PKI without tears

Stephen Wilson

This article was first published in the American Bar Association's eBlast journal, Jan 2003

Abstract

Traditional Public Key Infrastructure (PKI) is unnecessarily complicated. Largely as a result of early misconceptions that we needed an all-purpose digital passport to do business on the Internet, traditional PKI has become overloaded with invasive personal identity checks and complex legal arrangements. To make things worse, early software implementations brought out explicit details of digital certificates, necessitating unusually intense user training. To try to support stranger-to-stranger transactions, user agreements for general purpose certificates have required people to read and understand huge and forbidding Certification Practice Statements. And yet the business benefits of going to all this trouble remain controversial.

Most of the burden of orthodox PKI derives from trying to create the all-purpose digital identity. In day-to-day personal commerce, this is famously analogous to a drivers licence, but in the professions and in business, a single identity is uncalled for and unprecedented. PKI tends to deliver its greatest benefits—automatic paperless processing, reduced legal risk, lower cost of dispute resolution—in high value, high volume, specialist applications, where digital personae are application-specific.

There are new PKI models where the cryptography is embedded deeply into smartcards, to much the same extent that complex ferromagnetic technology is built into all the other plastic cards we take for granted. Application software can be engineered so that all digital certificate functions are automated; smartcards can be issued to professionals and business people under existing terms and conditions which reflect the users' standing. The user experience then becomes the same as with any conventional access card. We can do away entirely with the need to read and understand complex Certification Practice Statements and Policies, sign up to unusual Subscriber and Relying Party agreements, or undergo esoteric technical training. Thus the underlying PKI becomes true infrastructure, used purely to automate paperless transactions between parties who are already accustomed to dealing with one another.

This paper presents a fresh look at the business drivers and true benefits of digital signatures, and shows how application-specific PKI can deliver the benefits with better usability, zero registration overhead, reduced training costs, simpler liability arrangements, and streamlined accreditation. The paper is aimed at regulators, policy analysts and e-business strategists with an interest in the future of PKI.

About the author

Stephen Wilson, Chief Security Specialist at SecureNet in Australia, is a leading international authority on Public Key Infrastructure and information security. Stephen is a member of the ABA-ISC and the APEC e-Security Task Group, and is past chairperson of the Certification Forum of Australasia.

The shift from electronic passport to electronic business card

Since the mid 1990s we have seen major changes in the way Public Key Infrastructure (PKI) is applied to e-business. In PKI's early conceptions, digital certificates were proposed to authenticate non-descript transactions between parties who had never met before. Crucially, certificates were construed as the *sole means* for people to authenticate one another. Rarely, if ever, have traditional PKI formulations included any other context to a hypothetical electronic transaction that might help its receiver decide whether or not to accept it. The digital certificate was predicted to be your all-purpose digital identity; no other context was thought to be necessary.

Orthodox PKI has come in for fierce criticism. Many commentators have pointed to a stark paradox: online transaction volume and value are increasing rapidly, in almost all cases without the aid of PKI. Some find the orthodox proof of identity to be intrusive; others have lampooned the idea of forming new Internet contracts in reliance on digital signatures. The one-size-fits-all electronic passport has certainly failed to take off, yet PKI's critics frequently throw the baby out with the bathwater. They fail to imagine that registration processes, digital signature software, and governance models can all be radically improved.

Perhaps inevitably, in the absence of any specific context for its application, orthodox PKI emphasises proof of personal identity. Early certificate registration schemes simply co-opted familiar identification conventions like the intuitively appealing passport. In Australia, the "100 point check" of the Financial Transaction Reports Act 1988—where the applicant must furnish a number of identification documents such as birth certificate and drivers licence—became the de facto registration rule. Yet very few traditional business transactions require parties to sight one another's passports or other personal documents. The 100 point check bears little resemblance to the way we authenticate one another in regular business transactions. The requirement for PKI users to submit to strenuous personal identity checks over and above their normal business credentials is a major obstacle in the adoption of digital certificates.

Another impediment to adoption has been the legal complexity traditionally associated with PKI. Most users are given little comfort by typical PKI services and schemes as to their legal position. For instance, the Commonwealth Government's best advice to PKI users is that the legal relationships between Subscriber and Relying Party, and between Relying Party and the CA, are "unclear in Australian Law". This position is the outcome of two successive legal studies commissioned by the National Electronic Authentication Council (NEAC), both of which were inconclusive regarding liability in a general purpose PKI. These were sound, well researched reports, yet their terms of reference had digital certificates as the *sole means* of authentication, with no prior relationship between any of the parties and no other context to their transactions. It is not surprising that liability was difficult to pin down under such sparse and artificial circumstances.

It turns out that the "killer applications" for PKI overwhelmingly involve transactions with very specific contexts between parties acting with clear and formally defined authority. These parties might not know each other personally, but invariably they recognise and actually anticipate each other's qualifications, befitting their business relationship. As we shall see below, contemporary usage of PKI is characterised by closed communities of interest, prior out-of-band registration of members, and in many cases, special purpose application software featuring additional layers of security and access controls.

So digital certificates are much more useful when implemented as application-specific "electronic business cards", than as one-size-fits-all electronic passports. And by taking account of the special conditions that govern different e-business applications, we have the opportunity to greatly simplify the registration processes and liability arrangements that go with PKI.

The real benefits of digital signatures

There is a range of potential benefits of using PKI, including its cryptographic strength and resistance to identity theft (when implemented with private keys in hardware). Many of its benefits are shared with other technologies, but at least two are unique to PKI:

1. ***Digital signatures provide robust evidence of the origin and integrity of electronic transactions, persistent over time and over "distance", greatly simplifying audit logging, evidence collection and dispute resolution, and cutting the future cost of investigation and fraud.***

   If a digitally signed document is archived and later checked, the quality of the signature remains undiminished over many years, even if the public key certificate has long since expired. And if a digitally signed message is passed from one Relying Party to another and on to many more, passing through all manner of intermediate systems, everyone receives an identical, verifiable signature code with which to authenticate the message.

   Electronic evidence of the origin and integrity of a message can of course be provided by means other than a digital signature. For example, the authenticity of typical e-business transactions can usually be demonstrated after the fact via audit logs, which indicate how a given message was created and how it moved from one machine to another. However, the quality of audit logs is highly variable and it is costly to produce legally robust evidence from them. Audit logs are not always properly archived from every machine, they do not always directly evince data integrity, they are not always readily available down the track, they are rarely secure in themselves, and they usually need specialists to interpret and verify them.

   Digital signatures on the other hand make it vastly simpler to reconstruct and if necessary re-wind transactions, essentially anytime after the fact. As online fraud steadily rises, electronic service providers are looking to PKI to cut their systemic cost of investigation, forensics and dispute resolution.

2. ***Digital signatures and associated digital certificates are machine readable, allowing the credentials or affiliations of the sender to be bound to the message and authenticated automatically when received, enabling totally paperless transacting.***

This is an important but often overlooked benefit of digital signatures. By processing a digital certificate chain—including checking CRLs, Policy Identifiers and other extensions—Relying Party software can *automatically* tell:

i. that the message had not been altered since it was originally created
ii. that the sender was authorised to launch the transaction, by virtue of credentials or other properties endorsed by a recognised CA
iii. that the sender's credentials were valid at the time they sent the message
iv. that the authority which signed the certificate was accredited to do so.

One reason many overlook machine readability is that they have come to expect person-to-person e-mail to be the archetypal PKI application, thanks to e-mail being so often used by vendors to illustrate PKI in action. There is an implicit suggestion in much PKI marketing and training that *in regular use* we should manually click on a digital signature icon, examine the certificate, check which CA issued it, read the Policy Qualifier, and so on. Yet the overwhelming experience of PKI in practice is that it suits special purpose and highly automated applications where the receiver of signed transactions is actually a computer.

Characterising good applications for digital signatures

Understanding the basic benefits of digital signatures allows us to characterise the types of e-business applications that merit investment in PKI. Applications for which digital signatures are a good fit tend to have the following features:

- Reasonably high transaction volume
- Fully automatic processing (or straight through processing) of transactions
- Multiple recipients, or multiple “hops” between sender and ultimate receiver
- Significant risk of dispute or legal ramifications, necessitating high quality evidence to be maintained over long periods of time.

This fresh view of the technology helps to explain why many first generation applications of PKI were problematic. Retail Internet banking is a well known example of e-business which so far has flourished without digital certificates. A few banks did try to implement certificates, but generally found them difficult to use, for an uncertain improvement in security at the time; most later reverted to more conventional access control and back-end security mechanisms. Yet with hindsight, retail funds transfer transactions don’t have a great need PKI, since they can make use of existing back-end payment systems. Funds transfer is characterised by tightly closed arrangements, a single Relying Party, built-in limits on the size of each transaction, real-time or near real-time settlement, and well defined audit trails. A Threat and Risk Assessment would show that access to Internet banking can rest on simple password authentication, in exactly the same way as antecedent phone banking schemes do.

The analysis suggests that the following will be good applications for PKI:

- tax returns
- customs reporting
- e-healthcare
- financial trading
- insurance
- electronic conveyancing
- superannuation reporting
- patent applications.

Trading off *Complexity* against *Applicability*

Most of the overhead in orthodox PKI comes from mis-treating the technology as a general purpose proof of identity. As discussed, orthodox PKI is constructed around the tacit assumption that there is no specific context for the transactions it is intended to support, and that the digital certificate is the sole means for authenticating the sender. Consequently, the traditional schemes emphasise high standards of personal identity, exhaustive contracts, and unusual legal devices like Relying Party Agreements. They can also resort to arbitrary "reliance limits", which have little meaning for non-payments transactions the likes of which dominate the contemporary PKI applications listed above. Notoriously, traditional PKI contracts require users to read and understand Certification Practice Statements, and sign up to obscure undertakings to safeguard their private keys.

All this overhead stems from not knowing what the general purpose digital certificate is going to be used for. On the other hand, if particular digital certificates are constrained to defined applications, then the complexity surrounding their specific usage can be radically reduced.

Consider the American Express Blue credit card, a new chip-enabled credit card. When you sign up for an Amex Blue card, you agree to regular credit card terms and conditions; that is, you undertake to not reveal your PIN to others, not to let anyone else use your card, to promptly report its loss, and so on. You are not required to read a lengthy "Certification Practice Statement" (CPS); nor do the Ts&Cs impose novel requirements like safeguarding your private key. The Amex Blue card’s underlying PKI imposes no additional burden on card holders whatsoever.

The trade-off for this dramatic simplification is that any Amex Blue digital certificate is constrained in its application within a well defined scheme. For instance, it could not be used to sign or encrypt generic e-mails, nor to authenticate the client in generic SSL connections. It is likely that in future, only software applications approved by American Express will be able to access the digital id functions in the Blue card.

From this experience we can abstract a more powerful, generalised meaning of a digital certificate. Rather than making representations about someone’s personal identity, a digital certificate can stand for the holder’s membership of some defined community, such as a group of credit card holders, registered medical practitioners, chartered accountants, or even the board of directors of a company. Each community will have an associated class of e-business applications, with Ts&Cs to match.

Contemporary usage of PKI is context rich

The role of PKI in all contemporary “killer applications” is fundamentally to help automate the online processing of electronic transactions between parties with well defined roles and credentials. This is in stark contrast to the way PKI has historically been portrayed, where strangers Alice and Bob use their digital certificates to authenticate context-free general messages, often presumed to be sent by e-mail.

In reality, serious business is never conducted stranger-to-stranger in the complete absence of context and cues as to the parties’ legitimacy. Using generic e-mail to convey a business message would be like sending a fax on plain paper to someone you’ve never met before. Instead, serious business is usually highly structured:

- Parties have an expectation that only certain types of transactions are going to occur between them and they equip themselves accordingly (for instance, a Medicare office is not set up to handle like tax returns).
- The sender is authorised to act in defined transactions by virtue of professional credentials, a relevant licence, affiliation with an employer or other authority, and so on. And the receiver recognises the source of the sender’s credentials.
- The sender and receiver typically use prescribed forms and/or special purpose application software with associated user agreements and licence conditions, adding context and additional layers of security around the transaction.

When PKI is used to help automate the online processing of transactions between parties in the context of an existing business relationship, we should expect the legal arrangements between the parties to still apply. For business applications where digital certificates are used to identify users in specific contexts, the question of legal liability should be vastly simpler than it is in the general purpose PKI scenario where the issuer doesn’t know what the certificates might be used for.

Comparing orthodox and contemporary PKI models

While orthodox PKI has proven difficult to implement and use, many of its underlying elements should be preserved as we move to a more flexible model. In particular, most of today's standards (like X.509 and RFC 2527), commercial RA/CA products, and backend CA services can be re-applied with little or no change.

To illustrate, the following two diagrams compare and contrast the traditional PKI model, where general purpose identity certificates are supplied over the counter, with the more contemporary model, where certificates are embedded into applications and managed as part of a broader scheme.

As shown in Figure 1, it has been traditionally assumed that each user would apply in person to a Registration Authority for their general purpose certificate, supplying passport-strength evidence of identity, and signing a Subscriber Agreement. The archetypal certificate application is person-to-person e-mail, where receiver Alice is expected to examine the certificate of the stranger Bob, and ascertain for herself Bob's veracity. The scope of PKI accreditation or licensing typically encompasses just the RA and CA; in particular, it usually ignores any specific applications or context for the certificates, or additional controls that govern their usage.

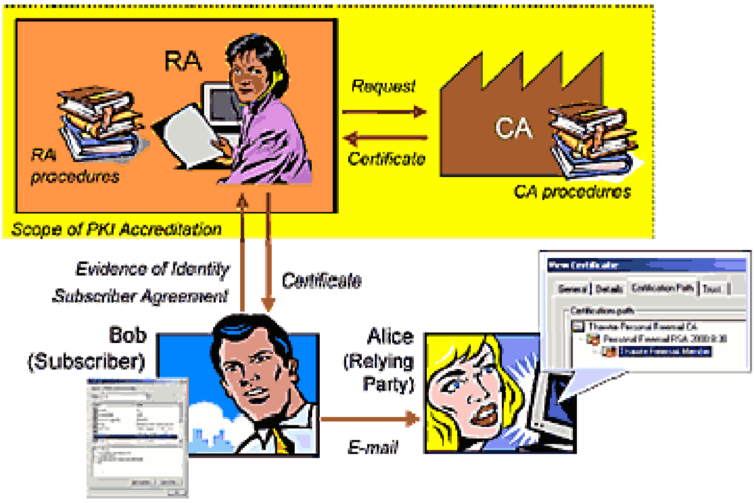

Figure 1: Orthodox PKI

When certificates are embedded in smartcards, the PKI can look like Figure 2. In this case, user Bob is a member of some community of interest and subject to its membership provisions and other scheme rules. As a current member, Bob can be sent a smartcard from the scheme's administrator, more or less automatically. Such smartcards are produced as per a conventional PKI, by a backend Certificate Authority and smartcard provisioning bureau.

Depending on the scheme, the smartcard might work for instance as a purchasing card, a business licence, a professional membership token, or an employee card. In each case, Bob uses his card to access associated e-business software, happily unaware of the embedded digital certificate and underlying PKI. Typical functions include healthcare transactions, statutory B2G reports (like securities commission returns), purchase orders and so on, received and processed usually by machine. The scope of PKI accreditation or licensing should now encompass not only the RA and CA but also the intended use of the smartcard.

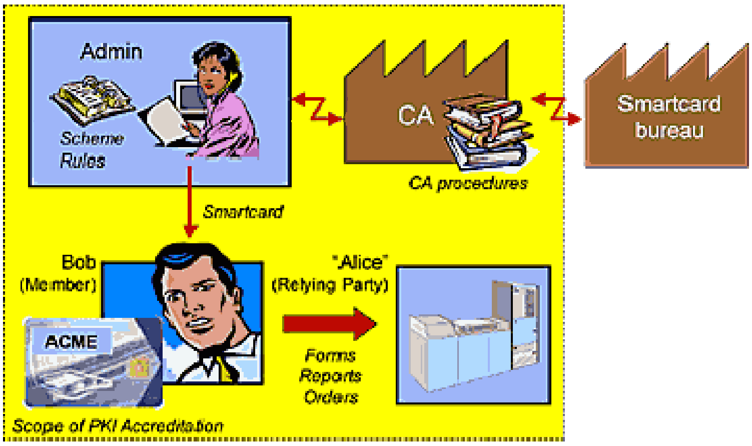

Figure 2: Contemporary PKI

To summarise, Table 1 compares and contrasts the two models.

	Early PKI circa 1998	**Contemporary PKI c. 2003**
Metaphor	Electronic passport	Electronic business card
Meaning	Personal identity	Credentials and/or affiliation
Value proposition	"Non-repudiation"	Persistent identity
Intended use	General purpose, non-descript stranger-to-stranger e-business	Special purpose, well defined applications between parties with well defined roles and credentials; NB: receiver is often a machine, not a natural person
Communities of Interest	General public	Professions (e.g. doctors, pharmacists, lawyers, accountants), business licence holders (e.g. customs agents, stock brokers, real estate agents, conveyancers), employees etc.
Implementation	Explicit keys & certificates; single all-purpose certificate separate from applications	Embedded keys & certificates; multiple certificates, specific to and bundled with applications
Registration process	Explicit, based on strenuous evidence of personal identity	Implicit and automated, based on existing membership status and rules
Regulatory environment	Crypto export restricted; defence agencies prominent in policy formulation; presumed impact on the national interest	Crypto export relaxed; defence agencies less concerned.

Table 1: PKI now and then

A new vision for PKI

We have seen that Public Key Infrastructure is not necessarily a centralised general purpose identification system. The new vision is for there to be a number of different more or less independent PKIs, each dedicated to particular e-business applications (or classes of application). Public key functions will be increasingly embedded in smartcards and application software, and key management processes—registration, revocation, renewal and so on—will be aligned with established membership rules and legal relationships. Thus the user's experience of PKI-enabled e-business should be exactly the same as that of any conventional PIN-protected plastic card application.

The table below elaborates how a dedicated PKI could be implemented in the health sector. The system is described from the users' perspective, with some of the underlying technical details noted in italic text in the margin.

A hypothetical Healthcare PKI	
An authoritative body for the sector would issue secure, PIN-protected smartcards to registered medical practitioners, representing their right to practice – nothing less and nothing more. The smartcard would allow its holder to conduct the same sorts of transactions online as they do off-line, under precisely the same legal and regulatory arrangements. Much of a doctor's rights and responsibilities in the Australian health system for instance is embodied in a government-allocated "Prescriber Number"; such numbers could be linked to the smartcard. In the main, smartcards would be minted in bulk, on behalf of all registered providers, from details held on the health body's definitive database. No application form is required. Smartcards could be distributed by post with PIN mailers to follow, or collected in person. Cards could be "topped up" with additional credentials at authorised service points. For example, a hospital administrator could add credentials specific to doctors practicing at that hospital.	*The health body acts as the main RA, and would work with an outsourced CA operator and a smartcard personalisation bureau. Specialty colleges, licensing bodies, institutions and so on could also act as RAs for particular types of credentials.* *The prescriber number could be coded into the certificate profile. Where prescriber numbers vary according to medical practice location, multiple certificates could be issued onto the one smartcard, to codify the different situations; prescriptions written in different practices would link to different certificates, just as they are written on different script pads conventionally.* *Authentication key pairs would be generated securely on the smartcard. Confidentiality keys could be loaded from outside, or generated onboard and exported for archive, for disaster recovery.* *Today's smartcards typically have capacity for four or more certificates, protected under individual or common PINs, at the*

	user's discretion. The Policy OID is the simplest way to distinguish (or "brand") certificates issued for different purposes. See below. *Using the health body's CA to sign special purpose certificates requested by RAs under different Policies leverages the body's PKI investment and cuts overall systemic costs.*
The smartcards would be compatible with a range of medical software applications, available on the open market.	*The health body's CA public key and the certificates' unique Policy OID(s) would be recognisable by application software, so that qualified medical practitioners can be distinguished from all other users. Software can search the card's certificate store for special credentials and automatically invoke them in the context of the application, such as hospital admissions and discharge, or prescribing of controlled drugs.*
Typically, a doctor or pharmacist for instance would log onto their clinical software application by inserting their smartcard into a reader and entering their PIN. After a period of inactivity, the software would require the user to re-enter their PIN. Users would be trained to take their card with them when they vacate their workstation.	*Once the user is logged on, the software is able to execute PKI-enabled functions on behalf of the user (create signatures, check signatures, decrypt messages etc.) with no further intervention from the user. The user need not know that any cryptographic processing is occurring at all.*
The software would provide a range of standardised forms (including prescriptions and insurance claim forms), reports (such as children's immunisation and other public health reports) and letter writing functions, producing files that can be sent to other healthcare workers, saved locally in patient databases and so on.	*Digital signatures would be generated automatically by the software and associated with the file as appropriate (e.g. sent with transmitted messages, filed with stored messages, logged with audit logs etc.). In cases where multiple credentials are stored on the one card, the software would automatically select the key and certificate most appropriate for the transaction.*
Messages sent by the healthcare provider would bear their credentials, picked up from the secure smartcard.	*The user's digital certificate, demonstrating their official credentials bestowed by the health body, would be sent along with all digitally signed transactions.*

In many cases, medical application software would utilise web mail for communicating with other healthcare workers. After connecting to the Internet, the software would receive notice of incoming messages, and notify the user with a dynamic hotlink from which the correspondence can be downloaded. Incoming data, in the correct format, can be automatically transferred into local patient databases and other applications.	*Web mail has several advantages including* • *positive receipt of delivery* • *ability to destroy undelivered messages* • *e-mail client independence* • *relatively good portability* *When incoming messages are downloaded, any attached digital signatures are automatically checked and associated certificates verified. Only on exception need the user be alerted to this process; for instance if the signer's credentials are not recognised, or their certificate has been revoked.*

Recommendations for regulators

To deliver the benefits of the new vision, the following general recommendations are offered to PKI regulators:

1. Shift from the orthodox position that sees digital certificates as representing absolute personal identity, to one that allows them to stand for membership of a defined community of interest.
2. Anticipate that many digital certificates will be application or context-specific, rather than intended for general purpose e-business.
3. Grant certificate issuers, acting on behalf of a community of interest, the discretion to define their own registration procedures—especially the evidence of identity rules—fit for the intended purpose of their certificates.
4. Require issuers of application-specific certificates to rigorously specify their registration procedures (typically in the Certificate Policy).
5. Require that application-specific certificates carry unique Object Identifiers that map to the registration rules and intended purpose, so that Relying Party software can be configured to automatically recognise application-specific certificates.
6. Include the intended purpose and the detailed context of specific applications in the terms of reference for accreditation/licensing of certificate issuers. As part of each issuer's evaluation, examine their broader scheme rules (if applicable) and how well they have matched registration procedures to the intended purpose.
7. For embedded PKI applications where the certificates are not visible—such as smartcard applications, or software that bundles the private keys and certificates into the executable—allow for Subscriber Agreements (and Relying Party agreements as applicable) to be subsumed into conventional types of user agreement.
8. Consider using X.509 Critical Extensions to enforce application-specificity, so that software not intended to utilise a given certificate is likely to automatically detect misuse.
9. In general, maintain the rigorous standards applied traditionally by schemes like Identrus and WebTrust for CAs to all backend operations, including RA/CA technology, cryptographic modules, data centre security, operational personnel security, and financial security.

Conclusion

The new vision for PKI means the technology and processes are no more of a burden on the user than any regular plastic access card. Rather than imagine that all public key certificates are like electronic passports, we should deploy multiple, special purpose certificates, and treat them more like electronic business cards. A certificate issued on behalf of a community of business users and constrained to that community can thereby stand for any type of professional credential or affiliation.

We can now automate and embed the complex cryptography deeply into smartcards, so that all terms and conditions for use are application focused. As far as users are concerned, a smartcard can be deployed in exactly the same way as any magnetic stripe card, without any need to refer to—or be limited by—the complex technology contained within. This approach increases usability, eliminates the onus on users to read and understand any CP/CPS, cuts the training burden, and allows legal liabilities for the use of the card to be determined under existing relationships and arrangements.

Application-specific smartcards can be issued under rules and controls that are fit for purpose, as determined by the community of users or an appropriate recognised authority. In particular, regulators should allow communities discretion to determine evidence of identity requirements for issuing their cards, instead of externally imposing personal identity checks, thereby dramatically cutting the overheads traditionally associated with digital certificate registration.

Finally, if we constrain the use of certificates to particular applications (or classes of applications) then we can factor the intended usage into PKI accreditation processes. Accreditation could then allow on a case-by-case basis for particular PKI scheme rules to govern liability. By "black-boxing" each community's rules and arrangements, and empowering the community to implement processes that are fit for purpose, the legal aspects of accreditation can be simplified, reducing one of the more significant cost components of the whole PKI exercise.