Site Search to find websites?](#)

[Can I search the Archive?](#)

[Do you collect all the sites on the Web?](#)

[Why isn't the site I'm looking for in the archive?](#)

[How can I have my site's pages excluded from the Wayback Machine?](#)

[How can I use the Wayback Machine's Site Search to find websites?](#)

[Why is the Internet Archive collecting sites from the Internet? What makes the information useful?](#)

[Do you archive email? Chat?](#)

[How can I get a copy of the pages on my Web site? If my site got hacked or damaged, could I get a backup from the Archive?](#)

[Is there any personal information in these collections?](#)

[Can I add pages to the Wayback Machine?](#)

[How do I contact the Internet Archive?](#)

[Where is the rest of the archived site? Why am I getting broken or gray images on a site?](#)

[Why are some sites harder to archive than others?](#)

The Wayback Machine

Can I link to old pages on the Wayback Machine?

Yes! The Wayback Machine is built so that it can be used and referenced. If you find an archived page that you would like to reference on your Web page or in an article, you can copy the URL. You can even use fuzzy URL matching and date specification... but that's a bit more advanced.

Who was involved in the creation of the Internet Archive Wayback Machine?

"The original idea for the Internet Archive Wayback Machine began in 1996, when the Internet Archive first began archiving the web. Now, five years later, with over 100 terabytes and a dozen web crawls completed, the Internet Archive has made the Internet Archive Wayback Machine available to the public. The Internet Archive has relied on donations of web crawls, technology, and expertise from Alexa Internet and others. The Internet Archive Wayback Machine is owned and operated by the Internet Archive."

How was the Wayback Machine made?

Alexa Internet, in cooperation with the Internet Archive, has designed a three dimensional index that allows browsing of web documents over multiple time periods, and turned this unique feature into the Wayback Machine.

How do you archive dynamic pages?

There are many different kinds of dynamic pages, some of which are easily stored in an archive and some of which fall apart completely. When a dynamic page renders standard html, the archive works beautifully. When a dynamic page contains forms, JavaScript, or other elements that require interaction with the originating host, the archive will not contain the original site's functionality.

How can I use the Wayback Machine's Site Search to find websites?

The Site Search feature of the Wayback Machine is based on an index built by evaluating terms from hundreds of billions of links to the homepages of more than 350 million sites. Search results are ranked by the number of captures in the Wayback and the number of relevant links to the site's homepage.

Can I search the Archive?

Using the Internet Archive Wayback Machine, it is possible to search for the names of sites contained in the Archive (URLs) and to specify date ranges for your search. We hope to implement a full text search engine at some point in the future.

Do you collect all the sites on the Web?

No, the Archive collects web pages that are publicly available. We do not archive pages that require a password to access, pages that are only accessible when a person types into and sends a form, or pages on secure servers. Pages may not be archived due to robots exclusions and some sites are excluded by direct site owner request.

Why isn't the site I'm looking for in the archive?

Some sites may not be included because the automated crawlers were unaware of their existence at the time of the crawl. It's also possible that some sites were not archived because they were password protected, blocked by robots.txt, or otherwise inaccessible to our automated systems. Site owners might have also requested that their sites be excluded from the Wayback Machine.

How can I have my site's pages excluded from the Wayback Machine?

You can send an email request for us to review to info@archive.org with the URL (web address) in the text of your message.

How can I use the Wayback Machine's Site Search to find websites?

Can I find sites by searching for words that are in their pages?

The Site Search feature of the Wayback Machine is based on an index built by evaluating terms from hundreds of billions of links to the homepages of more than 350 million sites. Search results are ranked by the number of captures in the Wayback and the number of relevant links to the site's homepage.

Can I still find sites in the Wayback Machine if I just know the URL?

Why is the Internet Archive collecting sites from the Internet? What makes the information useful?

What is the Wayback Machine? How can I get my site included in the Wayback Machine?

Most societies place importance on preserving artifacts of their culture and heritage. Without such artifacts, civilization has no memory and no mechanism to learn from its successes and failures. Our culture now produces more and more artifacts in digital form. The Archive's mission is to help preserve those artifacts and create an Internet library for researchers, historians, and scholars. The Archive collaborates with institutions including the [Library of Congress](#) and the [Smithsonian](#).

What are the sources of your captures?**Do you archive email? Chat?****Why are some of the dots on the calendar page different colors?**

No, we do not collect or archive chat systems or personal email messages that have not been posted to Usenet bulletin boards or publicly accessible online message boards.

How does the Wayback Machine behave with Javascript turned off?**How can I get a copy of the pages on my Web site? If my site got hacked or damaged, could I get a backup from the Archive?'****How did I end up on the live version of a site? or I clicked on X date, but now I am on Y date, how is that possible?**

Our [terms of use](#) do not cover backups for the general public. However, you may use the Internet Archive Wayback Machine to locate and access archived versions of a site to which you own the rights. We can't guarantee that your site has been or will be archived. We can no longer offer the service to pack up sites that have been lost.

Where does the name come from?**Is there any personal information in these collections?**

We collect Web pages that are publicly accessible. These may include pages with personal information.

How do I cite Wayback Machine urls in MLA format?**Can I add pages to the Wayback Machine?**

On <https://archive.org/web> you can use the "Save Page Now" feature to save a specific page one time. This does not currently add the URL to any future crawls nor does it save more than that one page. It does not save multiple pages, directories or entire sites.

What is the Archive-It service of the Internet Archive Wayback Machine?**How do I contact the Internet Archive?****How can I help the Internet Archive and the Wayback Machine?**

All questions about the Wayback Machine, or other Internet Archive projects, should be addressed to info@archive.org.

Who has access to the collections? What about the public?**Where is the rest of the archived site? Why am I getting broken or gray images on a site?**

Broken images occur when the images are not available on our servers. Usually this means that we did not archive them.

How can I get pages authenticated from the Wayback Machine? How can use the pages in court?

You can tell if the image or link you are looking for is in the Wayback Machine by entering the image or link's URL into the Wayback Machine search box. Whatever archives we have are viewable in the Wayback Machine.

Some sites are not available because of robots.txt or other exclusions. What does that mean?

The best way to see all the files we have archived of the site is:
http://web.archive.org/*/www.yoursite.com/

There is a 3-10 hour lag time between the time a site is crawled and when it appears in the Wayback Machine.

What is the Wayback Machine's Copyright Policy?**Why are some sites harder to archive than others?**

If you look at our collection of archived sites, you will find some broken pages, missing graphics, and some sites that aren't archived at all. Some of the things that may cause this are:

- Robots.txt -- A site's robots.txt document may have prevented the crawling of a site.
- Javascript -- Javascript elements are often hard to archive, but especially if they generate links without having the full name in the page. Plus, if javascript needs to contact the originating server in order to work, it will fail when archived.

- Server side image maps -- Like any functionality on the web, if it needs to contact the originating server in order to work, it will fail when archived.
- Orphan pages -- If there are no links to your pages, the robot won't find it (the robots don't enter queries in search boxes.)

As a general rule of thumb, simple html is the easiest to archive.

Can I find sites by searching for words that are in their pages?

No, at least not yet. Site Search for the Wayback Machine will help you find the homepages of sites, based on words people have used to describe those sites, as opposed to words that appear on pages from sites.

Can I still find sites in the Wayback Machine if I just know the URL?

Yes, just enter a domain or URL the way you have in the past and press the "Browse History" button.

What is the Wayback Machine? How can I get my site included in the Wayback Machine?

The **Internet Archive Wayback Machine** is a service that allows people to visit archived versions of Web sites. Visitors to the Wayback Machine can type in a URL, select a date range, and then begin surfing on an archived version of the Web. Imagine surfing circa 1999 and looking at all the Y2K hype, or revisiting an older version of your favorite Web site. The Internet Archive Wayback Machine can make all of this possible.

How can I get my site included in the Wayback Machine?

Much of our archived web data comes from our own crawls or from Alexa Internet's crawls. Neither organization has a "crawl my site now!" submission process. Internet Archive's crawls tend to find sites that are well linked from other sites. The best way to ensure that we find your web site is to make sure it is included in online directories and that similar/related sites link to you.

Alexa Internet uses its own methods to discover sites to crawl. It may be helpful to install the free Alexa toolbar and visit the site you want crawled to make sure they know about it.

Regardless of who is crawling the site, you should ensure that your site's 'robots.txt' rules and in-page META robots directives do not tell crawlers to avoid your site.

What are the sources of your captures?

When you roll over individual web captures (that pop-up when you roll over the dots on the calendar page for a URL,) you may notice some text links shows up above the calendar, along with the word "why". Those links will take you to the Collection of web captures associated with the specific web crawl the capture came from. Every day hundreds of web crawls contribute to the web captures available via the Wayback Machine. Behind each, there is a story about factors like who, why, when and how.

Why are some of the dots on the calendar page different colors?

We color the dots, and links, associated with individual web captures, or multiple web captures, for a given day. Blue means the web server result code the crawler got for the related capture was a 2nn (good); Green means the crawlers got a status code 3nn (redirect); Orange means the crawler got a status code 4nn (client error), and Red means the crawler saw a 5nn (server error). Most of the time you will probably want to select the blue dots or links.

How does the Wayback Machine behave with Javascript turned off?

If you have Javascript turned off, images and links will be from the live web, not from our archive of old Web files.

How did I end up on the live version of a site? or I clicked on X date, but now I am on Y date, how is that possible?

Not every date for every site archived is 100% complete. When you are surfing an incomplete archived site the Wayback Machine will grab the closest available date to the one you are in for the links that are missing. In the event that we do not have the

link archived at all, the Wayback Machine will look for the link on the live web and grab it if available. Pay attention to the date code embedded in the archived url. This is the list of numbers in the middle; it translates as `yyyymmddhhmmss`. For example in this url `http://web.archive.org/web/20000229123340/http://www.yahoo.com/` the date the site was crawled was Feb 29, 2000 at 12:33 and 40 seconds.

You can see a listing of the dates of the specific URL by replacing the date code with an asterisk (*), ie: `http://web.archive.org/*/www.yoursite.com`

Where does the name come from?

The Wayback Machine is named in reference to the famous Mr. Peabody's WABAC (pronounced way-back) machine from the Rocky and Bullwinkle cartoon show.

How do I cite Wayback Machine urls in MLA format?

This question is a newer one. We asked MLA to help us with how to cite an archived URL in correct format. They did say that there is no established format for resources like the Wayback Machine, but it's best to err on the side of more information. You should cite the webpage as you would normally, and then give the Wayback Machine information. They provided the following example: McDonald, R. C. "Basic Canary Care." *_Robirda Online_*. 12 Sept. 2004. 18 Dec. 2006
[<http://www.robirda.com/cancare.html>]. *_Internet Archive_*.
[<http://web.archive.org/web/20041009202820/http://www.robirda.com/cancare.html>].
They added that if the date that the information was updated is missing, one can use the closest date in the Wayback Machine. Then comes the date when the page is retrieved and the original URL. Neither URL should be underlined in the bibliography itself. Thanks MLA!

What is the Archive-It service of the Internet Archive Wayback Machine?

For information on the **Archive-It** subscription service that allows institutions to build and preserve collections of born digital content, see <https://www.archive.org/about/faqs.php#Archive-It>

How can I help the Internet Archive and the Wayback Machine?

The Internet Archive actively seeks donations of digital materials for preservation. If you have digital materials that may be of interest to future generations, please let us know by sending an email to [info at archive dot org](mailto:info@archive.org). The Internet Archive is also seeking additional funding to continue this important mission. You can click the donate tab above or click [here](#). Thank you for considering us in your charitable giving.

Who has access to the collections? What about the public?

Anyone can access our collections through our website archive.org. The web archive can be searched using the [Wayback Machine](#).

The Archive makes the collections available at no cost to researchers, historians, and scholars. At present, it takes someone with a certain level of technical knowledge to access collections in a way other than our website, but there is no requirement that a user be affiliated with any particular organization.

How can I get pages authenticated from the Wayback Machine? How can use the pages in court?

The Wayback Machine tool was not designed for legal use. We do have a legal request policy found at [our legal page](#). Please read through the entire policy before contacting us with your questions. We do have a [standard affidavit](#) as well as a [FAQ section for lawyers](#). We would prefer that before you contact us for such services, you see if the other side will stipulate instead. We do not have an in-house legal staff, so this service takes away from our normal duties. Once you have read through our policy, if you still have questions, please [contact us](#) for more information.

Some sites are not available because of robots.txt or other exclusions. What does that mean?

Such sites may have been excluded from the Wayback Machine due to a robots.txt file on the site or at a site owner's direct request.

What is the Wayback Machine's Copyright Policy?

The Internet Archive respects the intellectual property rights and other proprietary rights of others. The Internet Archive may, in appropriate circumstances and at its discretion, remove certain content or disable access to content that appears to infringe the copyright or other intellectual property rights of others. If you believe that your copyright has been violated by material available through the Internet Archive, please provide the Internet Archive Copyright Agent with the following information:

- Identification of the copyrighted work that you claim has been infringed;
- An exact description of where the material about which you complain is located within the Internet Archive collections;
- Your address, telephone number, and email address;
- A statement by you that you have a good-faith belief that the disputed use is not authorized by the copyright owner, its agent, or the law;
- A statement by you, made under penalty of perjury, that the above information in your notice is accurate and that you are the owner of the copyright interest involved or are authorized to act on behalf of that owner;
- Your electronic or physical signature.

The Internet Archive Copyright Agent can be reached as follows:

Internet Archive Copyright Agent
Internet Archive
300 Funston Ave.
San Francisco, CA 94118
Phone: 415-561-6767
Email: info at archive dot org