The Deconstructed (or Distributed) Journal – an emerging model?

J. W. T. Smith

The Templeman Library, University of Kent, UK

Preamble

The reason for the rather disjointed title is that although I like the original title of my proposed publishing model (the Deconstructed Journal) it does sometimes distract from the actual form of the model. It is a distributed model, and I sometimes wish I had called it the Distributed Journal, but it might not have been such a memorable title.

Introduction

The original idea that developed into the Deconstructed Journal (DJ) model was floated at a meeting at the Royal Society in 1993.[1] It proposed a web site that contained subject-relevant information and/or pointed to other sites that contained further relevant information. This early model now exists in the form of subject gateways like the SOSIG social science Hub[2] or the EEVL engineering Hub,[3] both part of the JISC-funded Resource Discovery Network. Further thinking about what form of academic publishing was most suitable in a networked world led to a series of insights about the nature of academic publishing, what it was for and who it was for.

A series of insights

The means/ends confusion

When traditional journal publishers moved from being paper-oriented to network-oriented (or at least network-aware) they assumed that what they needed to do was simply move the journal from paper to e-form. This was a classic means/end confusion. The journal is a 'means', a way or method of achieving a goal or 'end'. In this case the end is a range of requirements or roles related to academic research and scholarship. The journal is not the academic publishing industry – it is the product of that industry. For a real revolution in academic publishing we should not be looking at replacing the journal *per se* but replacing the entire industry that produces it. This is what I mean by a new publishing model.

Lessons to be learnt from the current model

We can learn from the journal model before we discard it. It has evolved, and more importantly, survived, over centuries to play an important role (or roles) in the world of academic research. As I stated five years ago: 'any replacement must therefore play the same roles or, to rephrase it more strongly, it must satisfy the same needs.' (Smith, 1999a).

The roles of the journal

Following Smith (1999a) there are nine of these:

- *Editorial*. Subject selection with some quality control.
- *Quality control (content).* Carried out by the referees.
- *Quality control (form).* Copy editing, etc.
- *Conferring recognition of work done*. Editorial board and referees.
- *Marketing/making aware.* Marketing of the journal/ articles to possible readers.
- *Delivery/dissemination.* The delivery of the information (in the form of the physical item).
- *Subject definition.* A journal helps to define the areas it serves. It does this overtly through invited review papers and/or editorials and covertly by the papers it publishes or rejects.
- *Community definition*. Done through its readership.
- *Archiving (maintaining a record for posterity).* Strictly speaking, it is not a role of the journal to archive the results of work done. What the paper-based model does is produce a physical object that others archive.

Publishers not necessary?

Having analysed the roles played by the academic journal it becomes clear that: (1) the publisher, like the issue, is a product of the industry model required to produce the paper journal, and (2) publishing models without a central publisher are possible with net-based publishing

The message of the medium

The internet is distributed and non-hierarchical. Any node is conceptually equal to any other. It seems logical therefore that in order to make best use of its attributes any publishing model that uses it should be similar in structure or be such that it takes advantage of this structure.

The DJ model

Looking at the roles listed above it is clear that these roles are partly or wholly independent. Quality control is not dependent on distribution or vice versa. This intuition is reinforced by the idea that we want to make best use of the distributed nature and flexibility of the internet. The DJ model therefore proposes that each of the roles that need to be played to satisfy the needs satisfied by the paper-based journal can be played by an independent agent loosely co-operating with the other agents that satisfy the other needs. It is also clear that although there is some chronological order in the required co-operation, in many cases it is not necessary. For example, there is no reason why an item cannot be made available (suitably tagged) before quality control and recognition has taken place.

The nine roles listed above can be refined to four essential ones. The table below (taken from Smith, 2003) summarizes these and shows who plays these roles in the paper-based model and in the DJ model.

Quality control (content)	Referees, organized by publisher	Independent 'certification agents' or CAs ('evaluator organizations' in Smith, 1999a)
Conferring recognition of work done	Referees and journal editorial board	Independent 'certification agents' or (less directly) editorial boards of overlay journals,[4] ('subject focal points' in Smith, 1999a)
Making available	Publisher – printing the article in an issue and distributing it	Placing of material in local or centralized freely accessible electronic archives or repositories[5]
Making aware or marketing	Publisher – marketing of the journal to libraries and other customers	Overlay journals, general or specialized search engines, web directories, subject portals, weblogs.

Note that in order for the DJ model to exist there must be agents playing these roles and that they must be co-operating in the ways displayed by the model.

An emerging model?

The traditional academic publishing model has a co-ordinator (the publisher) organizing or bringing together the various sub-processes that constitute it. However, the DJ model (because it has no central co-ordinator) needs the spontaneous appearance and then the subsequent co-operation of its parts to come into existence. This may sound almost metaphysical, but it is the way all evolving systems work. For example, the eye did not pop into existence overnight. The various elements (light-sensitive cells, a nervous system, a data processing system, etc.) all evolved separately and eventually came together to form a proto-eye.

The same behaviour appears to be happening with regard to the development of the DJ publishing model. Many of the various elements needed to form the model are coming into existence, but not with the intention of being part of it. In this sense, the model is emerging or evolving from existing or innovative activities on the net. However, it cannot evolve directly from the current academic publishing model, because there are inherent contradictions.

The main problem is the idea of ownership by the certifying organization (publisher), which requires payment (subscription or 'pay per view') to allow access to the full text. This prevents the formation of virtual journals, which are an essential part of the DJ model. However, as we will see below, a variant of the traditional model, i.e. 'open access', could allow for this.

The current situation

This section considers what elements, or potential elements, of the DJ model already exist. We will look at each role and its required agency, using those listed in the table above.

Quality control (content)

Independent certification agents

As yet there are no fully independent certification agents. By 'independent' is meant providing certification separately from publication or 'making available'. There are journals that make articles they have published in e-form freely available either immediately or after a short period (6 or 12 months is common). BioMed Central (BMC)[6] makes the research articles in all of its e-journals freely available. Many conventional journals allow authors to place copies of final papers (the version as published in the journal) in e-print repositories and some authors do this even if the journal does not explicitly allow it.

It is assumed that certification agents (CAs) will operate using a payment model similar to that of the existing open access (OA) journal publishers. They will charge a fee for refereeing an article and attaching their 'seal of approval' to it. OA journals (including those of BMC) are discussed in detail below in *Making available*, while CAs are discussed in *What's missing?*

Conferring recognition of work done

Independent CAs

As noted above there are no examples of truly independent certification agents yet.

Editorial boards of overlay journals

By choosing to link to an article, the editors of an overlay journal are indicating they think the work 'cited' is of some value. However, using the current usual form of linking they can only point to an address (URL), they cannot guarantee that the item pointed to is the same one they originally chose. So the link is really saying: 'This is a good article (assuming it is the one we read when we made this decision).' This limits the extent to which they can confer recognition of work done. The problem of document integrity and authenticity in an electronic environment is discussed below in *What's missing?* and in the Appendix. Overlay journals are covered in greater detail below in *Making aware/marketing*.

Making available

Placing material in local or centralized freely accessible electronic repositories

There has been a steady growth of material in e-print repositories. Some have followed the Physics ArXiv[7] model; other subjects have invented their own, for example, CogPrints[8] (cognitive studies), NCSTRL[9] (computer science), RePEc[10] (economics) and PhilSci Archive[11] (philosophy of science).

Production of new e-print repositories has been made easier by the provision of free software to build them. Already a range of packages is available, for example, CDSware[12] from CERN, DSpace[13] from MIT, and Eprints[14] from the University of Southampton. All of these packages are OAI compliant (see *Making aware /marketing*).

Further impetus to the provision of e-print repositories was given by the Budapest Open Access Initiative[15] (BOAI) from OSI.[16] This promotes the use of open repositories and open journals to make the results of research freely available. Another recent promotion of the idea of OA repositories is the recent report from SPARC[17] in the US (Crow, 2002) which strongly promotes the idea of institutional repositories. Institutional repositories (IRs) are a variation on the basic theme. The repositories listed above all have a subject focus. IRs are operated by institutions (usually universities) and are intended to contain, and make freely available, all the research articles and reports they produce.

The value and possibilities of subject and institutional repositories is greatly enhanced by the output from the Open Archives Initiative[18] (OAI; see *Making aware/marketing*).

OA journals

The OA journal inverts the accepted 'subscriber pays' model of academic journal publishing and charges the author or author's employer a processing fee. The most fully developed commercial example of the OA journal model is employed by BioMed Central.

BMC currently hosts over 100 titles in the area of biosciences and medicine. It has titles of its own and also hosts journals being produced by independent groups of scientists. As with any conventional journal, all papers go through a selection and peer review process before they are published. There is an 'article processing charge' of around $500 per article accepted. This includes 'obtaining peer reviews and ... preparing the article for publication', the inclusion of a reference in PubMed, and archiving the article in PubMed Central.[19]

Another option is for institutions to join the institution membership programme, which enables all the relevant members of the institution to offer articles to BioMed Central for publication at no cost. In the UK, the JISC[20] (Joint Information Systems Committee), part of the Higher Education Funding Councils, paid a fee making all UK Higher Educational Institutions (HEIs) members of BioMed Central for an experimental period.

BMC is not the first publisher to try the OA model. The Institute of Physics has been publishing the *New Journal of Physics* using this approach since 1998.[21] The BOAI has published two detailed guides, explaining how to launch a new OA journal and how to convert a subscription-based journal to OA (Crow and Goldstein, 2003a, b). There was a further boost to OA publishing and the use of institutional repositories in the recent report from the House of Commons Science and Technology committee (House of Commons, 2004).

Making aware/marketing

Overlay journals

It is clear that there is a growing collection of freely available academic material either already quality certified or needing to be certified (or which would benefit for certification). This provides the target material for overlay journals which are discussed next.

The name 'overlay journal' comes (I believe) from a comment in Ginsparg (1996), where he discusses the possibility of information services provided as an 'overlay' within the Physics e-print archive. Such a service already existed in 1996 (Smith, 2000). An overlay (or virtual) journal is basically a list of evaluated and commented links to full text articles held elsewhere.

An excellent example of a working overlay journal is *Applications of Superconductivity*.[22] This title happily describes itself as a 'virtual journal' and it contains 'a multijournal compilation of developments in superconducting electronics, materials and largescale systems'. It shows exactly how an overlay journal can add value. In addition to links to relevant articles, it provides e-mail alerting of new items, the ability to search across the virtual journal and links to article supply services if the text you want is not freely available. Although it is currently free one can see how it could charge a small fee and be worth the cost. *Applications of Superconductivity* is one of a series of virtual journals *(Virtual Journals in Science and Technology*) developed jointly by the American Physical Society and the American Institute of Physics.

Weblogs

I see weblogs as precursors to new overlay journals, or even new forms of journal. At their most basic, weblogs are web pages containing lists of sites visited (hence 'web logs') with comments by the producer or editor. Their original form was like a diary recording interesting pages found. They are intended to be constantly updated with the latest addition being at the top of the list. There are variations on this format; for example, one might have a 'thought for today' approach with links on a theme embedded in a few lines or paragraphs of text, then the next day another theme would be explored.

A weblog might be devoted to a single theme with more and more links being added over an extended period. They have existed in their current form since 1997 (although some writers on their development claim the earliest web site listings produced by Tim Berners-Lee and others in the early 1990s were proto-weblogs) (Paquet, 2002). Such is the interest in producing weblogs there is now a site that offers comparisons of a range of weblog production tools.[23] Some writers have discussed the possibilities of weblogs for researchers (Paquet, 2002; Mortensen and Walker, 2002).

There have been a few articles discussing weblogs in the general library and information literature over the past few years, but no one appears to have spotted that weblogs have all the basic attributes of full-scale overlay journals. With very little (if any) modification one could take one of the weblog production packages and build a passable overlay journal quite quickly. As I pointed out five years ago (Smith, 1999b), almost all the genuine innovation in e-publishing has come from net users, not from the commercial publishing world. Also, end users often use tools designed for one thing for something the designers didn't envisage. When Tim Berners-Lee originally invented the web he was thinking of hyperlinked technical documents, not the web as we see it today.

Finally, it is interesting that weblogs started as online diaries or journals (in the original meaning of the word 'a record of the days activities').

RSS feeds

RSS is a method of making announcements about the content of one web site available for easy inclusion in another, or for reading using an RSS reader.[24] It could also be used for listing additions to a weblog or an overlay journal. If there were, say, three overlay journals that regularly listed items of interest and provided an RSS feed, you could have a reader program, e.g., FeedReader,[25] that regularly checked all three for new items.

General or specialized search engines and web directories/portals

OAI-PMH based services

A major step forward in the area of 'making aware' has been the Open Archives Initiative (OAI). Despite its name, the OAI is not directly about open access archives; it is about finding what is in institutional and other repositories, whether they are OA or not. What the OAI has produced is a protocol that enables the operators of a repository to make public the metadata describing the contents of their archive. This is in a standard format so others can harvest this information and build indexes that enable users to view the contents of a number of repositories, potentially all the repositories in the world that make their metadata available in this way.

This is known as the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). The OAI-PMH is a simple but profound idea. It allows a paper in the smallest college repository anywhere in the world to be as visible as one in a large university repository or major publisher's site. In addition to possibly enabling a revolution in academic publishing in the long term, this idea has major implications in the nearer term for researchers in the developing world (Smith, 2004; Chan and Kirsop, 2001).

Currently there are 183 OAI-compliant repositories. These are contributing records to 17 search service providers.[26] A particularly interesting service is provided by DP9[27] from Old Dominion University, which forms a link between traditional search engines and the contents of the OAI-compliant repositories, allowing the former to index the latter. So we may one day be able to search Google (and others) for academic articles.

Conventional search engines

Although we have all the new tools discussed above, there is a continuing need for the traditional search engines like Google and AltaVista. There are also more specialist services like Scirus[28] concentrating on specific areas of knowledge. In the future one can imagine specialist search engines so focused that they border on being new overlay journals. There will also be a continuing need for the general purpose directories like Yahoo and the Open Directory Project[29] as starting points for less focused searches. The specialist directories and subject portals like those that form the RDN (Resource Discovery Network)[30] will possibly move towards becoming overlay journals over time.

What's missing?

As can be seen from the previous section the main elements of the DJ model are beginning to form, not in order to satisfy the requirements of the model, but simply as outcomes of other activities.

I stated above that in order for the DJ model to work all of the main agencies (or elements) that form it have to be extant. We already have the independent repositories in the form of institutional and central open repositories and the software packages to build more, plus the OA journals. We have the beginnings of a mechanism to provide detailed search and retrieval services, with the OAI metadata harvesting protocol and the services being built using this standard. Overlay journals already exist as such or in proto-form as web directories or subject portals. We may find ourselves with a surfeit of overlay journals if weblogs develop as I suspect they might. The only major element that is missing is the independent CAs.

Independent certification agents

These are critical to the DJ model because without the separation of quality control from making available (publishing) you still have remnants of the traditional journal model with articles only available from a specific source. It has to be admitted that the model adopted by BMC almost escapes this criticism, as copies are deposited in PubMed Central. We still, however, have a partly centralized model.

Who could be a CA?

Any person or organization that can claim expertise in a subject and is respected for that knowledge could set up as a CA. Learned or professional societies have a head start in this. They already have the necessary reputation and their members have the expertise. Commercial organizations could do it by 'buying in' or otherwise organizing such expertise. This is what commercial publishers already do. They persuade recognized academics to sit on editorial boards of journals or act as referees for papers. So existing publishers could just move to become CAs. Clearly, there is no reason why independent CAs as required by the DJ model should not exist.

The 'seal of approval'

The ideal envisaged in the DJ model is that a document can be anywhere (including the possibility of multiple copies in more than one place) and the CA can be anywhere. What is needed is a mechanism whereby the CA can attach a 'seal of approval' to the document that guarantees this is a true copy and it was certified by this CA. Once we have such a mechanism the document can be placed anywhere on the net with no continuing connection to the CA.

This leads us to the problems of document 'integrity' and ' authentication'. There are existing solutions to these problems. Integrity can be guaranteed using a 'message digest' which is a almost unique fixed length string calculated from the contents of the document using what is known as 'a hash function'. The chances of two files, no matter how similar, having the same message digest are very low. This, combined with public key cryptography techniques, would give a digital signature (DS). These topics are covered in greater detail in the Appendix.

What is important is that there is an existing mechanism, the DS, which enables a CA to attach a seal of approval to a document. There is no technical reason why this final step in the emergence of the DJ model cannot happen.

The implications?

The main implication of all these developments is that it is possible that the current academic publishing model could dissolve into the many co-operating agents of the DJ model with no loss of functionality.

Jobs

Since there would be more small organizations, the DJ model could employ more people. There would be room for entrepreneurs to start small businesses and even invent new kinds of business.

Since the agencies would be relatively small, the need for senior managers would decrease and there would be no need for the very large publishing companies that appear to do well in the current academic publishing environment. However, the basic work would still need to be done, there would still be a need for copy editors, illustrators, etc.

Cost

The proponents of the open publishing model (author pays, reader access free) claim that overall it will be cheaper than the current model. However, since the two systems will be running in parallel and both will need paying for, there will be a period of extra cost. With a distributed model it might be possible to disperse some of the activities to agents who have no need to make a profit, or for whom the provision of a web server (for example) is at notional cost (because they already run one for other purposes). In this situation, further savings can be made on the overall cost of the industry.

Solutions for problems with the old model

Before we adopt a new way of doing things (or seeing things) that new way should not just be theoretically better; it should have practical advantages, or at very least solve problems inherent in the old model.

One of the major problems with the old model is that the publisher needs to ensure a return on the effort and cost required to publish the paper journal and thus needs to 'own' the copyright of the item. This means it cannot appear in another journal even though it might be relevant to more than one subject. Any open publishing model that has the equivalent of overlay journals can solve this. A similar advantage is that an overlay journal can be both new and have a history because it can point to earlier relevant articles. This was called 'full grown birth' in Smith (1999a).

Paper journals were limited in size; good papers could therefore be rejected on space grounds. This idea has carried over into the e-world, but any model of e-publishing can escape this limit. In the paper model or the paper-influenced e-model an article is either in or out, but with the DJ model CAs can rank rather than just say in or out.

Notes

1. 'E-Journals – Exchange of Experience Meeting', 26 February 1993, The Royal Society, London (organized by the BLR&DD).
2. www.sosig.ac.uk
3. www.eevl.ac.uk
4. The phrase 'overlay journals' has become the preferred name used by other writers in this area to indicate what were previously called 'virtual journals'. Overlay journal does describe the way in which they operate well, although I still like 'virtual journal' as I feel this best describes these agents.
5. The term 'archive' was for some time the usual one used to refer to collections of e-prints, although here it is not used in its sense of a store whose main aim is preservation rather than making items available. The term 'repository' is currently the preferred name for these collections.
6. See www.biomedcentral.com
7. See arxiv.org
8. See cogprints.soton.ac.uk
9. See www.ncstrl.org
10. See netec.mcc.ac.uk/RePEc
11. See philsci-archive.pitt.edu
12. See cdsware.cern.ch
13. See www.dspace.org
14. See software.eprints.org
15. See www.soros.org/openaccess/read.shtml
16. See www.soros.org
17. Scholarly Publishing and Academic Resources Coalition, see www.arl.org/sparc
18. See www.openarchives.org
19. See www.pubmedcentral.nih.gov
20. See www.jisc.ac.uk
21. See njp.org
22. See www.vjsuper.org
23. See BlogComp, www.urldir.com/bt/
24. For a simple introduction to RSS see searchenginewatch.com/sereport/article.php/2175271
25. See www.feedreader.com
26. See www.openarchives.org/service/listproviders.html
27. See arc.cs.odu.edu:8080/dp9/about.jsp
28. See www.scirus.com
29. See dmoz.org
30. See www.rdn.ac.uk

References

Chan, L. & Kirsop, B. (2001). Open archiving opportunities for developing countries: towards equitable distribution of global knowledge. *Ariadne*, 30 December. www.ariadne.ac.uk/issue30/oai-chan/

Crow, R. (2002). The case for institutional repositories: a SPARC position paper. www.arl.org/sparc/IR/IR_Final_Release_102.pdf

Crow, R. & Goldstein, H. (2003a). *Guide to Business Planning for Launching a New Open Access Journal*, Open Society Institute, Edition 1.0. www.soros.org/openaccess/pdf/business_planning.pdf

Crow, R. & Goldstein, H. (2003b). *Guide to Business Planning for Converting a Subscription-based Journal to Open Access*, Open Society Institute, Edition 1.0. www.soros.org/openaccess/pdf/business_converting.pdf

Ginsparg, P. (1996). Winners and losers in the global research village. Paper presented at the conference 'Electronic Publishing in Science', UNESCO HQ, Paris 19-23 February, 1996.

arxiv.org/blurb/pg96unesco.html

House of Commons Science and Technology Committee (2004). *Scientific Publications: Free for all?* Tenth Report of Session 2003-04 (HC 3991). www.publications.parliament.uk/pa/cm200304/cmselect/cmsctech/399/399.pdf

Mortensen, T. & Walker, J. (2002). Blogging thoughts: personal publication as an online research tool. Chapter 11 of the *Proceedings of SKIKT-Researchers' Conference: Researching ICTs in Context*, InterMedia, University of Oslo, 8 April 2002. www.intermedia.uio.no/konferanser/skikt-02/docs/Researching_ICTs_in_context-Ch11-Mortensen-Walker.pdf

Paquet, S. (2002). Personal knowledge publishing and its uses in research. radio.weblogs.com/0110772/stories/2002/10/03/personalKnowledgePublishingAndItsUsesInResearch.html

Schneier, B. (1996). *Applied Cryptography : Protocols, Algorithms, and Source Code in C*, 2nd ed., New York: Wiley.

Smith, A. (2000). The journal as an overlay on preprint databases, *Learned Publishing,* vol. 13, no. 1, pp. 43-48. www.ingentaselect.com/vl=747003/cl=42/nw=1/fm=docpdf/rpsv/catchword/alpsp/09531513/v13n1/s6/p43

Smith, J. W. T. (1999a). The Deconstructed Journal, a new model for Academic Publishing, *Learned Publishing,* vol. 12, no. 2, pp. 79-91.

library.kent.ac.uk/library/papers/jwts/d-journal.htm or www.ingentaselect.com/vl=11603875/cl=25/nw=1/fm=docpdf/rpsv/catchword/alpsp/09531513/v12n2/s3/p79

Smith, J. W. T. (1999b). Prolegomena to any future e-publishing model, in the *Proceedings of the ICCC/IFIP Conference on Electronic Publishing '99 – Redefining the Information Chain, New Ways and Voices*, Ronneby, Sweden, 10-12 May 1999, pp. 293-298. library.kent.ac.uk/library/papers/jwts/Prolegomena.htm

Smith, J. W. T. (2003). The Deconstructed Journal Revisited – a review of developments. *ICCC/IFIP Conference on Electronic Publishing – ElPub03 – From Information to Knowledge*, Universidade Do Minho, Guimarães, Portugal, 25-28 June 2003, pp. 2-88. library.kent.ac.uk/library/papers/jwts/d-jrevisited.htm

Smith, J. W. T. (2004). The importance of access to academic publications for the developing world and the implications of the latest developments in academic publishing. Invited paper presented at the *International Conference on Computer Communication: Core Platform for the Implementation of the Computer Society*, 15-17 September 2004, Beijing, China.

library.kent.ac.uk/library/papers/jwts/develop.htm

Appendix: Document integrity and authentication

If you print out a page of an article and put it in a drawer for a year you can be reasonably sure that it will still be there and readable when you look again although the ink may fade and the paper become brittle. One thing you can be absolutely certain about is that the words will not move around the page or some of them disappear without trace or be replaced by others.

This is not true with electronic documents. They are just computer files and can easily be altered intentionally or unintentionally. The integrity of computer files (and hence electronic documents) has always been a problem. It is less of a problem as long as the document stays on the same computer, because it is possible to track any changes and be sure that a file has remained unchanged in terms of content even if its physical representation has changed. However, once the document is made available on the network and can be downloaded to other computers this basic certainty is lost.

Fortunately, there are ways to ensure integrity of the contents of a computer file (and hence an electronic document). One of these is to use what is known as a 'one way hash function' to compute information about the file which can be used later to see if the file content has changed. Any hash function takes an input string of a variable length (like a file containing an electronic document) and returns a fixed length string which is usually

much shorter. A one way hash function takes this a step further such that it is very hard to reconstruct the original string given the fixed length string. It is also very hard to construct another input string that hashes to the same output string.

The output string is given a range of names, e.g. message digest, fingerprint, cryptographic checksum, or message integrity check. The most commonly used name seems to be 'message digest'. For a detailed description of this (and other related techniques) see Schneier (1996). Hash functions are not secret, so given a file, the message digest and the name of the hash function used to produce the original message digest, it is possible to re-calculated the message digest of the file you have and compare it with the one given with the file. If they are the same, you can be sure the file you have is identical to the original.

So now we have a way of ensuring that the file is unchanged. How can we be sure the sender is who they claim to be, or, in our case, that this is the file certified by the relevant CA? One way to do this is to use a Digital Signature (DS). The use of a hash function as described above to check the integrity of a file is the first half of a DS. A DS also uses public key cryptography (PKC) to ensure the authenticity of a message by ensuring the sender (or the person or organization who 'signs' the message) is who they claim to be.

With PKC there are two encryption keys, one private and one public; a message encrypted with the one has to be decrypted with the other. This has the added advantage that only the sender knows the private key and the public key only decrypts messages encrypted with the matching private key. So you can be sure that if someone's public key decrypts a message it must have been sent (or encrypted) by them.

We could prove both the integrity and authenticity of a message (or document) by encrypting the whole thing, but encryption and decryption are computationally expensive and so a DS combines the use of a hash function with PKC to make it easier.

The procedure is as follows. A message digest is calculated for the document; this is encrypted using the sender's private key. The document and digest are bundled together, for example in an e-mail message. The recipient takes the document and calculates the message digest, then finds the public key for the sender and decrypts the accompanying message digest. If the two message digests are the same, this is the document sent (or certified) and the sender (or certification agent) is who they claim to be.

Simple, isn't it? Unfortunately, it isn't.

Although the elements that enable DSs to work are all known, there appears to be no agreed standard for how they are put together. It is possible to buy DS programs that run on PCs which automatically do the calculations and encryption and package up the file ready to send or to be downloaded. However, the recipient has to have the same software for the unpacking and verification to be done automatically. It is as if it was agreed that all cars have to have a steering wheel and brakes (and also agreed how these things work) but there is no agreement on which side the steering should be on or whether the brake is the middle or left pedal.

Any competent computer scientist could carry out the process. I am assured it is not particularly difficult, but we are not all computer scientists. It is possible that in time commercial packages will converge on a common standard, at least to the extent that someone using one DS program will be able to accept and process a file processed and packaged by another. Maybe what we need is an initiative similar to the OAI, which designs a simple standard sufficient for academic publishing needs.

There may also be simpler ways to achieve our goal. Since all we want is to be sure that the document we have is the one originally certified we could just calculate the message digest, send this to the claimed certification agent (or a secure site that maintains a list of certified documents). The CA would return the title (or bibliographic record) of the document. This may not be as secure as using a DS, but we are not dealing with national secrets or sensitive personal information.

Contact

J. W. T. Smith
The Templeman Library
University of Kent
Canterbury
Kent CT2 7NU
UK

J.W.T.Smith@kent.ac.uk
www.library.kent.ac.uk