# A networked registration scheme to support open science

J Adrian Pickering, Christopher J Gutteridge and David De Roure
*School of Electronics and Computer Science, University of Southampton, Southampton, SO17 1BJ, UK*
*jap@ecs.soton.ac.uk*

## 1. Introduction

The Open Source and Open Science movements have demonstrated the success of distributed collaborative experimentation and intellectual property (IP) development. While those contributing to the effort may do so without seeking to secure IP rights, it is clear that credit and attribution are crucial to the scholarly lifecycle because they underpin reputation – when IP is created it is only fair that 'credit is given where credit is due'. We propose that there need to be systems in place, independent of the project, where the evidence of 'prior art' can be registered. The authors' thesis is that simply having such a system available will ensure proper behaviour between collaborators and foster higher productivity.

Repositories such as EPrints and myExperiment, which focus respectively on publications and digital 'research objects' [1], can readily use such a system – the intellectual assets stored digitally in the repository can be registered by their owners. To achieve this with the necessary guarantees we need an appropriate registration scheme and architecture.

## 2. Current schemes

A key concept in doing business is that each party declares its position in a way that is non-repudiatable (NR). The subsequent action by the counterparty is constrained if it knows that NR techniques are in use. Proper behaviour is induced because of the implicit threat of action using robust, NR evidence. This could be by laying the evidence before peers or, in an extreme situation, taking legal action. In the past this state would be captured on paper and openly marked (signed) by those bound by its content. When the parties become virtual and timeliness is crucial, some other mechanism is required. Maintaining trust in these circumstances is the challenge. Today, experimental data is extensive and electronic and, when created, it is not clear what elements may be important. The 'log book', electronic or paper, is not suitable anymore. There needs to be a robust and economical way of enabling NR over this type of data.

The idea of using cryptographic hashing algorithms to assist time-stamping a digital document was published by Haber and Stornetta [3]. The motivation was the ability to easily declare the existence of a document to a third party without disclosing its content. Cryptographic hashing algorithms are designed to produce a short 'digest' (or 'hash') of a digital document (file) which is (a) collision-free i.e. no two documents will generate the same digest and (b) it is infeasible to synthesise a collision i.e. generate another document that has the same digest as another. Since the digest is, in general, shorter than the original document, there is less information there than in the original. Together with the nature of the algorithm, this means nothing can be construed about the original document from its digest. The accepted cryptographic hashing algorithms are public and are continuously subject to scrutiny by cryptanalysts since they underlie electronic signing mechanisms.

The principles have been used in a number of registration systems, notably the digital notary service operated by Surety [6], which is based upon Haber and Stornetta's concept and patents. Also, since 1995 a UK Jersey-based company has been operating its 'Stamper' time-stamping service based on PGP signing (IT Consulting [5]). More recently, in the UK, Codel have been promoting their Codelmark service [4].

The Surety and Codel services are both subscription-based services. Subscribers need to have faith in the company and trust that their processes are rigorous since the underlying registrations are not open to users' scrutiny. Daily, they digest the registration data and publish the resulting 'master hash' in a newspaper of record (Codel publish in the Financial Times). 'Stamper' attempts to be more open by publishing its signing summaries on the web and over Usenet.

Since the users' digests are not disclosing anything, it is not clear why it should not be possible to openly declare the registrations. This reduces the trust barrier to using the system: the users can see their registrations, their context and watch the scheme function. Further, if they are concerned about the robustness of the service, they can take copies of sufficient data for safe-keeping elsewhere.

Though the digests do not declare anything interesting about the user, other data recorded with the digest could. Subscription services need to know whose data it is in order to secure their income stream. They can undertake traffic analysis on registrations and assign it to users. Even if this data is not made public, the registration service needs to be trusted not to ever misuse this data. The only sure way of avoiding this risk is to allow anonymous requests for service.

Thus there is a need for an openly-available registration scheme whose only function is to accept and publish sequences of digests received from anonymous users. If the user needs to assert that they, or their company, are the only owners of the data at the time, then it is for them to incorporate some secret in the data before its digest is registered (e.g. using a signed HMAC, Eastlake and Hansen [2]). That is an optional, separate issue from registering the existence of the data.

## 3. The registration scheme

Open declaration of evidence is optimal. Digest hashes are of fixed length, short and are not expensive to store. Publishing these is cheap and web technology provides the ideal 'notice board' where the public can observe them.

Anyone with an electronic document that they wish to register posts its hash on a registration server of their choice using a protocol that supports anonymity. Because of the properties of cryptographic hashing algorithms, the registrand should be able to demonstrate later that only they had the means to create the registration at that point in the journal. The registration server's vital task is to journal the registrations chronologically. The server can annotate the registrations with other data. An obvious and useful choice is a timestamp. However, such timestamps are only indicative: they are not essential. The scheme fixes the time order of registrations, which is necessary and sufficient.

The power and scaleability of the scheme lies in realising that a registration server is itself generating material that needs evidential protection. Thus, periodically, it hashes a journal segment and registers that with another disinterested server. Provided there are sufficient, randomly cross-registering, independently operated servers, the time order of registrations across the server network can be adequately resolved. Any timing information embedded in the journals is useful, supplementary evidence, particularly if the timestamps (or other form of time anchor) come from reputable sources.

As an example, Figure 1 illustrates what happens when a user registers 'My Document' with their chosen 'Yellow' server. The hash the user generated for My Document is stored with other registrations in Segment 6 of Yellow's journal. It returns a receipt to the user that is kept with My Document. Shortly later, Yellow closes its Segment 6 and registers its hash with itself in new Segment 7 and the Green server. This process continues at indeterminate times among other, disinterested, cross-registering servers. Note that the effect of the registration (network 'flow', in graph terms) is felt twice in Yellow Segment 7. This achieves stronger connectedness between the variously-owned segment nodes and distinguishes this scheme from those that use hash trees to store registrations e.g. [6].

If the user wishes to assert they were in possession of My Document at the time claimed the registration process is replayed but comparing the registration with the evidence in Yellow server Segment 6, as indicated by the saved receipt (see Figure 2). For clarity, the illustration shows the user making just one registration with one server. For robustness, it is wise to register with at least two servers.

Once the hash of a server journal segment is released to the network it will rapidly be subsumed within further cross-registrations. This locks all the dependent data into time order. No other data on which the hash depends can be altered without potential detection.

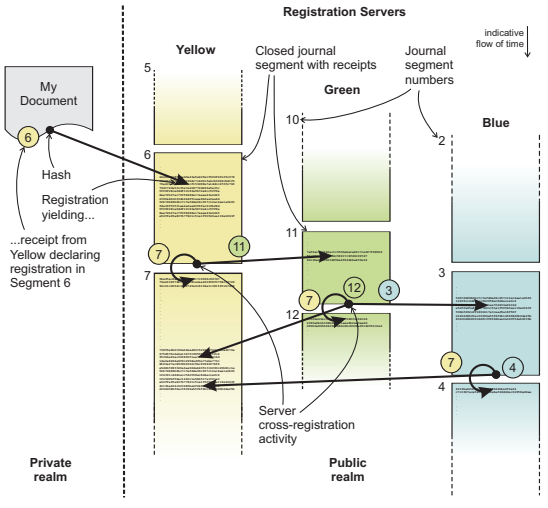


**Figure 1. Registering among net-connected servers**

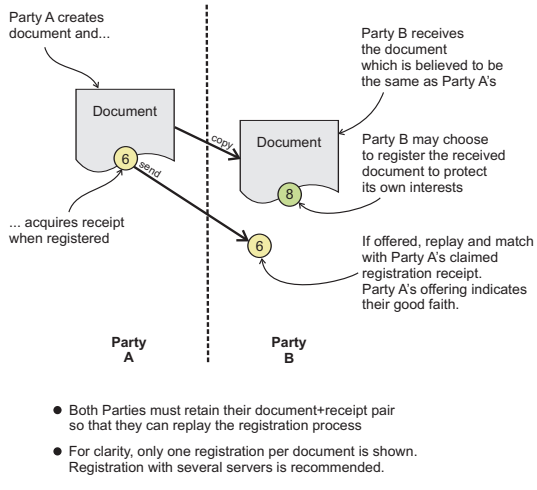


**Figure 2. Confirmation of the state of a document**

Many, unknown users—notably, the registrands, will observe the scheme. Further, these will be taking copies of relevant segments so that they (or their representative) can replay the algorithms later if required. Server operators will not know their users through normal use. Any inexplicable post-hoc alteration or loss of journals will cause potentially irreparable damage to their reputation.

Since there must be formal disinterest between servers and clients, there cannot be any service contract and, therefore, the service must be free. This poses some security and resourcing challenges. Fortunately, hashes are small, and networks and storage are comparatively cheap. We also have precedents for the evolution of large, mutual self-interest based systems—Internet and email.

The holder of some potential evidence would now have complete freedom to register its hash with servers of their choosing. Since services might disappear, it is wise to register with several. Service operators will have a service policy that would guide a user in their choice.

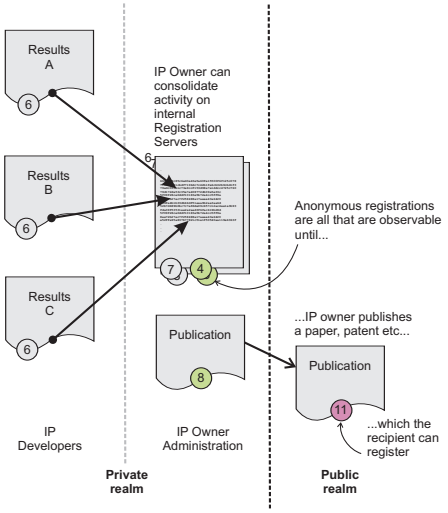


**Figure 3. Gathering IP prior-art evidence before publication**

Figure 3 shows how a research organisation (e.g. a university) can run a local registration service to concentrate the registration traffic before using a third party service. The advantages of doing this are (a) encouraging the IP developers to use the service regularly and (b) mitigating traffic analysis by outside parties. Because the primary server is internal to the organisation, its evidential robustness would be more in question. However, the IP Administration can dynamically adapt its own registration policies to suit the risks.

It is always the users' responsibility to have the means to replay the registration algorithms to a third party to prove that they possessed the data at issue at the time claimed. This is why it is important that the scheme is open and simple. In a dispute, it is still for the parties to interpret the meaning the data that was registered. What will not be in doubt is (a) document possession and (b) its time-order context.

## 4. Conclusion and next steps

An open scheme that should enhance the ability of citizens or workers to collaborate with trust is proposed. The time is right for its widespread adoption so that it can start to enable the benefits claimed. A very simple demonstrator can be accessed at Probity[1] [7] which readers are invited to try. This uses HTTP to effect registrations.

Work is in progress in developing open-source client and server prototypes to be used within the UK e-Infrastructure as part of the broader architecture used by both EPrints and the myExperiment project. In its next phase, myExperiment is further developing the notion of Research Objects and also exploring the core services required for future e-Laboratories. There is scope for 'added value' in embedding the registration primitives within tools where NR could be useful. Those with server resources and interest are invited to join the effort and get the scheme working for the research communities' benefit.

## 5. References

[1] D. De Roure, C. Goble et al "Towards Open Science: The myExperiment approach" Concurrency and Computation: Practice and Experience (in press)
[2] D. Eastlake and T. Hansen, "US Secure Hash Algorithms (SHA and HMAC-SHA)" IETF RFC 4634, July 2006
[3] S. Haber and W.S. Stornetta, WS "How to time-stamp a digital document", *Journal of Cryptography*, Vol 3 No 2, 1991, pp 99-111
[4] R. Hill, "The Codel Authentication System" Codel Technical White Paper, Codel Ltd. *www.codelmark.co.uk/white-papers*, 14 February 2007
[5] IT Consulting, PGP Digital Timestamping Service *www.itconsult.co.uk/stamper/stampinf.htm*, 2002
[6] Surety, *www.surety.com,* checked 2009
[7] Probity, Demonstrator at *www.probity.org/demo,* 2009

[1] **P**ublic **R**egistration **O**n the we**B** of **I**ntellectual proper**TY**