Public Key Superstructure

It's PKI Jim, But Not As We Know It!

7th Symposium on Identity and Trust on the Internet, NIST, Gaithersburg, MD, USA 2008

Stephen Wilson
Lockstep Consulting Pty Ltd
11 Minnesota Ave Five Dock NSW 2046 Australia
swilson@lockstep.com.au

ABSTRACT

While PKI has had its difficulties (like most new technologies) the unique value of public key authentication in paperless transactions is now widely acknowledged. The naïve early vision of a single all-purpose identity system has given way to a more sophisticated landscape of multiple PKIs, used not for managing identity *per se*, but rather more subtle memberships, credentials and so on. It is well known that PKI's successes have mostly been in closed schemes. Until now, this fact was often regarded as a compromise; many held out hope that a bigger general purpose PKI would still eventuate. But I argue that the dominance of closed PKI over open is better understood as reflecting the reality of *identity plurality*, which independently is becoming the norm through the Laws of Identity and related frameworks.

This paper introduces the term "Public Key Superstructure" to describe a new approach to knitting together existing mature PKI components to improve the utility and strategic appeal of digital certificates. The "superstructure" draws on useful precedents in the security printing industry for manufacturing specialised security goods without complicated or un-natural liabilities, and international accreditation arrangements for achieving cross-border recognition of certificates. The model rests on a crucial re-imagining of certificates as standing for *relationships* rather than identities. This elegant re-interpretation of otherwise standard elements could truly be a paradigm shift for PKI, for it normalises digital certificates, grounding them in familiar, even mundane management processes. It will bring profound yet easily realised benefits for liability, cost, interoperability, scalability, accreditation, and governance.

1. HOW DID PKI GET SO HARD?

PKI has been a notoriously disappointing technology. Much of the difficulty experienced bringing it to market can be traced back to the earliest PKI simply coming too soon. In the absence of well specified applications, an intuitive but ultimately distracting metaphor was allowed to dominate the agendas and rule the thinking of developers, product managers, policy makers, lawyers and standards setters. Ironically, while the metaphor was deceptively simple, it bred almost unlimited complexity.

Technically of course, an X.509 certificate does little more than bind a name to a public key value. This sort of arcane service hardly makes for compelling advertising, so naturally the early CAs and web browser vendors needed a simpler bit of imagery. What, they asked, was an X.509 certificate *like*? Whoever first suggested it was *like a passport* deserves a special place in the PKI hall of infamy.

1.1 The digital passport red herring

The notion of a digital passport had extra traction in the mid 1990s as it was already widely appreciated that creating "trust" in and on the Internet was going to be a crucial challenge.[1] So the world was much enamored with the idea that proving one's identity with a universally recognised passport is literally the key to doing business online. Perhaps the charm of the passport metaphor distracts people from the reality that most business is actually done in a local context. Furthermore, personal identity is not usually paramount, at least not in business; as a general rule in all walks of life, the less identification needed the better.

The implied objective of a one-size-fits-all digital certificate was perhaps the single biggest compli-

[1] Peter Steiner's cartoon "On the Internet, nobody knows you're a dog", the exemplar of the 'trust problem', appeared in New Yorker magazine in July 1993, thus predating almost all e-commerce as we know it today.

cating factor in all of PKI. By trying to make one certificate type meet the needs of all possible transactions, the legal arrangements became almost entirely unmanageable. A good question is why the futility of the universal PKI project wasn't spotted sooner.

The first Certification Authorities set up shop years before any meaningful e-commerce was on offer. Imagine trying to draft a subscriber agreement when you have no idea what a certificate is going to be used for. Any reasonable Threat & Risk Assessment has to explicitly relate to the application and its context. In the absence of any actual details, the only possible risk mitigation ploy is to enforce strenuous proof-of-identity checks on certificate subscribers so that in the event that something goes wrong, there is the prospect of sheeting home some blame.

If any trustworthiness at all could be vested in this type of certificate, then it is premised entirely on the rigor of the CA's certification practices. And so in turn the quite artificial situation arose where CAs, all of them brand new "trust" businesses, competed on the quality of their arcane Certification Practice Statements, as if customers could really be expected to read and care about these tomes.

So the digital passport idea, divorced as it was from any actual application, led immediately to legal complexities. The metaphor all on its own is likely to also be responsible for several operational quagmires, as follows.

1.1.1 Cost to the end user

Retail digital certificates are famously expensive and inconvenient to obtain. In many jurisdictions, the *de facto* proof-of-identity test was precisely (and arbitrarily[2]) the same as that of a passport. Such a level of identity vetting is highly unusual in everyday business. In Australia, an opt-in national PKI for healthcare professionals was met with strong opposition on this basis; administrators long complained of PKI being a "slow and unwieldy process" because of the personal identity vetting, and have cited "resistance from doctors and staff to fill out [registration] forms" as a major reason for the slow uptake of certificates [7].

1.1.2 The failure of Post Office CAs

The national postal authorities of several countries, including Australia, Belgium, Hong Kong, Malaysia, the UK and the US, rather quickly started up CA businesses on the strength of their existing privileged positions as passport registrars. Most post office CAs failed to generate any sustainable free market customer base for their certificates.

1.1.3 Cross Certification

The most unfortunate (albeit subtle) side effect of the passport metaphor in my view was the way it helped to inspire Cross Certification and certificate "policy mapping" as the dominant frame for creating PKI interoperability. Cross Certification is the orthodox way for certificates issued in different domains to be assessed for 'compatibility'. What's really going on here is a determination as to whether or not one CA's detailed processes – especially their registration policies – are *equivalent* to another's.

Leaving aside the practical matter that it shouldn't even be necessary for both counterparties to carry a certificate and belong to a CA, the deep limitation of Cross Certification is its inability to recognise different certificates as being fit for different purposes. Consider whether *it even makes sense* to ask if the certificate of for instance a Taiwanese doctor is "equivalent" to the certificate of an American immigration official. Cross Certification together with its offspring, the Bridge CAs, are premised on the assumption that one identity is all we need. As we shall see later, that notion has been repudiated several times over in the decade or more since PKI got its false start.

1.2 E-mail not a killer application for PKI

A total lack of real applications would explain why e-mail became by default the most talked about PKI application. Many PKI vendors to this day continue to illustrate their services and train their users with imaginary scenarios where our heroes Alice and Bob breathlessly exchange signed e-mails. Like the passport metaphor, e-mail seems easily understood, but it manifestly has not turned out to be a 'killer application', and worse still, has contributed to a host of misunderstandings.

The story usually goes go that Alice has received a secure e-mail from stranger Bob and wishes to work out if he is trustworthy. She double clicks on his digital signature and certificate in order to identify his CA. And now the fun begins. If Alice is not

[2] In Australia, the identity vetting protocol for passports and the related know-your-customer rules for opening a bank account were codified in legislation in 1988. The same identification standard was uncritically adopted by default eight years later when Standards Australia made its first efforts to standardise PKI [23]. Yet there is no logical connection in fraud mitigation measures between face-to-face retail banking and online transactions, and no obvious reason for the same identity vetting standards to have been carried over.

immediately trusting of the CA (presumably by reputation) then she is expected to download the CP and CPS, read them, and satisfy herself that the registration processes and security standards are adequate for her needs.

Does this sort of rigmarole have any parallel in the real world? A simple e-mail with no other context is closely equivalent to a letter or fax sent on plain white paper. Under what circumstances should we take seriously a message sent on plain paper from a stranger, even if we could track down their name?

In truth, the vast majority of serious communications occurs not between strangers but in a rich existing context, where the receiver has already been qualified in some way by the sender as likely being the right party to contact. In e-business, routine transactions are not usually conducted by e-mail but instead use special purpose software or dedicated websites with purpose built content. Thus we see most of the digital signature action in cases such as e-prescriptions, customs broking, trade documentation, company returns, patent filing and electronic conveyancing.

Several important simplifying assumptions flow from the fact that most e-business has a rich context, and these should be heeded when planning PKI:

1.2.1 Emphasise straight-through processing

In spite of the common worked example of Alice and Bob exchanging e-mails, the receiver of most routine transactions – such as payment instructions, tax returns, medical records, import/export declarations, or votes – is not a human but instead is a machine. The notion that a person will examine digital certificates and chase down the CA and its practices is simply false in the vast majority of cases. One of PKI's great strengths is the way it aids straight-through processing, so it has been a great pity that vendors, through their training and marketing materials, have stressed manual over automatic processing.

1.2.2 Play down Relying Party Agreements

The sender and receiver of digitally signed transactions are hardly ever un-related. This is in stark contrast to orthodox legal analyses of PKI which foundered on the supposed lack of contractual privity between Relying Party and CA. For example the Australian Government's extensive investigation into legal liability in digital certificates after 111 pages still could not reach a firm conclusion about whether a "CA may owe a duty of care to a [Relying Party] who is not known to the CA" [22]. The fact is, this sort of scenario is entirely academic and should never have been given the level of attention that it was. The idea of a "Relying Party Agreement" to join in contract the RP and the CA is moot in all "closed" e-business settings where PKI in thriving. It is this lesson that needs to be generalised by PKI regulators, not the hypothetical model of "open" PKI where all parties are strangers.

1.2.3 Play down certificate path discovery

The fact that in real life, parties are transacting in the context of some explicit scheme, means that the receiver's software can predict the type of certificate that will most often be used by senders. For instance, when doctors are using e-prescribing software, there is not going to be a wide choice of certificate options; indeed, the appropriate keys and certificates for authenticating a doctor issuing a prescription will likely be installed at both the sending and receiving ends, at the same time that the software is (see also a worked example at subsection 4.4). When a doctor writes a prescription, their private key can be programmatically selected and invoked to create a digital signature, according to business rules enshrined in the software design. And when such a transaction is received, the software of the pharmacist (or insurance company, government agency etc.) will similarly 'know' by design which certificates are expected to verify the digital signature. All this logic in most transaction systems can be settled at design time, which can greatly simplify the task of certificate path discovery, or eliminate it altogether. In most systems it is straightforward for the sender's software to attach the whole certificate chain to the digital signature, safe in the knowledge that the receiver's software will be configured with the necessary trust anchors (i.e. Root CA certificates) with which to parse the chain.

2. BIG PKI: ONLY EVER A STRAWMAN

"Big PKI" should have always been seen as a strawman, one that was construed with no real compelling need. Instead, in the vain attempt to allow stranger-to-stranger e-business, PKI inevitably grew ever more bloated and vulnerable to criticism.

Just consider the conventional sort of definition of PKI. NIST defines PKI as "personnel, policy, procedures, components and facilities to bind user names to electronic keys so that applications can provide the desired security services".[3] Microsoft considers it to be "the combination of software, encryption technologies, processes, and services that

[3] See http://csrc.nist.gov/nissc/1999/program/isso/tsld005.htm.

enable an organization to secure its communications and business transactions"[4] (the definite article at the start of the definition rather extravagantly seems to admit no other way for an organization to transact other than PKI). From the outset, this language sets PKI apart from any other authentication system. Traditional PKI requires an enterprise to commit itself to establishing novel and incredibly complex policies and procedures, in addition to deploying public key components. Allowing any new technology to so impact a business is plainly asking for trouble.

From the late 1990s a succession of critics sought to demolish PKI, usually on the basis of the mirage of a universal digital passport. The best known popular critique was probably that of Ellison and Schneier in 2000 [13] which detailed ten risks that we were supposedly "not being told about". On closer examination however, most of their concerns apply to the quality of security policies and the safekeeping of cryptographic keys in any setting, not just PKI. And when Ellison and Schneier do focus on PKI, it is actually the special case of a global infrastructure that they have in mind. For example, their argument that PKI doesn't resolve "which John Robinson is he" is unimportant in closed PKIs where communities of interest already have – indeed, *must* have – reliable mechanisms for guaranteeing unique handles in their local namespace. No PKI implementation should ever change the way users are known by the parties they deal with.

Another much cited assault on PKI came from academic law professor Jane Winn in her catchy 2001 exposé of "the shocking truth" [28] about digital certificates. Winn lampooned the prospects of forming new contracts over the Internet purely on the strength of strangers' certificates. Yet far from producing the definitive critique of PKI in general, she herself wrote that "what is now becoming apparent is that a more important [application] for digital signatures than 'open' Internet commerce among strangers may be 'closed' Internet commerce systems among parties *already in contractual privity with each other or to a system administrator*" (emphasis added).

It is this point that helps explain why, in the face of such widespread disillusionment and cynicism, PKI through the early to mid noughties continued to grow steadily and thrive in pockets. Well known examples include the Johnson & Johnson corporate PKI,[5] the Pan Asia Alliance trade documentation system,[6] Swedish financial sector's *BankID*,[7] the US Patent & Trademark Office online patent filing system,[8] the pharmaceutical industry's SAFE *Biopharma* scheme,[9] and Skype.[10]

It's possible that the florid ambitions of early PKI were amplified by dot-com mania. One analysis of the underwhelming demand for third party CA audit services suggested:

> *"[During] the Internet boom there was a belief that e-business was going to release a massive pent-up demand to conduct stranger-to-stranger commerce. But truly un-vetted business introduction is rare"* [15].

In parallel with the dawning realization that PKI works best when parties are not strangers, several other shifts in the identity management landscape have informed contemporary thinking, as follows.

2.1.1 The need for more than one certificate

The chair of the IETF PKIX Working Group Dr Stephen Kent has criticised the rigidity and unreality of orthodox "big CAs". In 2005 he told a conference of the Asia PKI Forum:

> *"For many big CAs, there is an assumption that a single certificate is all a user should need. This assumes that one identity is sufficient for all applications, which contradicts experience. For personal privacy and security, multiple independent certificates per user are preferable"* [18].

2.1.2 Supply chain perspective of certificates

In 2005 the OASIS PKI Technical Committee developed a new *digital certificate supply chain* to help better describe various cost components that impact on return on investment in PKI [27]. The supply chain recognises separable components of a PKI:

- the toolkits, libraries, services and so on used to PKI-enable software applications
- end user support
- digital certificates themselves

[4] See http://tinyurl.com/3ahtaw.

[5] See http://www3.ietf.org/proceedings/04nov/slides/easycert-1/easycert2.ppt.

[6] See http://www.paa.net.

[7] See http://www.bankid.com.

[8] See http://idtrust.xml.org/entrust-us-patent-office-success-story.

[9] See http://www.safe-biopharma.org.

[10] See http://share.skype.com/sites/security/2006/02/zui_and_the_skype_pki.html.

– Registration Authority services, costs and overheads
– Certification Authority operations
– key media (e.g. smartcards, SIM cards).

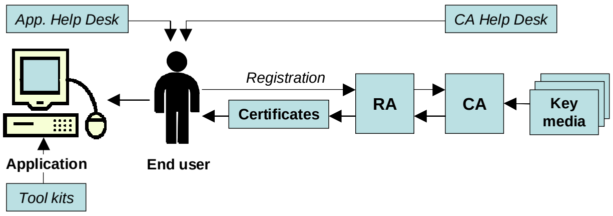

Figure 1: Digital Certificate Supply Chain

One important upshot of the supply chain perspective is that it more vividly underscores the separation of CA and RA, which most often are legally treated as the one entity. For instance, the most detailed legal analysis yet to be carried out on PKI in Australia, by law firm Clayton Utz in 2000, assumed that the CA carries out the functions of RA [22]. Decoupling the CA from the RA can be usefully extended further to create a *wholesale* approach to certificate production, as we shall see later.

2.1.3 Relationship Certificates

Greater separation of CA and RA helps the fresh formulation of "Relationship Certificates", originally developed by me for the Australian Government's *Gatekeeper* PKI program [5].

Orthodox digital certificates representing the personal identity of their Subjects are issued after an RA performs identity proofing on the applicant. They therefore represent an affirmation by the RA that the Subject has passed certain documented threshold tests relating to evidence of identity. A *Relationship* Certificate simply represents a different type of affirmation, namely that the Subject has a particular type of relationship with the RA. By extension, a Relationship Certificate can thereby stand for the Subject's rights or entitlements to participate in certain transactions sanctioned by the relationship. In a great many cases, significant and powerful credentials derive directly from membership of chartered professional associations, or simply from being employed by a company, and so can be instantiated by the relationship the user has with the organization's administration. Under these circumstances, a Relationship Certificate issued by the administrator means *nothing more and nothing less than the fact that the Subject is a member of the organization*; in particular, this type of certificate makes no formal representations about the subject's identity outside the organization. Relationship Certificates should lose their meaning outside the context of the relationship.[11]

Relationship Certificates have philosophical parallels with the idea of "authorization PKI" which has been floated sporadically as an alternative to "authentication PKI". For example, a recent IETF draft for secure Internet routing suggests that "[if] issuers need not verify the right of an entity to use a subject name in a certificate, they avoid the costs and liabilities of such verification" [6]. I believe that Relationship Certificates represent a fundamental shift in the way we think about PKI mainly because they break the nexus between authentication and authorization. A Relationship Certificate can evince its Subject's authorization to act in a given role on a given transaction domain *without needing to separately establish the person's "identity"*. In this regard, Relationship Certificates differ from two superficially similar constructions:

– ***Attribute Certificates***. Classically, Attribute Certificates do not bind public keys to users but rather only bind authorizations to names. They therefore cannot be used on their own to validate digital signatures, but instead are generally used in conjunction with some other general purpose public key "identity" certificate (a degree of complexity that appears to have inhibited the take-up of Attribute Certificates in commercial PC applications). Relationship Certificates on the other hand are just regular public key certificates. They stand alone to assert the Subject's role or responsibilities, and can be processed by conventional software. That is, Relationship Certificates can substitute for conventional X.509 certificates in standard applications without any software modifications; only the "business rules" for interpreting what a certificate means need updating.

– ***SPKI ("Simple PKI")***. SPKI [12] was formulated in the late 1990s in response to some of the challenges summarised above, as a way of mapping an authorization directly to a key, thereby skipping the cumbersome mappings of names to keys (using regular Identity Certificates), and authorizations to names (using Attribute Certificates). In this regard Relationship Certificates closely resemble SPKI Authorization Certificates. Unfortunately, SPKI has not

[11] In the real world, all credentials have context, and the appropriate credential depends on the transaction. For example, if a doctor were pulled over by a traffic cop and asked to show her drivers licence, she should get nowhere trying to present her medical qualifications.

penetrated the market as far as hoped, perhaps because it positions itself implicitly as an adjunct to the name-key mapping. SPKI is often associated with a needlessly complicated triangle formed from Identity, Authorization and Attribute certificates; see for example [20].

Further operational details of how Relationship Certificates could be implemented are provided in Section 4 and a worked example provided in subsection 4.2.3.

2.1.4 *Smart key media*

The historical complexity faced by users in managing keys and certificates is being almost entirely put to bed by an increasingly rich array of smart key media. With key pair generation integrated into their chips, and certificate lifecycle management being absorbed into card management systems, PKI enabled smart-cards in particular are set to transform PKI. They are exactly as easy to use as any conventional magnetic stripe card.

2.1.5 *New thinking about identity*

Finally, we can look to the new wave of thinking about identity in general as an indication of better ways to utilise PKI. Much of the focus of "identity 2.0" (as promoted by organisations such as sxip[12]) is on the multiplicity of things that we say about ourselves, and the things that others say about us. That is, identity 2.0 begins with a realization that the usefulness of online identity depends on context, and it must be responsive to the natures of the diverse relationships we have with those we transact with.[13]

It seems to be increasingly accepted that people live with multiple identities. The preeminent exposition of modern identity theory is probably Kim Cameron's Laws of Identity [10]. The laws include a new definition of *digital identity* as "a set of claims made by one digital subject about itself or another digital subject". Cameron knows that this sort of relativist definition might not sit comfortably with everyone:

> *"We recognise [that our definition] does not jive with some widely held beliefs – for example that within a given context, identities have to be unique. Many early systems were built with this assumption, and it is a critically useful assumption in many contexts. The only error is thinking it is mandatory for all contexts."* [10]

But Cameron is certainly not alone, not anymore. Other researchers have reached the view that there may be a many-to-one mapping of identities onto entities; see e.g. Jøsang and Pope:

> *"An identity is a representation of an entity in a specific application domain. For example, the registered personal data of a bank customer, and possibly also the customer's physical characteristics as observed by the bank staff, constitute the identity of that customer within the domain of that bank. … A person may of course have different identities in different domains. For example, a person may have one identity associated with being customer in a bank and another identity associated with being an employee in a company."* [17]

The notion of what I would call *identity plurality* is not merely a semantic or philosophical point. A simple example demonstrates that in business we clearly conduct ourselves according to multiple identities, and that we seamlessly switch between them without trouble. Furthermore, when we exercise a context-dependent identity, we beneficially mask our biological one. Imagine that a company Acme Inc. has a corporate bank account with A Banking Corporation (ABC). The Acme company secretary, Alice, would be a signatory to the Acme bank account and would have custody of an ABC key card for the purpose. Alice might also hold a personal account with ABC. Now, when she banks on behalf of Acme, Alice exercises a different identity compared with when she banks on her own behalf, even if she happens to access both accounts during the same visit to the branch or ATM. The distinction is both emotional – Alice probably won't feel any real attachment to the millions of dollars she routinely handles for the company – and legal. Corporate law says clearly that the Acme account holder is not Alice but the company.

One deep implication for PKI of identity plurality is that it inverts the expectation that closed PKI is a compromise while open PKI is the proper long term goal. On the contrary, we should now appreciate that open PKI would be a special and highly theoretical instance. It is the closed PKIs – each with its own arrangements and business rules – that represent the general case.

[12] See http://www.sxip.com.

[13] However, many in the identity 2.0 movement go from this background to a position of desiring all diverse relationships to be federated into a multi-context transcendent identity. In my opinion this is one step too far. Dick Hardt's famous identity 2.0 conference presentation is frankly utopian in the way it advocates linking all our reputations together, for it overlooks the privacy problems arising when linked records are exposed by accident, when wrong doers exploit the linkages, or when a user seeks to sever one of their relationships for some reason.

3. WHAT IS PKI REALLY GOOD FOR?

I contend that clearly the best use of PKI is to help automate electronic transactions in a particular context between parties that already have a formal relationship.

The orthodox textbook accounts of the benefits of PKI invariably list authentication, integrity and something called "non-repudiation". These high level properties may actually be delivered by all manner of technologies, a fact that made early PKI's over-inflated marketing claims seem frankly silly, even at the time.[14]

3.1 A clearer benefit description for PKI

We need a more sophisticated shared understanding of what makes PKI unique. I suggest its unique benefits would be better told as follows:

– Digital signatures create long-lived, tamper-resistant evidence of 'who did what to whom', which is so critical to electronic transactions carrying high legal risks or compliance requirements.

– PKI, when deployed with hardware key media like smartcards, is recognised as "the only practical solution [to eaves-dropping and account hijacking] today" [9]; digital signatures originated by the end user protect against Man-in-the-Middle attack, while smart key media offer a sufficiently compact logic engine to be certifiably resistant to malware.

– Digital certificates can convey *authority information* – like credentials, licences, affiliations and so on – and digital signatures bind that authority information directly to messages, to decentralise and greatly simplify transaction processing.

PKI digital signatures are persistent over both time and 'distance', meaning the separation of sender and receiver. At essentially any time in the future[15] a digitally signed transaction can be easily re-validated to prove where it originated: all that is needed is a trusted copy of the root public key, the certificate chain, and the relevant CRLs (all of which are routinely available from any decent CA). In addition, authority information about the sender can be sealed into their certificate at the time of issue, and this authority information also has great longevity, thanks to the digital signature of the Certificate Authority on the certificate.

The integrity of digitally signed data is not reduced by being copied or forwarded across systems or across borders. In contrast, other authentication technologies rely heavily upon audit logs to prove 'who did what to whom'; forwarding non-PKI transactions from one system to another complicates and dilutes the strength of the audit trail. So PKI is uniquely suited to complex transaction environments, like healthcare, pension fund management and trade documentation, where there are multiple relying parties, formal authorizations, and/or long lifetimes.

3.1.1 *The challenge of persistent credentials*

Verifying transactions originating from professionals is a case in point. Consider a lawyer who signs conveyancing documents relating to a land sale. When the contract is settled, all parties (the buyer, the seller, their respective banks and so on) will be acutely interested to know that the lawyer's credentials are valid. It is straightforward to check credentials online, at the time, by looking up a database of qualified practitioners. But in electronic conveyancing, what becomes important is the ability to check the credentials of a lawyer who signed documents in the past. Unless special measures are taken to archive practitioners' databases, it is difficult to obtain definitive machine readable information about the state of someone's credentials at a given time in the past. With digital signatures and digital certificates on the other hand, the matter becomes trivial: if Relying Party software knows the relevant Policy OID and has a trusted copy of the root public key, then it can verify the credentials of the lawyer at the same time as it verifies the digital signature, no matter how old it is (with reason, as qualified by the long term risks mentioned previously).

[14] PKI has no monopoly on "non-repudiation" despite the term only being coined in connection with it. PKI marketing too often suggests that only PKI delivers non-repudiation. If this were true, credit card holders who use their cards online could try to mischievously repudiate any one of their payments on the basis that it was *not* digitally signed and therefore did *not* have "non-repudiation"!

[15] Several very long term risks ultimately threaten the validity of old digital signatures. Brute force attack by future computers on asymmetric cryptographic algorithms, exploitation of likely weaknesses in MD5 and SHA-1, and eventually the possibility of quantum computing, all mean that any digital signatures intended to remain valid for a few *decades* should have their keys and certificates comprehensively archived. Note that the stability and usability of archive media over the decades is another quite separate challenge.

3.2 PKI in plain English

The steadily improving automation of digital signature and certificate management operations means that the way we describe PKI to lay people can now side-step the technical details of asymmetric cryptography, hash functions and so on, and focus instead on what it actually *does*. A fresh, plain English description might run as follows (assuming smartcards are the key media).

> *A smartcard plus application software combine to produce* digital signature *codes for electronic transactions. Unlike any other electronic signature method, digital signature codes are unique to the owner and also to each transaction. Digital signatures operate as if a personalised electronic stamping machine was inside each smartcard, creating a specific tamper resistant 'mark' on each message or file created by the card holder. Digital signatures remain valid indefinitely; at any time in future, the 'mark' can be easily verified to prove its origins.*
>
> Digital Certificates *are electronic notices that bind identities to such devices as smartcards.*[16] *Certificates can thereby bind individuals to transactions signed using their smartcards. A digital certificate can identify the card holder and can also hold any other information that the issuer is qualified to declare. If the issuer is authoritative over information such as professional credentials, then that information can be sealed within its digital certificates and thus bound to each card holder plus the transactions they sign.*
>
> *To process digitally signed transactions, the receiver's software requires a copy of the sender's certificate, plus a special "master code" – known as a root certificate – which is used to mathematically validate all certificates in a given PKI scheme. Different master codes define different PKI schemes, be they sector-specific, national or general purpose such as SSL website authentication. Application software can ship with all necessary master codes, or can have them installed later.*
>
> *Digital certificates can be electronically revoked at any time. Revocation may be requested by the holder in the event that they lose their smartcard. Alternatively, revocation of a professional's certificate may follow automatically from their membership lapsing, their qualifications being cancelled, or their employment changing.*

3.3 Modern PKI success stories

Many of the more recent PKI success stories resonate with the concepts of identity plurality and digital certificates having more to do with multiple relationships than a single identity. Examples follow.

- A large public hospital in Australia developed a new "Known Customer" certificate to be issued on smartcards to several thousand of its staff [8]. The intended digital signature applications include electronic medical notes created by nurses, electronic hospital discharge notes, and online employee self-service access to pension fund administration, leave forms and so on. The hospital's human resources department will operate a delegated Registration Authority workstation. A commercial back-end CA will independently manufacture customised certificates on request from the RA, and inject them onto smartcards. The same CA will be able to produce similar but distinct Relationship Certificates for other communities of interest in the health sector. Overlap between healthcare communities is commonplace, with geographically related area health services often sharing information management resources and infrastructure. "Interoperability" of the hospital's staff certificates with other local applications will be easily fostered simply by promulgating knowledge of the certificate Policy OIDs.
- The Australian government has been exploring how digital certificates can act as electronic credentials for a number of different types of professionals. A state association for legal professionals has researched how digital "practicing certificates" can be issued to attorneys.[17] The most compelling application for digital signatures in the practice of law is electronic conveyancing. E-conveyancing is forecast to provide direct savings of AU$70 per transaction for vendors and purchasers, and an overall saving to industry of AU$33 million p.a. by 2010, assuming 66% of transactions are by then done online [24].
- Most e-health projects around the world anticipate the use of digital certificates. The use of digital signatures in the pharmaceutical industry has been fostered by the Food & Drug Administration's Title 21 Code of Federal

[16] This simplified account deliberately but without loss of generality suppresses the intermediate detail that the certificate actually binds the identity to a key pair, which is separately bound to the smartcard by way of hardware key management.

[17] See news about this project at http://www.galexia.com/public/projects/projects-Law.html (accessed 31 Jan 2008).

Regulations (21 CFR Part 11) in respect of Electronic Records and Electronic Signatures. Health smartcards in France[18] and Germany[19] are currently being upgraded with PKI-capable chips so as to support a new wave of applications that require patient signatures, such as e-prescribing. The Australian federal Department of Health and Ageing in 2006 commissioned independent security analysis that strongly endorsed digital certificates for e-prescribing [2].

4. PUBLIC KEY SUPERSTRUCTURE

Having painted a newly optimistic picture for the future of PKI, one that resonates with broader identity management trends, I will now describe a number of fresh ways to better knit together extant mature building blocks – X.509 and similar certificates, RAs, CAs and PKI audit services – to deliver better, more flexible transaction authentication.

4.1 Relationship Certificates in practice

Relationship Certificates, as described in subsection 2.1.3, are best managed within an arrangement where a defined "community of interest" deploys digital certificates that represent membership of the community. Operationally, Relationship Certificates are issued with the administrator of the community acting as a delegated RA.

Such arrangements have been studied extensively and piloted by both the legal and medical professions in Australia, as mentioned in subsection 3.3. In response to market demand for PKI-based digital credentials that convey richer information about professional qualifications without being burdened with artificial registration requirements, the Australian Government PKI program recently introduced a special category for Relationship Certificates [5].

4.1.1 The Relationship Certificate profile

To be most effective, Relationship Certificates would have information in their X.509 (or similar) profile to specify the precise nature of the relationship between RA and Subject, allowing straight-through processing by any Relying Party software application configured to recognise the validity of the relationship. The best way to codify the meaning of a Relationship Certificate is probably in the Policy OID, which can be specified at design time.

Ideally, technical controls should be implemented as well to make it difficult to misuse a Relationship Certificate outside its intended context. One way to implement technical restrictions on misuse would be to include a *Critical* extension in the profile. Recall that the X.509 standard requires any software processing a certificate which has an extension marked as *Critical* to reject that certificate unless it expressly recognises the extension. Since special purpose software (as opposed to general purpose web and e-mail clients) is usually used in PKI-enabled transaction systems, within communities of interest, programming in awareness of *Critical* extensions is easy. And by the same token, it is safe to assume that if a given software program does not recognise the *Critical* extension, then it is proper behaviour to reject the certificate, on the grounds that such certificates are not supposed to be used outside special purpose applications. *Critical* extensions proved unpopular in the past because they were thought to harm interoperability. But if a special purpose Relationship Certificate is only intended to work with certain applications, then "interoperability" is more or less moot, since no other applications should be expected to accept it.

4.1.2 Practical benefits of Relationship Certificates

Relationship Certificates would bring major simplifications over third party identity certificates in several areas:

- Overheads associated with registering for certificates are greatly reduced; customers already known to the administrator in a community of interest will be able to receive certificates almost automatically without having to present in person at an unfamiliar RA.
- Certificate Subjects will require no legal relationship with the backend CA; any important new obligations introduced by PKI – such as responsibility to safeguard one's smartcard and promptly report its loss – can be folded into the administrator's formal contractual relationship with its members, rather than expressed in the traditional CA's "Subscriber Agreement".
- Users will no longer be required to pay up-front for a certificate from a third party CA in order to use PKI-enabled applications.
- Furthermore, the price of certificates should fall towards "wholesale" levels, because the cost of identity proofing associated with traditional identity certificates will be eliminated.

[18] http://www.sesam-vitale.fr/programme/programme_eng.asp.

[19] http://www.die-gesundheitskarte.de (in German).

– Support overheads and complexities may be lessened by having just one help desk for all business, application and certificate-related matters.

4.2 Wholesale certificate production

Historically, CAs have been tied legally into the whole of the certificate management process, no matter how they might operationally involve RAs in the registration and certificate lifecycle management processes. CAs tend to be joined in liability arrangements and contracts to potentially any wrongdoing or misadventure associated with certificates. Certificate policies, practice statements and user agreements have been correspondingly difficult to construct. To date, the separation of roles of RA and CA has done little to quarantine the two functions from one another, nor to simplify liability arrangements. Accreditation remains complex and sensitive to the slightest changes at either the RA or CA.

There might be a new way however of looking at backend CAs, likening them to conventional *security printers*, and dramatically simplifying the way that legal liability is apportioned when something goes wrong with a digital certificate.

4.2.1 The business of security printing

For decades it has been well known that in order to combat fraud, special care must be taken in printing certain documents: blank checks and prescription pads in particular, as well as business forms, concert tickets, gift vouchers, barcodes and so on. A whole industry has been built around special printing technologies, including watermarks, holograms, reactive inks that detect photocopying, and micro-printing. Moreover, a coherent business model has been built around security printing bureau services. In many sectors, standards have been introduced to cover the necessary security of premises and processes, and formal accreditation schemes govern compliance with these standards. For example, since 1 January 2005, written prescriptions for controlled substances in California must be on tamper resistant security prescription forms produced by a printer approved by the State Board of Pharmacy.[20] And the United Kingdom's payment clearing regulators APACS introduced a formal Check Printer Accreditation Scheme (CPAS) in 1995.[21]

4.2.2 The governance of security printing

Many of the standards governing security printing closely resemble those for traditional CAs. For example, check printer accreditation typically covers assurance of the following aspects of the operation:

- *'Equipment & Materials*
- *'Premises Security (external security for prevention of unauthorised access, and internal security with appropriate restrictions on access to different areas*
- *'Process Security (end-to-end process controls from raw materials, through to end product, full audit trail in place for each print job, destruction of unused/damage stock, protection of confidential information, employee screening and confidentiality clauses)*
- *'Order Processing*
- *'Quality Assurance*
- *'Dispatch & Delivery (secure & auditable dispatch system, sign-off for delivery, secure transport arrangements, secure packaging, appropriate labeling not to identify as checks, processes for lost/stolen consignments)'* (adapted from [14]).

Accredited printers under the British CPAS variously emphasise their personnel screening, internal segregation of access-controlled security cages, and perimeter fences and monitoring systems. Clearly a similar degree of effort is involved physical, procedural and personnel security for security printing operations as for well run CAs such as those typified by accreditation under *Identrust*, *WebTrust for CAs*, Australia's *Gatekeeper* PKI scheme, or the UK's *tScheme*.

4.2.3 Worked example: barcodes and certificates

A practical worked example of how digital certificates could replace a conventional paper security mechanism helps to further develop the comparison between backend CAs and security printers.

Consider a stock exchange that arranges for statutory announcements made by listed companies to be communicated by fax and secured by means of barcodes. Each listed company is provided with a

[20] See http://www.ag.ca.gov/bne/security_printer_list.php.

[21] See http://www.apacs.org.uk/payment_options/cheques_accreditation_scheme.html.

roll of self adhesive barcode labels. The barcodes uniquely identify the company and are individually serial numbered. When a statutory announcement needs to be made in accordance with the stock exchange's Listing Rules, the announcement is printed, signed by a duly authorised company officer, and has a barcode label affixed to it, before being faxed to the announcement processing centre. When received, optical character recognition software scans the fax, extracts the announcement, and verifies the barcode, before broadcasting the news across stockbrokers' screens. See Figure 2 below.

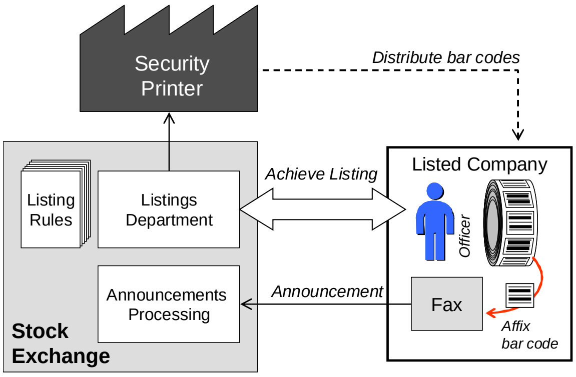

Figure 2: Authenticating faxed company announcements by means of secure barcode

The barcode label is an authentication token. Inclusion of a barcode on a fax is taken as reasonable evidence that the sender is a listed company, operating under the stock exchange's rules. Clearly such barcode labels are precious items. They need to be produced by a reputable security printer, with the ordering and distribution processes being subject to strict controls.

Now let us consider how the announcement processing system could be reengineered to use PKI and electronic messaging in place of fax machines. Figure 3 shows a nearly identical system, where the listings unit operates an RA (not shown), and instead of ordering barcode labels from a security printer, it orders digital certificates from a backend CA. In order now to make an official announcement, the company officer would use the certificate to digitally sign an electronic message.

Note that the certificates issued in this particular scheme are an instance of *Relationship Certificates*. They are not intended in any way to stand for the "identity" of company officers. Rather, they represent nothing more and nothing less than the fact that each Subject is an officer of a company listed on the stock exchange.

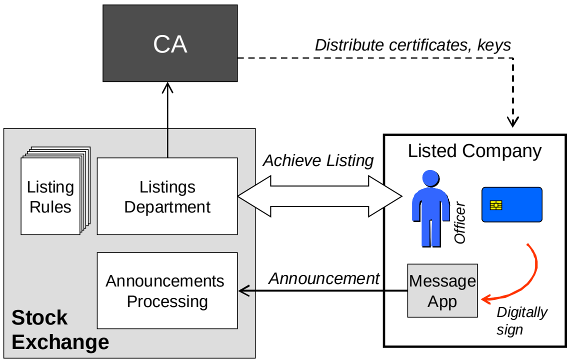

Figure 3: Authenticating electronic company announcements by means of digital certificate

Let's compare the security requirements of announcement methods that alternately use barcodes or digital certificates. Regardless the authentication method, the announcement system requires a common set of security controls:

1. The company listing process (which is where the relationship between a company and the stock exchange is established) must be robust and difficult to subvert.
2. It must be difficult to fraudulently order barcodes [or digital certificates].
3. Barcodes [or private keys and digital certificates] must be difficult to counterfeit.
4. The security printer [or backend CA] must be difficult to subvert.
5. Barcodes [or private keys] must be distributed and stored carefully.

If the conventions of orthodox PKI were to be applied to this operation, then a number of additional complexities would be imposed from outside on how the stock exchange runs its business. In particular, most PKI regulators today would expect standardised identity proofing for all certificate recipients at a level equivalent to passport application, irrespective of how the existing listing rules operate.[22]

Furthermore, the company officer, as certificate Subject, would generally have execute a user agreement *with the CA*. In contrast, requiring them to

[22] In the current climate of concern for homeland security, anti-money laundering, improved corporate governance and so on, it happens that many organizations are looking to strengthen their identity vetting processes, but nevertheless, that is an exercise that is not logically connected with PKI *per se* and the two should not be confused.

sign up with the security printer responsible for the barcodes would be unthinkable! Finally, changing backend CA typically triggers major re-accreditation of any regulated end-to-end PKI solution, because peak documents like the CP and CPS tend to intertwine all of the operational aspects. In the "real world", if the stock exchange gets a better deal from a competing printer, the changeover in backend operational matters would be completely invisible to the listed companies.

Comparing the digital certificate approach to the barcode system is suggestive of a more streamlined approach to PKI operations and accreditation. First note, referring to the list of security requirements on the previous page, that security controls can be clearly separated according to whether they relate to (1) the risk of *impersonation*, which tends to be managed by process or (2) the risks of *counterfeiting* or *theft*, which tend to be managed by technology:

– ***Impersonation related risks***
  - the company listing process must be robust and difficult to subvert.
  - it must be difficult to fraudulently order barcodes or certificates.

– ***Counterfeiting and theft related risks***
  - barcodes, private keys and certificates must be difficult to counterfeit.
  - the security printer or backend CA must be difficult to subvert.
  - barcodes or private keys must be distributed and stored carefully.

At the front-end of this authentication scheme, where the stock exchange deals with its companies, there is no logical difference between using barcodes or digital certificates, so we should expect the security of existing stock exchange registration processes to carry over to the PKI implementation without change. And at the backend, there is no need for either a security printer nor a CA to be concerned with the details nor even the integrity of the company listing and customer service processes, so long as there are controls in place to mitigate against wrongful ordering.

4.2.4 *Implications of security printing for CAs*

So, why couldn't we treat backend CAs in the same way as we treat regulated security printers? If a CA was set up as a service bureau, responsive to a particular set of RAs with which the CA has a specific arrangement, producing certificates on instruction more or less automatically via standard certificate request protocols, then a number of major simplifications to PKI management and governance could follow:

– The CA need have no interest at all in the semantic contents of the certificates it produces on instruction from a contracted RA. So long as there are safeguards in place to mitigate against false certificate requests being injected between the RA and the CA, the CA need not know anything at all about the RA's business process, nor the intended application of the certificates. The CA's business model and detailed processes could be held entirely constant over a wide range of different PKI applications. Protecting against injection of false certificate requests is a standard feature of most CA-RA products and is an express part of most if not all PKI product certification.

– There is no need for a contract or other legal arrangement between end users of certificates and the backend CA (just as there is no need for end users of checks and barcodes to have any relationship with the respective security printer).

– The CA's liabilities are straightforward to analyse and codify. For example (and in stark contrast to orthodox RA/CA arrangements) it seems clear that a CA would not normally be joined in legal action resulting from an RA being negligent in registering an impostor for a certificate. On the other hand, acts of omission or commission by a CA in producing poor quality certificates which led to harm on the part of message recipients, could be identified and prosecuted as such, and isolated from the RA.

– The meaning of the root key – which in orthodox PKI has led to so much confusion – can be likened to a unique watermark featured in all products from a given security printer. The chaining of a certificate back to the CA root would represent the simple fact that the certificate has come from an accredited facility (see subsection 4.4.1 for more details). The Root CA signature means only that it is extremely unlikely that a certificate has been forged, and does not impart any approval or endorsement of the contents of the certificate.[23]

– It should be much easier to novate backend CA service arrangements from one supplier to another.

[23] When couched in this way, the certificates issued by a Root CA can be seen recursively as special instances of Relationship Certificates.

Note that the security printing model would essentially preserve the physical, procedural, personnel and technological security controls of most current CA accreditation schemes, in order to protect against counterfeiting and subversion of the backend process. In particular, the benchmark of Common Criteria EAL4 rated CA and RA products would probably be retained, to help prevent fraudulent ordering of certificates.

4.3 Revisiting certificate interoperability

As discussed above, the historical focus on cross certification appears to have been a well-intended but misguided attempt to determine the equivalence of certificates issued in different domains. If we take time to revisit the business need for accreditation of PKIs, we can formulate a more powerful and yet lower cost approach to interoperability.

4.3.1 How should certificates "interoperate"?

Is there a topic in PKI more important and yet more confused than interoperability? A senior finance sector executive captured the uncertainty perfectly:

> *"[PKI] interoperability is something of a will-o'-the-wisp. You think you understand what people mean by it, and then quickly realise that you don't. In my experience, it's possible when discussing interoperability to be at cross-purposes for all of the time. Interoperability between members of the same PKI is axiomatic. Certificates issued by one bank should be recognizable by another. Interoperability becomes an issue when it is between different PKIs … But this still leaves the basic question of interoperable in respect of what?"* [21].

The best place to start thinking about interoperability is to unpack with a functional focus how digital certificates can help with authentication. A fine definition of authentication comes from the APEC eSecurity Task Group: *"The means by which the recipient of a transaction or message can make an assessment as to whether to accept or reject that transaction"* [2]. In the case of digital certificates, from the perspective of the receiver or Relying Party, the central question is really very simple: What information is available, in the certificate chain and elsewhere, to help the receiver decide whether to accept or reject the certificate and hence a digitally signed message?

There are three main things the receiver needs to know about a certificate in order to tell if it is fit for purpose:

1. ***What representations does the certificate make about its Subject?*** Or equivalently, was the certificate intended to be used in the transaction concerned? With Relationship Certificates standing for specific credentials or memberships conferring particular authorizations, each will bear a unique Policy OID indicating its intended applicability and context.
2. ***Is the certificate valid (i.e. not revoked)?*** Note that while revocation status is usually thought of as a question posed in real time, sometimes it will be back-dated; that is, we may need to know if the certificate Subject was valid at the time they launched the transaction (see e-conveyancing discussion in subsection 3.1.1).
3. ***Was the certificate issuer acting in compliance with applicable standards and regulations?*** Relevant standards will vary from one domain (or PKI scheme) to another; examples include the Australian government's *Gatekeeper* program, the finance sector's *Identrus* and the more general purpose *WebTrust for CAs*.

All of the information that an application needs in order to accept or reject a certificate could be found in the certificate chain, under the right circumstances. Compared with orthodox PKI which referred vaguely to "chains of trust", we need to be more precise about what certificates issued to CAs represent. If they represent each CA's compliance with standards (like *Webtrust for CAs* or *Identrust*) then when an end user certificate chains back to the root we can be sure that all intermediate CAs are doing what they're supposed to do. And if the end user certificate's Policy OID matches our expected value, then the certificate can be relied upon. For more details, see subsection 4.4.1.

4.3.2 Cross recognition versus cross certification

When transactions cross between jurisdictions or communities of interest, users must be able to determine whether or not to accept a transaction signed using an certificate issued elsewhere. This then is the fundamental issue in electronic authentication, rather than the quite arbitrary question of whether counter parties' certificates happen to be equivalent, as discussed in subsection 1.1.3.

In contrast to cross certification, *cross recognition* is defined as *"an interoperability arrangement in which a relying party in one PKI domain can use authority information in another PKI domain to authenticate a subject in the other PKI domain"* [1].

Users in a community of interest require information and guidance from their community leaders about the fitness for purpose of whichever external certificates can be expected to be received with incoming transactions. With a range of CAs issuing certificates for different uses, it is essential that a Relying Party can tell if an incoming certificate is acceptable for the transaction concerned; ideally their software application should be able to decide online and automatically whether to accept or reject a given certificate.

If a CA has been accredited under an external PKI scheme, then the issue boils down to whether or not that accreditation is acceptable to the local community of interest for the intended use of the certificates [26]. This is perhaps the simplest statement of the problem of cross recognition of PKIs.

4.4 How to convey "fitness for purpose"

Where a CA is audited or accredited under a particular scheme, its standing under that scheme should be made available to Relying Parties online. *Webtrust for CAs* does this to some extent by way of a web seal on the CA's site, but this requires out-of-band examination by the Relying Party, at least on occasion. That is, the fact of accreditation is not machine readable. It would be far better for a Relying Party application to be able to recognise programmatically the fact of accreditation.

4.4.1 Rendering CA audits machine readable

A more powerful and interoperable way to represent accreditation is to use a conventional X.509 certificate issued by (or on behalf of) the auditor. In 1999, I proposed an "accreditation based" way to construct PKI, in which "the X.509 certificate issued by an intermediate CA to a user CA is interpreted explicitly as a compliance certificate, directly analogous to the paper certificate issued by a [quality standard] certifier to a compliant organisation" [25]. The basic thrust of this proposition was adopted by the Australian Government PKI program in its *Gatekeeper Accreditation Certificate* CA initiative, which, while not yet operational, is envisaged will:

> *"… issue a digital certificate – the Gatekeeper Accreditation Certificate (GAC) – to each Gatekeeper Accredited CA. Issuance of the GAC would confirm that the CA has satisfied the Australian Government's requirements for Gatekeeper Accreditation. In issuing a GAC, Finance will* ***not*** *be acting as a Root Certification Authority as it does not impose a policy regime on digital certificates issued by subordinate CAs"* (emphasis in original) [4].

(Note the evident reluctance to act as a Root CA, an inhibition that will be considered in more detail in subsection 4.4.2.)

A key advantage of the accreditation based PKI model is that the audit certificate can be parsed and interpreted by entirely conventional X.509 software. The existence of a valid and current certificate chain extending from an end user back to a recognised auditor can be interpreted to mean that the user certificate is fit for purpose as circumscribed by the scope of the audit, and that the certificate was issued by a CA that was, at the time of the last inspection, found to be in compliance with its own policies and procedures as well as any other prescribed standards.

For an illustration, see Figure 4 which depicts a digital signature of a qualified doctor chaining through a Relationship Certificate to an issuing CA, the certificate of which is signed by a Root CA representing a health sector scheme.

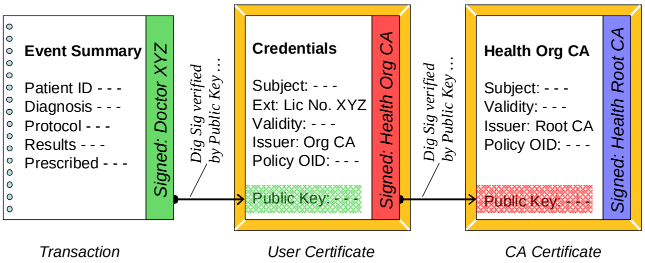

Figure 4: Imputing fitness for purpose from a certificate chain

If we adopt a somewhat fresh interpretation of what it means for various CAs to sign certificates, then the chain in Figure 4 allows fitness for purpose to be imputed as follows.

In this example, the user certificate issued by (or on behalf of) a reputable health organisation represents nothing more and nothing less than the fact that the Subject has a meaningful medical qualification. The Policy OID in that certificate directly represents a transaction domain on which the digital qualification is deemed to be valid. For example, if the health organisation were a medical practitioner registration board, then it could be that its certificates confer the authority to sign e-prescriptions and other transactions under certain legislation (and the Certificate Policy would call out that legislation explicitly and incorporate, probably by reference, all associated rights, responsibilities, terms and conditions). On the other hand, if the health organisation were a hospital, then its certificates might have a more restricted scope of meaning, such as the authority to admit patients for procedures according to a contract between the hospital and a doctor. Similarly, a certificate issued by a Health Maintenance Organisation (HMO) or private health insurer could confer authority to order particular tests electronically. In these cases the Certificate Policy would call out the applicable contract from which the certificate Subject's authority obtains.

So, if the digital signature on a health transaction chains correctly to a user certificate issued by the authoritative Health Organization CA, then the receiver can be assured of the veracity of the provider's credentials. It is straightforward for receiving software that deals with a whole class of healthcare transactions to be configured with the Policy OIDs of all issuers of certificates deemed to be authoritative for those transactions (obviating the need for complex certificate path discovery, as discussed in subsection 1.2.3).

Turning to the Health Organisation CA itself, it has been issued a certificate by a Health Sector Root CA. We interpret the signature of the Root CA as conferring membership of a health sector scheme. For a CA to hold a valid certificate signed by the Root means that the CA has been deemed to have met the scheme rules, and has passed whatever audits are specified by those rules. If the conditions of member-ship of the scheme are ever breached then the CA's certificate can, as an ultimate sanction, be revoked.

4.4.2 Root CA as "conformity assessment" anchor

Now we can consider re-inventing the role of Root CA. Orthodox formulations of the role and responsibility of Root CAs have historically been confusing (if not confused). It has been difficult to avoid unspecified legal liabilities growing as we move up the "chain of trust" from CA to Root.

But what exactly is it that a Root CA does? Or what should it do? As we saw in the previous subsection, there is a presumption that Root CAs "impose a policy regime on digital certificates issued by subordinate CAs" [4]. At least one PKI scheme – Australia's *Gatekeeper* – is reluctant to have Certificate Policy imposed by the Root CA. Instead, it prefers to allow member CAs to remain entirely autonomous in the way they construct their OID trees.

Operationally of course, a PKI needs a top-most CA that spawns other operational CAs, and provides a "trust anchor" to which certificates can chain. Relying Party software needs a dependable copy of the Root CA Public Key, and when a chain of certificates is established that terminates at a self signed Root CA certificate, it is said that they can be "trusted". In terms of certificate parsing and processing, this much is conventional wisdom and yet the *point* of chaining CAs together has not been obvious, and confusion has reigned over the types of bodies thought most apt to have custody of Root CAs.

The plain English synonym for Root CA certificate, "trust anchor", happens to be suggestive of a precise and powerful new type of role for Root CAs, which derives from existing audit and control structures. Obviously, most PKIs already embody various forms of audit, against standards that vary in rigor from one industry to another. But regardless of the details of a particular audit methodology, it is possible to gauge whether or not the audit has been conducted in a manner that is suitable for its environment. In the field of technical inspection or "conformity assessment", the pivotal question of 'who audits the auditors?' has long been addressed by a nested system of international inspection and accreditation standards. Across a very wide range of technical domains – from traditional materials testing to independent software validation – the standard ISO 17025:1999 *General Requirements for the Competence of Calibration and Testing Laboratories* [16] has been applied in the accreditation of inspection bodies.[24] The outcome of such accreditation is an assertion that a given inspection body is independent and competent to carry out audits in its field of expertise, no matter what that field may be, and regardless of the peculiarities of the standards that apply to that field. This outcome of auditor accreditation coupled with the fact that national ISO 17025 accreditation bodies have established multilateral *Mutual Recognition Arrangements* (MRAs) enables a new way of achieving "interoperability" between PKIs, as we shall soon see.

Multilateral MRAs are nowadays administered by overarching "co-operations" to which numerous national accreditation bodies are signatories. Examples include the Asia Pacific Laboratory Accreditation Cooperation (APLAC) and the International Laboratory Accreditation Cooperation

[24] In the first expression of the accreditation-based PKI model [25], I suggested that quality management systems were a suitable model for what CA auditors do, and that the peak standard ISO/IEC Guide 62 *General requirements for bodies operating assessment and certification/registration of quality systems* could be applied. More recent research indicates that ISO 17025 (form-erly known as ISO/IEC Guide 65) is a better fit for PKI auditors because it tests the competence of auditors and is generally more apt for technically demanding fields. It should be noted that a number of superficially similar accreditation standards exist including ISO/IEC 17020:1998 *General criteria for the operation of various types of bodies performing inspection*. To build a working PKI using these standards, a technical choice of high order standard would be made by expert standards bodies.

(ILAC).[25] Just as there are standards like ISO 17025 for 'auditing the auditors', recently another level has been added to govern accreditation bodies. For example, ISO 17011:2004 *General requirements for accreditation bodies accrediting conformity assessment bodies* is used by APLAC as the basis for its MRA. Accreditation bodies can join the MRA under a number of headings according to the broad focus of their activities, such as *inspection*, *calibration*, *testing*, or *reference materials production*.

Note that the accreditation of inspection bodies can be fine tuned to meet particular needs. When a certain field demands special skills and qualifications, ISO 17025 needs to be interpreted in respect of the type of inspection at hand and any special techniques involved. When a specialist field has its own conformity requirements, as might be set by local standards, accreditation bodies can produce *supplementary requirements* for accreditation of particular types of inspection bodies, to augment the general requirements of ISO 17025. Thus for example, Australia's National Association of Testing Authorities publishes detailed Accreditation Requirements, all based on ISO 17025, but fine tuned to a wide range of domains, including information technology.[26]

In relation to PKI governance, it is especially noteworthy that the information security community has already successfully applied ISO 17025 accreditation to overseeing the ISO 15408 Common Criteria evaluation scheme. The Common Criteria arrangement [11] requires security evaluators to be accredited in accordance with ISO 17025.

4.4.3 Scaleable global PKI from an ISO 17025 MRA

One reason that the Common Criteria scheme as been uniformly adopted with relative ease across 25 countries[27] is the existence of MRAs. The practical result of an MRA is that assessments done by accredited inspection bodies in one country can be readily accepted by interested parties in other jurisdictions. International trade in particular benefits from MRAs because goods that are subject to safety and type testing, such as electrical equipment, can be evaluated once in the country where they're made prior to export, and subsequently accepted by a large number of importers around the world without the need for repeat testing. APLAC describes this as the "free-trade goal of 'tested/inspected once, accepted everywhere'".[28]

The ability to accept and rely upon a specialised technical audit done in another country is surely the key to international PKI, if it is accepted that different digital certificates can mean different things, as argued throughout the earlier parts of this paper. ISO 17025 MRAs represent a hugely important asset in this regard because they can accommodate a flexible range of audit matters. Member accreditation bodies can be empowered under an MRA to implement new supplementary accreditation guidelines within a broadly defined scope such as "inspection",[29] and have the outcomes of their accreditations recognised in other countries, without them having to review in detail the substance of those guidelines. In turn, the outputs of organisations that have passed inspection under those new guidelines can be accepted across borders in other participating jurisdictions. Therefore, cross-border PKI could be constructed as follows.

The new PKI model treats digital certificates as the products of a special class of manufacturers, namely, RAs and CAs working in concert. The model also rests on the fact that from one jurisdiction (or industry) to another, reasonable decisions have already been made and broadly accepted regarding the appropriate standards that should govern RAs and CAs (including for example X.509, the PKIX series, RFC 3647, or the detailed *Identrus* technical requirements). Some jurisdictions and industries have gone one step further to select or design for themselves an appropriate PKI conformity assessment program. Examples include *Webtrust for CAs* (which was adopted by Microsoft as a pre-condition for being included in Internet Explorer's list of *Trusted Certification Root Certification Authorities*), *Gatekeeper*, *Identrus* accreditation (which

[25] ILAC members include American Assoc. for Laboratory Accreditation (A2LA; http://www.a2la.org), ACLASS Accreditation Services (http://www.aclasscorp.com), International Accreditation Service (IAS; http://www.iasonline.org), the United Kingdom Accreditation Service (UKAS; http://www.ukas.com) and Australia's National Association of Testing Authorities (NATA; http://www.nata.asn.au).

[26] See http://tinyurl.com/2yevj8.

[27] See List of Common Criteria Recognition Arrangement members at http://www.commoncriteriaportal.org/public/consumer/index.php?menu=4 (accessed 8 Jan 2008).

[28] See http://www.aplac.org/aplac_mra.html.

[29] See the various scopes of recognition for each of the APLAC MRA signatories at http://www.aplac.org/aplac_mra.html (accessed 8 Jan 2008).

has come to be recognised by *Gatekeeper*[30]), or various countries' local regulations implemented under the European *community framework for electronic signatures*[31].

Referring to the three preconditions for being able to accept a certificate cross-border, as defined in subsection 4.3.1, the proposed PKI model will deliver two[32] of them:

1. the conformity of the certificate issuer with agreed standards would be assessed by an approved PKI auditor, and the results of that assessment conveyed to the receiver, and
2. the intended purpose of the certificate would be precisely specified by its Policy OID (the uniqueness of which on the relevant domain would be enforced by the audit).

As described in subsection 4.4, to convey the results of the conformity assessment in this proposed PKI, special digital certificates will be issued by (or on behalf of) PKI auditors to each approved CA. The meaning of each these certificates is simply (but *precisely*) that the Subject has passed an audit according to detailed procedures and standards that would be uniquely indicated by a Policy OID. Different audit regimes and different auditors would map onto different OIDs.

Now, it is just as important that the status of the PKI auditors also be conveyed to receivers, and for that purpose, a special digital certificate will similarly be issued by (or on behalf of) an accreditation body to each accredited PKI auditor. In accordance with the regular provisions of ISO 17025, accreditation bodies would be chiefly concerned with the independence and the competence of PKI auditors. It may be the case that additional considerations, specific to PKI, are needed to be applied to make this determination, but as discussed, ISO 17025 accommodates this nicely. We should expect supplementary guidelines to be developed during the course of establishing the PKI model.

The chain of digital certificates corresponding to the different levels of conformity assessment could terminate at the accreditation body, with a self signed certificate. And yet the existence of international accreditation co-operations and MRAs presents the tantalizing prospect of cross border PKI resulting almost as a by-product of existing arrangements, with a conceptually simple switch from paper-based audit and accreditation certificates to digital certificates representing the same thing.

To operationalise the PKI, national accreditation bodies would act as jurisdictional Root CAs. It would not be necessary for these bodies to actually build and operate the CAs themselves; rather, they could outsource certificate production and concern themselves only with an RA. When looking at the potential legal liability of an accreditation body taking on the role of digital certificate issuer, we should be reminded that their proposed role in PKI is the identical to their role in traditional conformity assessment schemes; that is, they 'audit the auditors'. As such, their potential liability is well understood in industry, and tends to be well contained. If the digital certificates issued by accreditation bodies in this proposal are understood to mean nothing more and nothing less than the fact that the certificate Subject is an accredited PKI auditor, then the fact that the accreditation body is acting as a "Root CA" shouldn't introduce any new liabilities.

Figure 5 illustrates the proposed ISO 17025 MRA based PKI. The grey boxes represent the chief participants in the PKI, from CA through to international cooperation association (note that every one of these players already exists). The nested boxes show the scope or community of each participant: the smallest boxes are for the members served by each CA, intermediate for the CAs inspected by each auditor, and largest for the auditors inspected by each accreditation body (note how the communities are nested at the levels of country, auditor and CA). The block arrows indicate the framework that governs the work of each participant. At the top level, a working assumption is that of the existing types of MRA, an appropriate heading for PKI would be "inspection" (as opposed to *calibration* or *testing*).

[30] See http://www.dbcde.gov.au/Article/0,,0_4-2_4008-4_116523,00.html (accessed 8 Jan 2008).

[31] See http://europa.eu/scadplus/leg/en/lvb/l24118.htm (accessed 8 Jan 2008).

[32] The third precondition had to do with the certificate not being revoked; the ability to perform a CRL or OCSP check is taken for granted here.

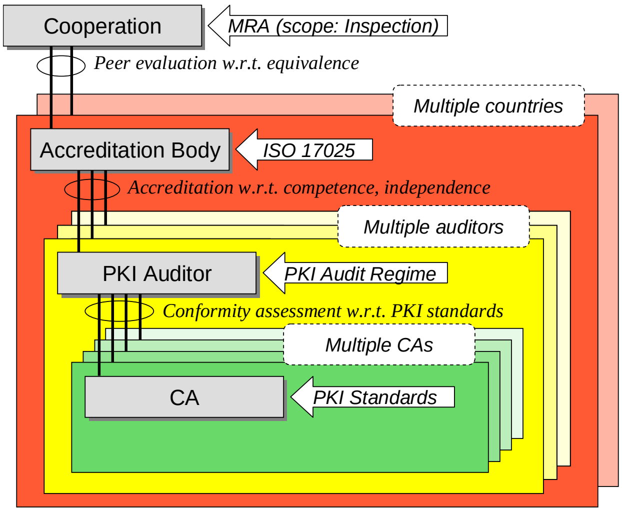

Figure 5: A PKI based on ISO 17025 Mutual Recognition

Note that the proposed PKI can be grown from the bottom up. It is not necessary for an international cooperation to come on board right away; in the interim it would be practical to have local self-signed trust anchors for each of the jurisdictional accreditation bodies. Whenever a new body joins an MRA and has its processes approved, it follows that all CAs within its jurisdiction automatically enter the fold of the international scheme. Thanks to their maturity and long established authority, having accreditation bodies in the PKI solves the hoary problem of infinite regress; that is, how far back do you go before you find a CA you can "trust"? The answer is you stop at a national body, or ultimately at an international cooperation like APLAC or ILAC. Most important of all, because there are existing protocols and agreements by which national accreditation bodies recognise and work with one another, this approach to PKI provides a natural and robust means for cross-border recognition of digital certificates.

In closing this account of an international PKI, let us remember what it is that a certificate chain can represent. If the receiver of a digital certificate knows what end user Policy OID is appropriate to the transaction at hand, and if the receiver's software has a trusted copy of the root key, then any certificate featuring that OID which chains to that root key can be taken to be fit for purpose, no matter which CA issued it. Certificate chains in this international PKI scheme would each embody an unambiguous cascade of dependable assertions:

- The end user was vouched for with reference to a certain CP by an RA authoritative in a given community, and was issued a certificate with a corresponding unique Policy OID, produced by a named CA.
- The CA was approved with reference to agreed standards, CPS etc. by a named PKI auditor, which issued (or had issued on its behalf) a digital certificate to the CA.
- The auditor was approved with reference to ISO 17025 by a named accreditation body, which issued (or had issued on its behalf) a digital certificate to the auditor.
- The accreditation body was approved with reference to a Mutual Recognition Agreement by a named international cooperation, which issued (or had issued on its behalf) a digital certificate to the accreditation body.

The certificate chain conveys the membership of all participants in the scheme as anchored by the root key controlled by the top level cooperation. And yet certificates that chain through different auditors and accreditation bodies are entirely autonomous. Neither the root nor the intermediate accreditation bodies would impose any arbitrary policies on the conduct of end user CAs and communities of interest. The uniqueness of the end user certificate Policy OIDs plus the separation of powers of auditors and accreditation bodies means that the one PKI could embrace any number of diverse communities, and could accommodate existing closed PKI programs like *Identrus* or *WebTrust for CAs* so long as their methods are transparent and compatible with ISO 17025.

5. CONCLUSIONS

There is no intrinsic reason that PKI should be as complex as it has been. Plenty of complicated physical principles have been successfully engineered and deployed as commercial technologies, such as magnetic stripe cards. PKI historically has been unwittingly burdened by well intended metaphors, such as that of the passport, that associated it with a vague ideal – universal identification of strangers online – which turned out to be hugely complicated and not even necessary. Meanwhile, smaller scale, closed PKIs have prospered in support of special purpose applications. This fact can now be appreciated as the natural state of affairs, resonant with the modern view of identity plurality.

We should build on the deeper lessons of successful closed PKIs, to regard certificates as evincing not absolute identity but rather any number of relationships, with special meaning in the contexts in which the certificates were issued and intended to be used. This simple change of aspect could herald a true paradigm shift, rendering digital certificates and their production much more mundane. Radical improvements would result on several fronts. Firstly, the practical application of PKI would be greatly simplified by breaking the nexus between authentication and authorization, for it allows X.509 formatted Relationship Certificates to stand alone in most transactions. Secondly, by localizing RA functions and more effectively decoupling certificate production, we could operate back-end CAs along the same lines as security printers, with vastly simpler legal arrangements than seen in orthodox PKI. And finally, existing nested frameworks for conformity assessment and accreditation provide the ready means for cross-border recognition of certificates, knitting together today's heterogeneous PKI applications, policies and audits into the one international Public Key Superstructure.

6. ACKNOWLEDGMENTS

The concept of "Relationship Certificates" was originally researched and developed by Lockstep Consulting under contract to the former Australian Department of Finance and Administration, represented by the Australian Government Information Management Office (AGIMO). The author gratefully acknowledges the permission of AGIMO to reproduce aspects of that work.

7. REFERENCES

[1] APEC Telecommunications Working Group – E-Authentication Task Group, Achieving PKI Interoperability (1999).

[2] APEC Telecommunications Working Group, Electronic Authentication: Issues Relating to its Selection and Use ISBN 981-04-7690-6 (2002).

[3] Australian Department of Health and Ageing, Electronic Signatures for Prescribing and Dispensing, eHealth Branch (2006) http://www.msia.com.au/esig_prescript_document.pdf (accessed 31 Jan 2008).

[4] Australian Government Information Management Office (AGIMO), Gatekeeper PKI Framework Cross Recognition Policy, Department of Finance and Deregulation (2008) http://www.gatekeeper.gov.au/__data/assets/file/0004/52276/Cross_Recognition_Policy.rtf (accessed 8 Jan 2008).

[5] Australian Government Information Management Office (AGIMO), Relationship Certificate Guidebook, Department of Finance and Administration (2006) http://www.agimo.gov.au/__data/assets/pdf_file/0016/52252/Relationship_Guidebook.pdf (accessed 19 Nov 2007).

[6] Barnes, R. & Kent, S. An Infrastructure to Support Secure Internet Routing, IETF Secure Inter-Domain Routing Working Group (2007) http://tools.ietf.org/id/draft-ietf-sidr-arch-00.txt (accessed 31 Jan 2008).

[7] Barnett, S. A pilot project: Sending encrypted specialist letters to GPs, *Health Openware Foundation Argus Forum*, (Canberra, Australia, 2004).

[8] Brewer, J. & Wilson, S., Smartcards and PKI at Medicare Australia, Australian Electrical & Electronic Manufacturers Association ICT Forums (2006) http://www.aeema.asn.au/ArticleDocuments/41/Smartcards%20and%20PKI%20at%20Medicare%20Australia%20-%2014Feb06.pdf (accessed 31 Jan 2008).

[9] Burr, W. Electronic Authentication in the U.S. Federal Government, *Asia PKI Forum*, Tokyo (2005) http://www.asia-pkiforum.org/feb_tokyo/NIST_Burr.pdf (accessed 23 Nov 2007).

[10] Cameron, C. The Laws of Identity, Microsoft Corporation (2005) http://www.identityblog.com/stories/2005/05/13/TheLawsOfIdentity.pdf (accessed 31 Jan 2008).

[11] Common Criteria, Arrangement on the Recognition of Common Criteria Certificates in the field of Information Technology Security (2000) http://www.commoncriteriaportal.org/public/files/cc-recarrange.pdf

[12] Ellison, C., Frantz, B., Lampson, B., Rivest, R., Thomas, B., & Ylnen, T. SPKI Certificate Theory RFC 2693, IETF SPKI Working Group (1999) ftp://ftp.isi.edu/in-notes/rfc2693.txt (accessed 31 Jan 2008).

[13] Ellison, C. & Schneier, B. Ten Risks of PKI: What You're not Being Told about Public Key Infrastructure *Computer Security Journal 16, 1* (2000).

[14] Forey, M., Cheque Printer Accreditation Scheme, *Xplor Document Management Conference* (Anaheim, California, USA, 2002).

[15] Freeman, R., Trust Services – A Market Appraisal, Mack Interact (2002) http://www.tscheme.org/library/tSi0156_01%20TSP%20market%20status%20report.pdf (accessed 19 Nov 2007).

[16] International Organization for Standardization, General requirements for the competence of testing and calibration laboratories, ISO/IEC 17025:1999.

[17] Jøsang, A. & Pope, S., User Centric Identity Management, *AusCERT Security Conference 2005* (Gold Coast, Australia) (2005).

[18] Kent, K. Global PKI: Status, Trends and the Future *Taipei International PKI Conference* (Taipei, September 2005) http://www.pki.org.tw/pkiforum2005/d_file/01_Stephen%20Kent.pdf (accessed 23 Nov 2007).

[19] National Office for the Information Economy, Liability and Other Legal Issues in the use of PKI Digital Certificates, Australian Department of Communications, Information Technology and the Arts (2000).

[20] Nazareth, S. & Smith, S. W. Using SPKI/SDSI for Distributed Maintenance of Attribute Release Policies in Shibboleth, Computer Science Technical Report TR2004-485, Dartmouth University (2004) http://www.ists.dartmouth.edu/library/spk1004.pdf (accessed 31 Jan 2008).

[21] Smith, P., Trust and Digital Certificates, *16th Payment Systems International Conference* (Belgium, 2000).

[22] Sneddon, M. Legal Liability and e-transactions, Australian Department of Communications, Information Technology and the Arts (2000). http://www.egov.vic.gov.au/pdfs/publication_utz1508.pdf (accessed 23 Nov 2007).

[23] Standards Australia, Strategies for the implementation of a Public Key Authentication Framework (PKAF) in Australia, Miscellaneous Publication MP 75 (1996).

[24] Victorian Office of the Chief Information Officer, Land Exchange (LX) Case Study, Government of Victoria (2004) http://www.egov.vic.gov.au/pdfs/Land%20Exchange-shh-30April-v1.0-CIO.pdf (accessed 21 Nov 2007).

[25] Wilson, S. New models for the management of public key infrastructure and root certification authorities. In *Proceed-ings of Information Security Management & Small Systems Security (IFIP WG 11. 1/2)* (Amsterdam, Sept 30 - Oct 1, 1999). Kluwer, Deventer, The Netherlands, 1999, 221-230.

[26] Wilson, S. Leveraging external accreditation to achieve PKI cross-recognition, *Australian Attorney General's Privacy and Security in the Information Age Conference* (Melbourne, Australia, 16-17 Aug 2001) http://www.ag.gov.au/www/agd/agd.nsf/Page/Privacy_PrivacyandSecurityintheInformationAgeConferencePapers (accessed 31 Jan 2008).

[27] Wilson, S. Guidelines on how to determine Return on Investment in PKI, OASIS PKI Education Sub-committee, V 1.4 (2005) http://idtrust.xml.org/guidelines-how-determine-return-investment-pki (accessed 14 Jan 2008).

[28] Winn, J. K. The Emperor's New Clothes: The Shocking Truth About Digital Signatures and Internet Commerce, *37 Idaho L. Rev. 353* (2001)