SCRIPSIT: A model for establishing trustable privacies

in

online public spaces

Paul Trevor-John Rodda (851853046)

Submitted in partial fulfilment of the academic requirements for the degree of

Master of Arts (Digital Media)

Centre for IT in Higher Education

University of KwaZulu-Natal (Howard College)

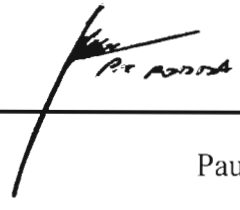
December 2004

Supervisor: P.A. Clarke (UKZN)

Declaration

I hereby declare that this dissertation represents original work by the author, and has not been submitted to another university. Where use has been made of the work of others, this has been duly acknowledged.

Signed



Paul Trevor-John Rodda, 15th of December 2004

Acknowledgements

I would like to acknowledge the contributions of following:

- Patsy Clarke, my supervisor, for her forbearance, guidance and encouragement throughout the process of discovery, reflection and development,

- Dr. Nicola Rodda, my wife, for unstinting tolerance and support of a mildly grumpy individual through the course of this project, and my children, Robert and Benjamin, for teaching me that all things are possible,

- Dr. Louise Corti of Qualidata (Essex University) for her commentary and encouragement, face to and via email,

- Drs. Heppell and Preece for participating in impromptu interviews at ED-Media 2004.

- Dr. Patrick Carmichael (Cambridge University) for his time and commentary both via email and in person in Cambridge.

- Prof. Gunnar Liestøl (Oslo University), Prof. Ron Krabill (University of Washington at Bothwell), Rose Quilling (UKZN), Elva de Sibandze Gomez (UKZN), Clarissa Muir (RAU), Dr. Dale Peters (UKZN), Katherine Dubbeld (UKZN) for their willing and valuable participation,

- A long list of "Others" that have variously contributed opinion, criticism (almost all constructive!) and encouraged pursuit of this line of enquiry.

- Prof. Alan Amory for his cunning encouragement and occasional goading. There are few who "read" people as well as he does.

- Not forgetting Dagmar Vader, for having a crucial argument (debate?) on aspects of secondary use of qualitative research data with me over a beer, one Friday afternoon, late in 2002. This event triggered a train of thoughts and questions which culminated in my original project proposal and, ultimately, in this dissertation.

Thank you to all of the above – I have thoroughly enjoyed my enquiries and pursuit of knowledge. I am the better person for having had to defend my position.

Paul Rodda

Abstract

This dissertation proposes a model supporting the creation of trustable privacies in public online spaces, with the model demonstrating the potential for supporting trustable data handling in the qualitative domain. Privacy and trust, from the pivotal perspective of the individual were identified as crucial intangibles in the qualitative research and personal trust domains. That both privacy and trust depend heavily upon credible mechanisms for privacy became clear during the literature review and interview processes.

Privacy, in its many forms, is a concept requiring greatly varying degrees of anonymity, confidentiality and control (Rotenberg, 2001; Lessig, 1998) and this was position was validated by literature and by qualitative comments by academic interviewees.

Facilitation of secondary users including academics, public and private organisations, communities, casual information browsers is a goal of this research. This goal of facilitation is supported by the model proposed, and is discussed in Chapter 6, where future work is discussed. The core requirement to address confidentiality, ethics, privacy, ownership and control of data (Corti, 2000) is satisfied by the model as proposed and discussed.

Expected outcomes of this research project are summarised as:

- Proposed model for the creation of trustable privacies in public spaces. [Primary outcome]
- Promotion of collaboration amongst domains and disciplines through improved universal access to archived data [Secondary outcome]
- Identification of application domains outside of the initially identified domain set [Secondary outcome]

Self-Contained ReposItory ProcesSIng Template (SCRIPSIT) describes a model supporting a decentralised, trustable set of structures and mechanisms. SCRIPSIT has its eponymous origin in the Latin word scripsit, meaning "*he or she wrote*".

Table of Contents

Tables

Figures

Glossary

Action Research

Applied research that join up practitioners with researchers in a research partnership. Emphasis here is on ongoing improvement of practice by the practitioners themselves. Rose (2000) provides applied examples of action research in information systems.

AGENT

A piece of software that runs without direct human control or constant supervision to accomplish goals provided by a user.

Applied Research

Research done with the intent of applying results to a specific problem. Evaluation is a form of applied research. This can be conducted as part of an action research approach.

Base64

7-bit encoded data, consisting only of printable characters. Less efficient that 8-bit encoding when used for binary objects. Preferred for use in SCRIPSIT for reasons of portability. See UTF-7.

Blogs

Online personal logs of absolutely anything. Usually for public consumption (abbreviated form of 'Web Log').

Browser

Program used to look at (or browse) WWW/Internet resources

CAQDAS

Computer Assisted Qualitative Data Analysis Software. A class of software tools facilitating the marking up/annotation of rich qualitative data sources (primarily textual) so as to support their use in qualitative research applications.

CGI

Common Gateway Interface. A standard means of extending Web functionality through execution of scripts on a Web server, in response to browser requests.

DCMI

Dublin Core Metadata Initiative. A range of networked metadata entities.
Digital certificate

DTD

Document Type Definition. Describes content and structure of a class of XML documents.

Encryption

Securely concealing the contents of a message in such a manner that the message is useless if intercepted by a party without the means (the key) to decrypt or decode the message. Encryption is fundamental to secure storage and transmission of data on the WWW.

ENGINE

The embedded script engine key to SCRIPSIT's peer-centric model

Fragment identifier

The part of a URI that allows identification of a secondary resource.

Grounded Theory

Grounded theory is a research method that seeks to develop theory that is grounded in data systematically gathered and analyzed.

Hashing algorithm

Checksum calculated on a private key, used to confirm that a piece of data has remained unaltered.

HIVAN

The HIV AIDS Network, based at University of KwaZulu-Natal (Howard College)

HTML

HyperText Markup Language. The language used to encode formatting, links and other features on Web pages.

HTTP

HyperText Transfer Protocol. Internet protocol used to manage communication between Web browsers (clients) and web servers.

HTTPS

HTTP Secure. An encrypted HTTP link.

Internet, The

Global computer network of connected server computers.

ICT

Information and Communications Technology

IVR

Interactive Voice Response system

Java

Platform independent programming language created by Sun Microsystems. Java can be used to create Java 'applets' or small applications on the Web. Java has achieved near-ubiquitous presence across many computing platforms.

JavaScript

Scripting language useful for handling interactive features in HTML. Scripts are executed from the browser on the client machine

KMS

Knowledge Management System. A codified means of making (usually) domain-specific structured information available to requestors. Usually requiring domain expert at one or more of the stages of encoding, requesting and interpreting content.

Link

A relationship between two resources where one resource refers to the other by means of a URI.

Metadata

Metadata (or "data about data") may be used to label and categorise data for searching and processing. A formal alternative is "structured descriptions of resources". Metadata form the matrix enabling large collections of data to function as organised libraries, which seldom exist as single instances.

Metastructures

Structures describing structures, usually in the abstract. Key to loosely associated collections of metalinked data on the WWW. Metastructure refers to the overarching framework supplying rules defining relationships amongst meanings within a domain or information category.

[XML] Namespaces

An XML namespace is a collection of names, identified by a URI reference, used in XML documents to specify element types and attribute names. XML namespaces are distinguished from namespaces used elsewhere computing disciplines by the fact that the XML versions have internal structure and are not sets in mathematical terms.

Navigation

The process of moving from place to place, particularly in a hyperlinked environment.

Online Public Spaces

Derived from the civil society domain, online public spaces refers to an individual's engagement with common or shared spaces in the digital world, viz. chat rooms, email, file servers and any other digital resource linked via the Internet.

Ontology

Set of concepts (things, events, and relations) specified in some way (such as specific natural language) in order to create an agreed-upon vocabulary for exchanging information.

OWL

Web Ontology Language (http://www.w3.org/2001/sw/WebOnt). See Semantic Web and *The House at Pooh Corner* by A.A.Milne (1926).

P2P

Peer-to-Peer

P3P

Platform for Privacy Protection.

Peer-centric

A term coined during the development of the conceptual base of SCRIPSIT. Peer-centricity describes the processing of data and assertion of control by the data consumer at a local level only. In a peer-centric data collection, processing occurs only *after* a data-bearing entity has reached the requester.

Peer-local

Peer-local refers to the fact that decrypted data only ever exists in a transient form on the requesting client.

PETs

Privacy enhancing technologies. Protocols, tools and processes which enhance individual privacy through measures which counter unauthorised attempts to intercept or otherwise abuse access to private data.

POTS

Plain Old Telephone System. Term for old-fashioned landline telephone systems

Protocol

Formally defined set of rules and formats. Computers use protocols to regulate communications

Privacies

Abstract area or domain where data remains intact and unrevealed except with the consent and permission of the nominated data owner.

Proxy server

Internet server acting as a firewall, mediating traffic between a protected network and the Internet.

Public key encryption

Encryption mechanism using complementary public and private keys to encrypt and decrypt data

Qualitative Research

The approach advocated by the interpretive school as a means to understanding social phenomena. Generally viewed as any kind of research that produces findings not arrived at by means of statistical procedures or other means of quantification, and includes in-depth interviews, observations and participant observation.

Quantitative Research

The approach advocated by the Positivist School. This approach measures social phenomena and obtains numerical values which can be analyzed statistically. Surveys using structured questionnaires and IQ tests are both examples of quantitative research.

QRD

Qualitative Research Data.

RDDL

Resource Directory Description Language (http://www.tbray.org/tag/rddl/rddl3.html)

RDF, RDF triples

Resource Description Framework (RDF) data consists of nodes and property/value pairs describing nodes. A node is any object that can be pointed to by a URI. Properties are attributes of nodes; values are either atomic values for the attribute or other nodes. Information about a research topic (a node), may include the property "Owner". The value for the Owner property may be a string of text, a URI pointing to another document or a persona definition. RDF defines metadata processing frameworks and data models based on triples (subject/resource, predicate/property, object/property value). Data graphs with unique identifiers may be formed with these

data triples. RDF forms the basis of tools able to link, classify and extend data and add subjective value. An example is the aggregation of a collection of XML documents into an RDF model. Document collections may be complete and fully formed, they may be data fragments and they may also be networks of multiply linked XML documents. This forms the essential basis of RDF/XML used as dynamic and extensible repositories. Semantically dependent queries against knowledge encoded in ontology are available via RDF/XML document networks.

Reification

Recasting of a statement. An example is found in recasting the statement that "John is a boy" as "John's mother believes that he is a boy".

Resource

A resource is anything that has identity. Examples include documents, images, services (news reports, weather information). Not all resources are retrievable (people, institutions and printed papers) across the WWW. The resource is the conceptual mapping to an entity or set of entities, and not necessarily the entity corresponding to the mapping at a specific time. A resource can therefore remain constant whether or not the entities to which it corresponds change over time. This is predicated on conceptual mapping of the resource remaining constant.

Resource Discovery

Process by means of which a specific resource or class of resources are discovered on the WWW.

Sandbox

A limited (software) environment which prevents programs from reading or writing files, initiating or accepting network connections with any system other than the originating server, running local programs, overwriting or emulating local program content and so forth.

Schema

A description of the structure of a database or other data source. In the context of XML, the schema refers to a definition of the structure of a class of XML documents.

SCRIPSIT

Self-Contained ReposItory ProcesSIng Template. SCRIPSIT describes a model intended to support trustable, resilient, persistent, peer-centric and serverless meshes or networks of encapsulated nodes.

Script

Code which may be directly executed by a program (or engine) that understands the language in which the script is written.

Semantic Web

The representation of data on the World Wide Web. A collaborative effort led by W3C with participation from researchers and industrial partners. Based on the Resource Description Framework (RDF), which integrates a variety of applications using XML for syntax and URIs for naming

Service discovery

The process of locating an agent or automated Web-based service that will perform a required function. Semantics will enable agents to describe to one another precisely what function they carry out and what input data are needed.

SGML

An ISO (International Standards Organisation) markup language for representing documents on computers. HTML is based on SGML.

SOAP

Simple Object Access Protocol (http://www.w3.org/tr/soap/)

SSL
Secure Socket Layer

Survey
A method of collecting information from a usually large sample of the population of interest. This is usually a quantitative method which allows statistical inferences to be drawn from the sample about the population.

Triangulation
Using multiple methods and/or data sources to study the same phenomenon. The idea here is for the weaknesses in any one method to be compensated for by the strengths of another. The researcher addresses the issue from different methodological positions, rather like taking photographs of the same subject from different angles to reveal a more valid picture of what the object actually looks like.

Trustable
Able to promote and/or inspire confidence in third parties that the service/resource will act to preserve confidentiality and privacy with respect to deposited data.

UDDI
Universal Description, Discovery and Integration (http://www.uddi.org)

URI
Universal Resource Identifier. URLs are the most familiar type of URI. A URI defines or specifies an entity, not necessarily by naming its location on the Web.

URI aliases
Two or more different URIs that that identify the same resource.

URI ownership

A relationship between a URI and a social entity, such as a person, organisation, or specification.

URI reference

An operational shorthand for a URI.

URL

Uniform Resource Locator. The familiar codes (such as http://www.sciam.com/index.html) that are used in hyperlinks.

UTF-7

7-bit character representation (printable characters only) – see Base64. UTF-8 and UTF-16 are 8- and 16-bit character set representations.

W3C

Worldwide Web Consortium (http://www.w3c.org)

Well-formed

Describing a document conforming to the syntax rules of XML.

WWW

World Wide Web. A large-scale, interlinked, global system of distributed hypermedia resources with a graphical interface that can be accessed and from which information can be selected for retrieval to a local computer.

VM

Virtual Machine – a self-contained environment in which applications (usually in a scripting language) may execute in a controlled manner. Java implementations are usually realised as VM environments allowing identical execution of scripts on a variety of computing platforms.

XML

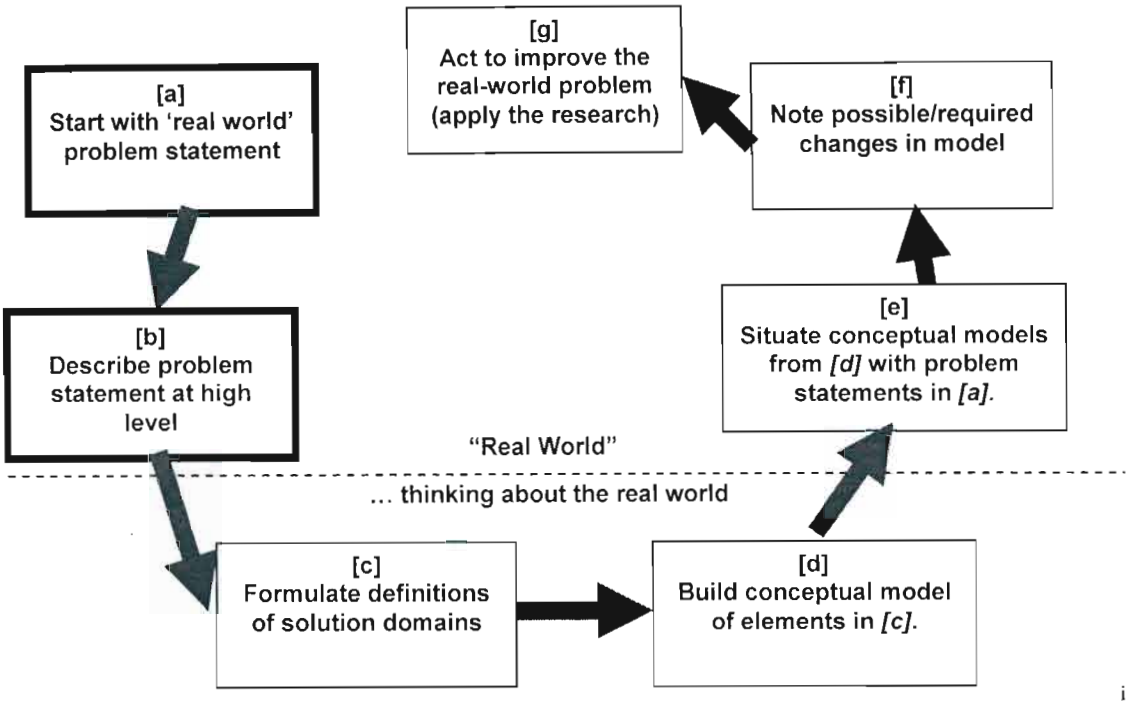
eXtensible Markup Language. A markup language allowing definition and use of customised document tags. Goldfarb (2000) succinctly defines XML as "… smart data HTML tells how the data should look, but XML tells you what it means but XML data isn't just smart data, it's also a smart document and you don't have to decide whether your information is data or documents; in XML, it is always both at once. You can do data processing or document processing or both at the same time"

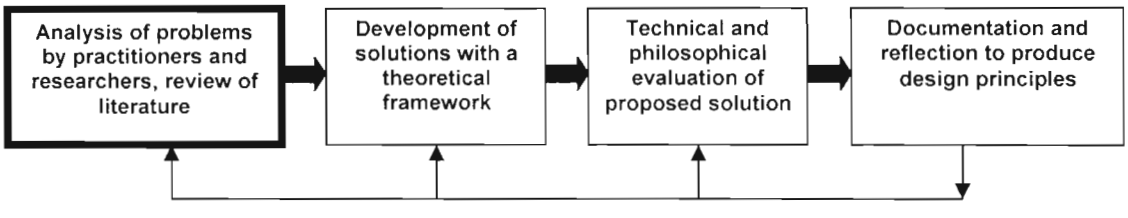
Xpath

XML Path Language. Intended for addressing parts of XML documents.

Chapter 1

Introduction to dissertation, background, problem statement and rationale





[i] The figures on this chapter preface page are repeated for each chapter in this dissertation. These refer to research methodologies described in Chapter 3. Please refer ahead for full discussion on methodologies.

1.1 Background and purpose

This dissertation presents the results of research that aimed to propose mechanisms, methods and models for the preservation of privacy, ownership, trust and anonymity, along with context and intended meaning, in published qualitative data. The initial scope of the research was defined by broadly perceived requirements of, and solutions to, issues surrounding secondary reuse of qualitative research data. A wider set of potential applications became apparent on reflection and research, hence the development of a working title of:

SCRIPSIT: A model for establishing trustable privacies in online public spaces

To describe privacy and trust, a common understanding must first be established. Brunk (2002) comments that privacy, in its many forms, is a concept embracing greatly varying degrees of anonymity, confidentiality and control.

Context and meaning in the minds of the original researchers (and/or users) and research subjects is largely inaccessible by secondary users. Secondary users include academic, public and private organisations, communities, casual browsers and data reusers. Corti (2000) refers to users of historical data repositories. Medical users, including doctors, nurses, midwives and health administrators are identified as an additional group of non-traditional secondary qualitative data users (Williams, 2000: cited in Fielding, 2000).

Adequate addressing of confidentiality, ownership, trust, ethics, privacy and management of data is of overriding importance. Barber (1983) asserts that trust is made up of a set of social expectations about self, institutions, nations and societal orders.

Considerations of privacy, of trust and the mechanisms which might support these will pivot on the central question asked in this dissertation, derived from Bromseth's

(2002) posing of the question of who ought to be responsible for the protection of an individual's privacy.

The central question asked, therefore, is who has the right and responsibility of protecting the privacy of the individual. Following on from this question are those which arise from investigations into mechanisms and models which might support such user-centric rights and responsibilities. These questions will guide the selection f much of the supporting literature to be reviewed.

1.2 Rationale for this research

The rationale and motivation for this research is based simply and primarily on the following considerations:

1. Secondary access to qualitative research data, especially in the social sciences, presents challenging situations with respect to data access and reuse (Corti, Day and Backhouse; 2000). Collaborative access to and use of qualitative research data includes a broad mix of producers and users. The nett result is a requirement for a self-contained and robust means of mediating, controlling and managing access to online and archived qualitative data (Walkerdine, Melville and Sommerville, 2002). Inadequate support for non-commercial applications (academic use and cradle-to-grave personal data, amongst others) prompted the initial investigations.
2. Successful applications of Knowledge Management Systems (KMS) tend to occur in vertical, domain-limited arenas (Eberhart, 2004). Users in these cases are either domain experts or use such experts as intermediaries. There is a shift in priorities and problems with research data users and reusers who are not domain experts. Reliance on trusted third parties is a significant and fundamental failing of existing mechanisms proposed and used for trustable archival purposes (Fitzgibbon and Reiter, 2003).

3. Relevance to the aims of regionally-located interests, including Research Africa and HIVAN, has been established in discussions with interested parties. Responses to a poster presented at *Idlelo - The First African Conference on the Digital Commons* conference in January 2004 in Cape Town (Rodda, 2004), and to unpublished parts of the research suggested that applications exist for the expected research products of this project. This poster was a follow-on to an earlier poster presented at the 5th Annual World Wide Web Applications conference at the University of Durban-Westville (Rodda, 2003).

An exploratory review of literature suggested a lack of models for user-centric privacy and trust which did not rely on third-party support and services.

1.3 Statement of problem

The practical problem attributed with catalysing the research presented here is that of determining a feasible and implementable model supporting the creation of private and inviolate spaces within the boundaries of universally and publicly accessible domains (or spaces).

The model to be developed is required to address the attributes argued as defining universal (and general) concepts of privacy and trust. There exists a problem with assertions and demands of trust, especially from self-appointed third parties. Removing requirements for placing of trust or control in third parties permits higher levels of real and perceived trust, and hence the birth of privacies which exist free of the fetters of organisational and political self-interest.

Extending the argument to the qualitative research data reuse paradigm, questions arise as to how to build, maintain and ensure levels of believable trust in security, reliability, persistence and access to private spaces touched by the research activities.

1.4 Research questions and objectives

Research objectives:

- Defining the extent of publicly accepted and understood privacy.
- Securing of privacy in a public arena
- Consider praxis in preservation and reuse of qualitative research data.
- Ethical considerations with respect to public access to private data.
- Assessment of gaps in praxis.
- Review of standards, testing suitability for use in model to be proposed.
- Preservation of context and meaning.
- Separation of data and storage mechanisms.
- Establishment of basis for proposing a model for trustable privacies.

Research questions

- Where is the most appropriate place for vesting of control over private data?
- Is it feasible to develop a model supporting creation of trustable privacies, with application across multiple domains?

Dissertation Objectives

- Draft model for establishment of trustable privacies in public spaces.
- Outline of the basis of further appropriate research in this domain.
- Consideration of potential future work.

1.5 Intended audiences and beneficiaries

Intended audiences for this research are listed in primary and secondary categories. Primary audiences are those who are directly linked to the qualitative researchers and research subjects originally identified. Secondary audiences are those secondarily or indirectly associated with one or more of the primary audiences listed above.

- Primary audiences
    - Academic communities
        - Qualitative researchers
        - Research subjects and communities
        - Archivists
        - Cross-disciplinary collaborations
    - Sectoral research institutions
        - Non-governmental organisations (NGOs)
        - Policy-making governmental bodies

- Secondary audiences
    - Displaced persons
    - Diasporic communities
    - Self-archivers
    - Informal online communities

1.6 Dissertation outline

Chapter 1 offers an overview of the problem statement and a brief exploration of the rationale behind the problem statement in this dissertation; research questions posed and intended audiences for this dissertation.

Chapter 2 reviews literature relevant to the domains touched upon by privacy, public access, data security, confidentiality and consent, knowledge management systems and public access to qualitative data. This literature review considers research concerning representations of qualitative data and protection of privacy and associated domains. A review of published standards for document structures, representation of context and privacy is included in this chapter. The problem is also situated within the domains of research, ethics, privacy and trust.

Chapter 3 discusses research methodologies employed. Discussion of theoretical frameworks follows, including applications of qualitative data and support of personal privacies in public spaces.

Aspects of theoretical frameworks:

- Research practices in qualitative data access and use.
- Extension of research paradigm for access and use of 'soft' data.
- Location of this research in the contemporary research arena.
- Key features of research models applied.
- Representation of meaning and context.
- Enquiry goals for this research.
- Framework for methodologies.
- Questionnaires and interviews.

Chapter 4 considers tools used for access to and manipulation of qualitative research data and mediation of public access to such data. Existing approaches to secondary use and mediation of access to qualitative data are considered here. Linking of Web Ontology Language (OWL), domain ontologies and the Semantic Web are considered with respect to resource discovery considerations. Knowledge Management Systems (KMS) are additionally challenged as appropriate models with respect to application as bases for both the representation of intended meaning and for the creation of trustable privacies. Other mechanisms available for access mediation and management of qualitative data are considered, with the intention of distinguishing fashion usage from fitness for purpose. Hendler (2003) offers a comprehensive Frequently Asked Question (FAQ) reference on OWL.

Tools and considerations with respect to access and use of Qualitative Research Data (QRD):

- Capabilities and restrictions in tools for QRD reuse.
- Capabilities and restrictions of Knowledge Management Systems.
- Contemporary and historical practice.
- Tools and reasons for their use.
- Gaps and the means to fill them.
- Ontological support in contemporary tools.
- Domain epistemology and representation thereof.
- Observations of contemporary initiatives and motivations.

Chapter 5 presents the results of the research in the form of a model for a peer-centric, decentralised means of publishing confidential data in public spaces. An overview of structures and model elements proposed as theoretical guidelines / solutions stated problems is included. A summary of arguments for and against aspects of contemporary praxis is presented, leading into an outline model in both theoretical and practical contexts. Expansion of the theoretical model is presented as an application of the model with respect to real-world requirements.

Describing the proposed peer-centric model:

- Model for combination of metastructures and metadata.
- Realisation of proposed model in a theoretical context.
- Realisation of proposed model in a practical context.
- Application of proposed model to a published requirement.
- Extension of proposed model paradigm.

Chapter 6 presents discussion of research and conclusions. The extent to which stated research questions are answered is considered with reference to the initial assumptions.

Practical implications and benefits coming out of this research are discussed. Notions of open access and protection of privacy and intended meaning are revisited in the context of the proposed model.

Summary of topics:

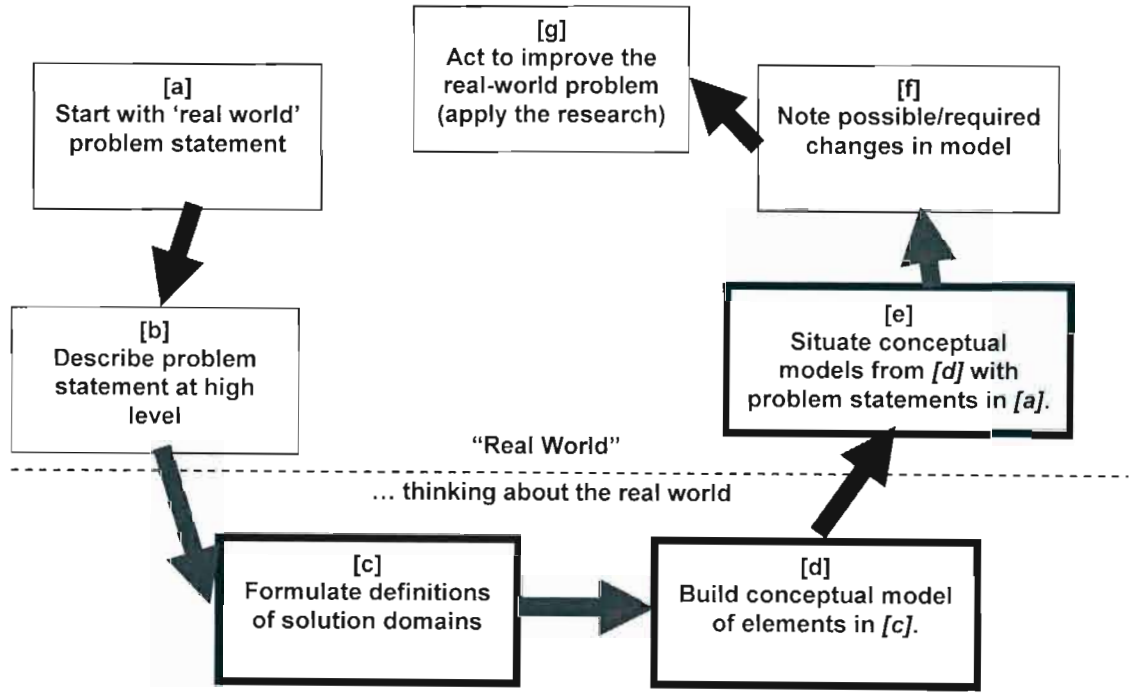
- Protection of Privacy and Trust.
- Metastructures for carriage and management of data.
- Discussion on research processes of this study.
- Limitations of this study.
- Implications of and future directions for this research.

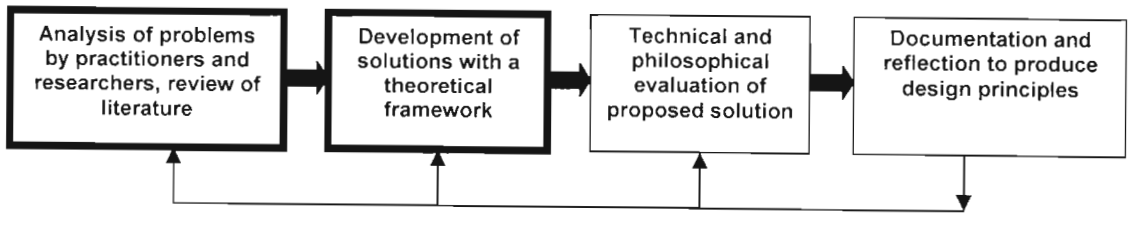
Metastructures for carriage and management of data are presented as appropriate mechanisms fulfilling the requirements of flexibility and robustness. There is reflection on processes and paths in this study. Finally, future direction is discussed.

Chapter 2

Literature Review

Consider current and past research with respect to representations of qualitative data and protection of privacy. Place the problem (as scoped for this project) within the domains of research, ethics and privacy, and related fields of endeavour.





2.1 Introduction

This dissertation considers literature in the domains of privacy, trust, ethical use of research data and derivation of privacy models.

Intangible currencies of privacy and trust underpin the requirements of mechanisms and models for creation of trustable spaces. These conflict with the accessibility and openness required of public data which ought to be in the public domain. Establishing trust requires a credible and robust protection of privacy (Castelfranchi and Falcone, 2000), no matter what the domain or environment. Privacy and trust in the digital world are perceptually and technically quite different from privacy and trust on a face-to-face level. There is almost always very little or no substantiation of the credentials of the so-called 'other party' in the digital world. Protection of individual privacy is to be discussed in a community context, with trust as a construction based on respect for and reliability of individual privacy.

Corti, Day and Backhouse (2000) discuss issues which impinge directly on privacy and trust. These are confidentiality, informed consent and anonymisation of data. A strong emphasis is placed on ethical actions around access to qualitative research data.

Qualitative data is that collected using qualitative methodologies (Corti, 2001). Such methodologies are usually characterised by an inclusive and accommodating nature. The qualitative perspective uses multiple approaches and methods, giving data sets which include diverse content, including structured and unstructured interviews, observations, audio and video material, still images. Secondary users of qualitative data engage in reappraisal and reanalysis of datasets. This may include reference to one or more related or unrelated data sets. Simple referential secondary use of qualitative data also occurs, with a wide number of user categories. Thompson (2000) comments at length on the conflicting uses and intentions of secondary users of qualitative research data.

The literature review concentrates on perceptual, conceptual and technical aspects of privacy, trust, resource discovery, ownership and related domains in individual and societal contexts. This is essential to establish some of the parameters and requirements for information representation and mapping at a technical level. All of this is required to establish conditions conducive to allowing individuals to be able to place a degree of trust in the model presented.

The flow through this chapter is listed here:

- Qualitative research and access to data in the public domain.
- Review of privacy and trust in qualitative data use.
- Review of privacy in the wider context.
- Review of qualitative data usage and issues arising.
- Information ownership and perceived control.
- Models and methods in privacy in public spaces.
- Abstraction of resource from underlying architecture.
- Standards supporting elements required in trustable privacies.
- Key themes identified in privacy in public spaces.
- Conclusions drawn from literature review.

2.1.1 Qualitative research and access to research data in the public domain

James and Sørensen (2000) argue the case for archiving social and behavioural data in a manner which allows other researchers relevant and appropriate reuse of archived data to answer further research questions and to provide historic access for future generations. A strong case is presented, with reference to the Murray Research Centre (James and Sørensen, 2000), for provision of reasonable protection of privacy of such data. Amongst the arguments presented is one supporting new prospective studies from old studies and archived data. It is argued that an important function of longitudinal studies involves the re-examination of attitudes and events, instead of

depending on retrospective commentary. It is further argued that longitudinal studies used for secondary analysis enables the use of data collected for one problem set to be used in the context of others, thus allowing the investigator to proceed without being bound by incomplete studies. Use in follow-up studies is also presented as an argument in favour of access to qualitative research data for reuse and reanalysis.

Corti, Day and Backhouse. (2000) provide similarly convincing arguments supporting archival of qualitative research data, with specific reference to the Qualidata initiative at Essex University.

Kuula (2000) comments on technical and information society issues and motivations for the Finnish Social Science Data Archive, started as recently as 1999. All of these initiatives are relatively young and not representative of the state of accessible qualitative research data. It is not in the scope of this dissertation to fully expand the arguments as to why qualitative researchers are notably reticent regarding publishing of such data.

2.1.2 Rights and responsibilities attached to qualitative and personal data

Being able to honour guarantees of anonymity at any stage of the research process is problematic for researchers. Initial publications arising out of analyses of qualitative data may be the novel publication of facts or stories concerning subjects under study.

This can be a severely challenging time for researchers — the rich nature of qualitative data lends itself to descriptions of the interviewees, their lives and their surroundings, and as such, the dilemma presented to the researcher is primarily that of how much detail to reveal.

2.1.3 Primary and secondary qualitative data

Qualitative data exists in two broad forms – primary and secondary. Primary is that gathered and assembled by the original researcher(s). Secondary is that research data gathered by other researchers subsequent to the original research and used for further work (not necessarily related to the original), and/or that gathered by other researcher(s) for use by the original researcher(s).

A summary of some of the forms taken by data potentially useful on a secondary research basis is presented below. For purposes of clarity, primary data is taken to be that which is assembled by the original researcher. Secondary data is that assembled by parties other than the original researcher for use in research.

A list of sources follows:

- Qualitative Sources
    - Biographies - subjective interpretation involved.
    - "Blogs" (online personal and publicly visible logs) and diaries.
    - Photographs, video and audio.
    - Recollections and communications (email, letters).
    - Mass print media (newspapers, magazines).
    - Period literature.
    - Official publications.

- Quantitative Sources
    - Demographics.
    - Administrative reports.
    - Surveys.
    - Commercial.
    - Market research.

- Archives and repositories
    - Finnish National Archives.
    - Institutional and national academic archives.
    - Qualidata at Essex University.
    - McMurray Repository (Canada).
    - Digital archives and closed-publication repositories.

- Reasons for use of secondary research data
    - Exploratory.
    - Supplementary.
    - Primary sources not available (e.g. 19th century Afrikaans authors).
    - Economic restrictions.

- Limitations
    - Divergent aims of primary and secondary use.
    - Divergence of intended meaning and definitions.
    - Longitudinal incompleteness and error.
    - Poor accommodation of bias .
    - Integrity – have data been cleaned or massaged unacceptably.
    - Validity of data.
    - Awareness that documents can affect perceptions as well as report on them.

2.1.4 Personal data records in (public) digital realms

Researchers and research communities of practice (CoP) readily acknowledge the critical role of ethical and sensitive attention to the rights of research subjects. Wenger and Snyder (2000) define a community of practice as

> *a group of people informally bound together by shared expertise and passion for a joint enterprise.*
>
> (Wenger and Snyder, 2000)

There are numerous examples of guidelines and CoP attempts at self-regulation with respect to ethical and sensitive handling of research data. References to some of these are made by Corti, Day and Backhouse. (2000) where the British Sociological Association and other bodies guideline and standards references are noted. Further examples are listed below in Table 1:

Table 1 - Examples of guidelines/codes of conduct

Organisation/research CoP	Reference to guidelines/codes of conduct
AIC	http://aic.stanford.edu/pubs/ethics.html
SARPN	http://www.sarpn.org.za/documents/d0000352/index.php
International Sociological Association	http://www.ucm.es/info/isa/about/isa_code_of_ethics.htm
Association of Social Anthropologists	http://www.asa.anthropology.ac.uk/ethics2.html

All of the Codes of Ethics (CoE) investigated merely proposed guidelines. Limited sanctions are available with some CoEs, but these are effective only within small and peer-monitored communities of research practise. No such limited sanctions afford the research subject(s) the luxury of granting additional rights, nor of rescinding any rights of access to their own data after the fact. This strongly indicates a need for a

mechanism which affords research subjects such facilities in a near-universal manner. Corti, Day and Backhouse. (2000) note that guidelines as published have many common attributes and features. These guidelines have evolved over many years and tend to reflect the prevailing societal norms relatively closely. Terminal responsibility for decisions relating to research projects is still the province of the primary researcher. This is arguably a significant factor in dissuading researchers from making much more than interpretive presentations of qualitative research available in any public sense.

2.1.5 Praxis with respect to secondary qualitative data analysis and use

Secondary analysis and reuse of quantitative data is *de facto* practice in the physical sciences and much of the social science arena. Subsequent accessing and reinterpretation of data is familiar and normal to most researchers (Corti, Day and Backhouse. 2000; Fielding, 2000; James and Sørensen, 2000; Kuula, 2000; Thompson, 2000).

When considering qualitative data, there is a considerable body of resistance, on the part of the original researchers (Brown, 2002; Fielding, 2000), to third-party secondary reuse of the archived qualitative research data. There are three primary reasons for secondary reuse of qualitative research data:

- Further or additional analysis.
- Cross-sectional or reduced data analysis.
- Analysis from alternative perspective(s).

Hammersley (1997) and Corti (2000) purpose reuse and reanalysis of qualitative research data as being significant, supporting argument in favour of the cumulative nature of qualitative research.

2.1.6 Computer Assisted Qualitative Data Analysis Software (CAQDAS)

Lewins and Silver (2004) categorise CAQDAS tools as being those which take a qualitative approach to qualitative data. This approach is considered to be one where interpretation of data is through identification of themes, contexts and communities of interest. Lewins and Silver propose a number of questions to be asked when selecting CAQDAS tools.

Following is a distillation of these questions, attempting to generalise these as areas of concern in the broader qualitative research context:

- Types of data and preferred ways of handling .
- Requirement for support for multiple methodologies .
- Requirement for thematic and quantitative access to data.
- Individual or collaborative approach to use of data.

It is the intention of this research to suggest that a model for trustable privacies has the potential to provide a means of extending the functional and application reach of any of the current crop of CAQDAS tools. CAQDAS tools include *QSR Nvivo* (QSR International, 2004a), *Atlas.ti*, *MAXqda*, *Qualrus* and *Kwalitan*. All of these tools have data import and export facilities; with XML (eXtensible Markup Language) being available on some. Simple support of XML and other generic data formats is not adequate as a means of extending reach in terms of a wider range of data resources accessible from a CAQDAS perspective. There is potential for future research in this direction.

2.2 Linking qualitative research data and questions of privacy and trust

This section of the literature review aims to demonstrate conceptual links between archival and accessing of qualitative research data and topics of privacy, confidentiality and, ultimately, trust.

2.2.1 Privacy, Confidentiality and Security

Privacy is taken to be the combination of the right of the individual to freedom from outside interference or intrusive monitoring and positive retained control of access to and use of personal data by that same individual. At no point does this definition of privacy infer access by any other party to this personal data. Access by any other party is by controlled and persistent consent. Privacy is moot without control remaining in the hands of the data owner, the individual.

Allowing limited and/or conditional access to personal data dilutes the strength of the privacy attribute. Simply formulating a confidentiality agreement negates much of the sanctity of privacy – by virtue of the fact that there are now at least two parties who have access to the personal data. At this point, the individual can but hope that confidential agreements are upheld.

Privacy is linked with personal data, whereas security has relevance across all and any data domains. It is a reasonable expectation that any data are secured. The corresponding expectation that any such data are treated as private. This expectation is not to be relied upon, as illustrated by the example of the Human Genome Project (Human Genome Information Project, 2003). The general breakdown of the human genetic code is not specific to an individual (although a researcher may wish to secure work-in-progress). The genetic footprint of an individual, and any information on genetic anomalies which that individual may have, are private data. These data, for purposes of arguments presented in this dissertation, fall in that domain defined by the individual's "*right to be let alone*" (Brandeis and Warren, 1890).

Kang (1998) avers that privacy tells us "*what to do*", and that security tells us "*how to do it*.". In the section following, data security is discussed with reference to retention of control and preservation of attributes and aspects of privacy.

2.2.2 Background to privacy in public spaces

Privacy eludes specific and unambiguous definition. Definitions range from the sociological, as postulated by Westin (1967), where privacy is argued to be the right of the individual to control and alter information about themselves and to determine the conditions under which any of this information is communicated to others, to Brandeis and Warren's (1890) legally-charged assertion that privacy was fundamentally the right of the individual to be left alone.

Margulis (1977) defines privacy as the partial or complete control of transactions between parties, where the ultimate objective is the strengthening of autonomy with reduction of the vulnerability of any party.

A new area of investigation is that of privacy in the domain of genetics being regarded as the right of the individual to restrict access to his or her own genetic information. Brief reference to this is found on the Human Genome Project Information (2003) website.

Privacy is a relatively recent presumed right and legal concept. Brandeis and Warren's (1890) definitions and scoping of privacy as a concept are perhaps the best-known and most commonly referenced benchmark for privacy. It is a concept which challenges immediate, unambiguous specification. Privacy exists within the confines of the mind, a home, a place of work or in the context of social community. It is primarily a resistance to perceived or real invasion of that space which we individually define as personally sacrosanct. The definition of public spaces has extended from the merely physical shared spaces to those spaces defined only by

shared interests. Users of public spaces range from individuals emailing, transacting and "blogging", through research subjects and researchers and on to diasporic communities and displaced individuals. Trust and privacy are inextricably interwoven - and concerns of confidentiality are shared by the broad spectrum of users. It is in this context that privacy and trust questions are considered in this dissertation.

2.2.3 Differentiating privacy and data security

An element of future perfection is evident in published works concerning user-centric privacy. Graham (1999) refers to user-centric privacy as a future technology. It is further stated state that a desirable state of affairs for online privacy is software negotiation based on user preferences and machine-readable statements of privacy policies. Suggestions that the Platform for Privacy Protection (P3P) alone might achieve this are flawed. Cranor (1997) established an early position of criticism of P3P, arguing that it was toothless guardian. This is addressed further in Chapters 5 and 6 in this dissertation.

Privacy and data security share a number of key attributes, amongst which are the following:

- Boundary definition, management and assertion.
- Selective access and disclosure mechanisms.
- Strong affinity between owners and perceived need for control.
- Dilution of perceived trust where control and ownership are separated.

2.3 Review of privacy

Privacy has emerged as a fundamental consideration with respect to access and the secondary use of qualitative data, not limited to data which originate in the research domain. As such, literature concerning privacy issues has grown to occupy a considerable portion of this chapter.

2.3.1 Privacy considered in broad terms

Rubenfeld (1989) proposes a theory of privacy as means of preventing institutions from becoming effectively totalitarian in nature. A limited definition of privacy by Rubenfeld is presented as the "*freedom not to have one's life too totally determined by a progressively more normalizing state.*". This convergent with Brandeis and Warren's (1890) reference to privacy in legal terms. The same perspective underpins the basis of this dissertation's anarchic philosophy - an anarchy in the pure sense, where control is the right and province of the data owner, and not of any other third party.

Public and private are colloquially understood to be separate and distinctly different in their characters. This is directly supported by Altman (1977), who observes that there is a fine and moving line separating privacy and publicity. Altman's theory of privacy may be restated as that of a conscious and deliberate act of controlling access to that which defines "self". This is central to the preservation of identifiable and believable privacy for the individual.

Privacy is therefore seen as a boundary or domain control problem. This view permits degrees of access and denial of access to be varied according to context and authority. While this is valid in a limited conceptual sense, the online world turns upside-down established views which seek to distinguish the public from the private.

The boundaries become indistinct and subject to capricious and arbitrary review by the data owner. This is the right of the individual as asserted here.

The argument is extended into the online domain by Lessig (1998) and Rotenberg (2001) through contextualised discourse on privacy. Both authors note the dynamic and complex nature of privacy and the technological and legislative drivers which help to shape privacy.

Privacy, while considered from the information security perspective, is fundamentally a question of the individual in community (Lessig 1998, Kubiatowicz 2003, Corti 2000) and as such, this literature highlights aspects of privacy considered from the reasonable person point of view.

2.3.2 A preferred definition of privacy

Selection of a preferred definition of privacy is required to establish the functional requirements of any proposed privacy and trust model. This selection is prefaced by considering literature on the concept of a right to privacy.

Common perceptions are that privacy, in all its forms, is a right granted at institutional level (by governments, constitutions) or a natural right which "simply is" - the Tao of Privacy. Clarke (1997) comments that privacy is better thought of as "*one kind of thing ... that people like to have lots of.*" This is perhaps far too loose a definition and will therefore be expanded upon.

The chosen method of describing privacy and trust is as follows:

- Describe basic principles of privacy. Survey views definitions of, privacy over the past century to provide a set of basic principles of privacy.
- Generalise principles sufficiently to allow a general statement of privacy. Discussion and survey of categories of privacy allowing distillation of fundamental aspects of privacy, which are assembled to form general statements of privacy. These form the conceptual basis of generic support of privacy, from the perspective of the individual.
- Establish a technical and procedural basis for assertion of privacy thus described. Taking the working definition of privacy established, technical concerns, requirements and attributes are discussed and the technical fundamentals required for creation of a privacy/trust model are laid down.
- Build a model supporting the concept of a trustable privacy. This is based on the working definition of privacy, as presented, and on the technical and procedural aspects addressed previously.

The steps outlined above have guided the development of both argument and model throughout this dissertation.

2.3.3 Privacy in public spaces

It is appropriate to ask whether or not the concept of privacy in public spaces constitutes a valid and persuasive privacy protection domain. This research began with the assertion that the perceptions of, and the facilities for the individual to voluntarily attribute trustability to a model constituted the essential basis of that required to create trustable privacies in public spaces. The nature of the research dictates a moderately restricted working definition of pubic spaces. *Public Spaces* are taken, for purpose of this research, to be public *digital* spaces. Parallels are drawn with physical public spaces in order to illustrate points.

The digital public spaces referred to include the Internet, organisational and institutional databases, personal data which is stored and/or extant in an online context, governmental data archives and any form of online accessible archived qualitative research data referring to individuals or communities.

Setting the focus, privacy as discussed and understood needs to be dissected. Privacy means many things to many people – the meaning given greatest consideration in this dissertation is that emerging from the needs and perceptions of the isolated individual. Public policy is constructed from a whole society standpoint.

A frequently quoted phrase in public policy is that of "*the greater good*". The Minnesota State legislature, along with the US Congress, is challenged in this regard by Brase (1999) in one example of hidden agenda policymaking.

In public debate and in the drafting of legislation, the greater good is a frequently used motivator for setting out overriding interests which are deemed to be appropriately used to advance the interests of society as a whole.

Moving on from loose assertions of "greater good" as diluter of essential privacy, Nissenbaum (2000) identifies three factors taken to effect dismissal of the perceived force of privacy in public, detailed below in Table 2:

Table 2 - Characterisation of privacy factors

Privacy factor	Characterisation of privacy factor
Conceptual	May be considered with reference to the terms "public" and "private" as used in political and legal context. Generally used to indicate clear delineation of that which is in the province of the individual or the family and that which is in the province of government or greater society.
Normative	Protection of privacy in public are targeted by objections based largely upon the assertion of overriding, competing interests challenging the appropriateness of maintaining agreeable levels of privacy. An example might be government collection of personal data for purposes of establishing spending habits (taxation) or communications habits (political affiliations) or any other purpose which requires assertion of alleged overriding interests.
Empirical	We are seen daily by many and often noticed by no-one. Most make the reasonable assumption that they are not noticed, or that an observer can retain only so many bits of information (usually sparse and fragmented). We do not worry about invasion of privacy during the course of our usual public passage each day.

Considerations of privacy and trust in this dissertation are primarily from the naïve position of the individual. This is seen from the perspective of Nissenbaum's (2000) *empirical* privacy and is moderated by the appreciation that arguments in favour of assertion of *normative* privacy are considerable.

A few questions arising around privacy conflicts are listed:

- Is it reasonable to pry into the financial affairs of an alleged fraudster to establish guilt or innocence?
- Are the contents of a paedophile's diary or hard drive considered available for scrutiny?
- Are the bounds of fairness and decency overstepped when searching a suspected shoplifter?

- Does the state have a right to an overall view of the collective scattering of data which define an individual's digital existence?

In all of the examples listed, the immediate tendency is to convict the individual before establishing factual links. An unproven suspicion does not provide grounds for negation, removal or abuse of assigned, attributed or natural rights. A datum or even a piece of structured information appears in a public digital space does not carry with it a transferable right of access by anyone other than the owner.

Article 19 of the United Nations Universal Declaration of Human Rights (United Nations, 1948) asserts that:

> *Everyone has the right to freedom of opinion and expression; this right includes freedom to hold opinions without interference and to seek, receive and impart information and ideas through any media and regardless of frontiers.*
>
> (United Nations, 1948)

It is inferred from this that the owner of the data has *a priori* rights to grant or rescind rights of access. Such rights reasonably lie with the owner. There are many more strong circumstantial arguments in favour of dismissing deeper concerns around privacy when viewed in parallel with a normative privacy mandate.

A fundamental concern which arises with the pursuit of normative privacy is that of gradual and unnoticed erosion of the ability of individuals to control access to their private domains. Lessig (1998) refers to this gradual loss when he notes that limits on searching are eroded further and further as a consequence of growing options for searching data without imposition of burden upon the targeted individual. Familiarity with ubiquitous technology breeds a careless contempt for the dangers of the personal data trail. Email, for example, is so much a part of life for many that the latent risks are very seldom considered.

Kang (1998) also refers to legislative history of the U.S. Electronic Communications Privacy Act (ECPA), which comments that

> *Privacy cannot be left to depend solely on physical protection, or it will gradually erode as technology advances. Congress must act to protect the privacy of our citizens. If we do not, we will promote the gradual erosion of this precious right*
>
> (Kang, 1998)

Taken in isolation, this is a commendable position for those custodians of the public good, the state, to take. Stepping back and considering the double-edged sword of state protection, that which is defined as a delineator of privacy is not infrequently turned to instrument of normative privacy. This brings with it the application of concerns which are used to override concerns of privacy of the individual.

There is much material presented by the Electronic Privacy Information Center (EPIC, undated) on their website[ii] highlighting many legislative attempts to assert normative privacy measures in the name of public safety and the "greater good".

Individual defence of own privacy is characterised by those aspects of life which, by and for that individual, are taken as neither subject to nor open to public scrutiny and criticism. Seen from the personal perspective, such private aspects of life are not shored up by any guarantee of protection. Personally-created records of anything deemed private, data captured and stored by researchers or any other agencies are all subject to abuse of privacy concerns where normative privacy is applied in isolation. Hoffman *et al.* (1997) provide a long list of privacy abuses prevalent in commercial Web interactions, but remain unable to suggest effective remedies for these abuses.

[ii] http://www.epic.org – the Electronic Privacy Information Center

2.3.4 Respecting privacy

There are as many variations in definitions of privacy as there are opinions and people. We live in an increasingly connected world where the issues are less about absolute and immediate intrusion in the physical world and far more about the trail of persistent data which left behind after every online transaction. Transactions include ordering goods online, online banking and the ubiquitous activity of emailing. Ubiquitous connectivity is both enabling and restrictive.

The broad concept of privacy has moved on from the time when privacy was what existed in our own heads (Lessig, 1998) and in our parents' bedrooms to one where our life data trails require protection to avoid abuse by the unscrupulous and the underhanded. Froomkin (2000) avers that individuals in society are inclined towards a kind of "privacy myopia" where they undervalue information about themselves to the extent that protecting such information is not perceived to be worth the effort of doing so. Froomkin's position has relevance to an initiative aimed at returning control over personal data to affected individuals is discussed in a *wired.com* article by Scheeres (2001). This initiative is described as a reaction to telemarketer abuse of personal information and quotes the originator, Tracy Coyle as saying that "...people can control it better themselves than by giving it to every Tom, Dick and Harry website that's out there.". Froomkin's comments concern the passive converse of the individual wishing to actively control his or her own data. This privacy myopia is present usually only where individuals have not had reason to doubt the integrity, privacy and trustworthiness of their interactions in the online world.

Every credit card transaction, every enquiry and reservation leaves a persistent trail. This is not of itself a negative consequence. The problems to appear when the individual is not accorded the right "*to be let alone*" (Brandeis and Warren, 1890).

Respect for privacy diminishes, and when governments and organisations fail to distinguish the possibility of criminal intent or activity from the reasonable human

right of individuals to create and maintain digital personae in public spaces, then achieving Brandeis and Warren's state of being 'let alone' becomes unattainable. The individual ceases to have or to exhibit privacy in any sense.

A complementary view is provided by Gavison (1980) who suggests that

> *... there are three elements in privacy: secrecy, anonymity and solitude. It is a state which can be lost, whether through the choice of the person in that state or through the action of another person.*
>
> (Gavison, 1980:421)

The essence of the concept of privacy in public spaces is to be afforded the opportunity to live discreetly in public and to store personal private spaces in the same public domain.

Trust is not automatic, nor may it be purchased – how do we trust third parties with our personal and confidential data? Are we able to determine the degree and type of access to our data? If there is no privacy, then trust is impossible.

> *We must attempt, at all times, to guarantee promises of confidentiality made to research participants, where possible. ... Whilst not all data subjects may be concerned about their anonymity, others are. For those subjects who wish to remain anonymous, for archiving we must seek to anonymise identifying information about them.*
>
> (Corti, Day, Backhouse, Dec 2000)

There is some room for the argument that a job worth doing is worth doing yourself. Bromseth (2002) questions who is responsible for protecting a user's privacy. This is noted in Chapter 1 as being the key question and motivator for this research. This dissertation suggests that the user is the appropriate candidate, with the proviso that adequate tools and mechanisms for the purpose exist and are available for use.

2.3.5 Privacy in the digital and online worlds

There is more than a passing relationship between the world experience in daily life, the physical world, and that experience online. The tendency is to unconsciously and seamlessly link experiences and perceptions of the online and physical worlds in which we live. Rheingold (1993) comments extensively on the evolution of online communities and on the blurring of perceived boundaries between the physical (or real) world and the online (or virtual) world. Lessig (1998) comments that the real and virtual worlds act as reciprocal modifiers.

2.3.6 Issues relating to privacy in the digital world

A number of aspects of data handling bear expansion, linked to issues of privacy in the digital (public) world:

- Security. The blocking of passive and active attempts to gain access to a datum or data available only with the explicit consent of the owner.
- Confidentiality. Wrapping up of a datum or data so that actual content is only available to an authorised and authenticated party. References to accreditation mechanisms are made in Chapter 5.
- Accessibility (or availability) is essentially self-explanatory. Confidence that data will appear, when sought, is fundamental to a privacy model being able to be trusted by users. There is an expectation that items deposited in a safety deposit box will be available on demand. When the user experience is counter to this, trust in the model (the bank safety deposit system) is eroded or destroyed.
- Integrity. Preservation of data as originally laid down by the owner, is the third trust-enabling aspect required of a feasible privacy/trust model.
- Anonymity, considered in two parts - complete anonymity and pseudonymity. Complete anonymity is the absence of identifying attributes linking data, or communications, to a specific entity (individual or system).

Chaum's (1981) MIX model was created to attempt to allow dynamic concealment of actual user identities in emails, chat rooms and discussion fora. Although Chaum was a relatively early thinker in terms of anonymity in the digital realm, his MIX model has formed the basis of many of the attempts to provide trustable privacy models for the WWW. Acquisti (2002) detailed a user-centric MIX-like protocol which attempted to return control over data to the owners, but still relied on trusted server entities. Like the overwhelming majority, MIX relies on third party processing for its ultimate realisation.

All models and implementations presented have relied upon intermediaries or third parties of one form or another. It is the position of this dissertation that all of these models are fundamentally and fatally flawed in this regard due to their demand for arbitrary placement of trust in an unknown third party.

Infomediaries was a term coined in the early part of the 1990s to describe, in mildly obtuse terms, the concept of using a trusted third party to retain enough confidential and personal information to successfully validate and authenticate transactions of any type between two parties. Maguire (1998) provides a popular overview of the communications between the Federal Trade Commission (FTC) and Microsoft over the challenges to its introduction of the allegedly privacy-enhancing Passport (.Net, 2003) scheme.

Microsoft's *Passport* uses a single email and password doublet to authenticate access to a variety of services. Included are identity and credential propagation services, ostensibly to ease the user's secure passage through the WWW. *Passport* is prototypical of the commercial drivers on the WWW which work against the protection of trustable privacies.

> *Inevitably, Passport is also controversial, because its purpose is to provide a workable balance among competing priorities: information versus anonymity; security versus usability; and the needs of companies versus those of the*

consumers they serve. The ongoing search for the right balance is taking place in a dynamic environment. Controversy, criticism, and debate are a natural part of that environment. ... developers listen closely ... feedback gets incorporated via ongoing improvements ... which results in a better online experience for users ...

(Microsoft, 2003)

The myth of the trusted third party is used as a foundation stone with such schemes. Point-to-point encrypted communication is frequently presented as a secure and trustable mechanism.

While it is frequently true that risk is low and security (data) is acceptable, the point-to-point encryption (SSL, SSH) requires agreement of sorts between the two parties communicating.

Point-to-point provides no guarantee of confidentiality on the serving node. *Freenet* (Clarke, Sandberg , Wiley and Hong, 1999) is philosophically similar to Chaum's MIX model, but succeeds in hiding the parties involved so well that the data owners lose control of the data published.

The network itself becomes the arbiter of what metaphorically lives and dies. This is in diametric conflict with the aims of creating model supporting trustable privacies in public spaces.

2.3.7 Erosion of privacy on the Internet

There are many risks encountered with use of the Internet. Privacy, so earnestly sought, is eroded considerably when the Internet is used. Perceptions of anonymity and privacy are founded on ignorance of the hidden mechanics of Internet transactions.

Early users of the Internet were technically expert users; the majority of current Internet users are probably better described as "appliance users". In other words, the expectation is that things happen. How and why things happen is of little direct interest unless we experience personal loss or invasion of our private spaces.

Opening the discussion on erosion of privacy on the Internet are the following items:

- Organisational monitoring of personal activities and data.
- Inappropriate trust building.
- Profile building by third parties.
- Identity fraud and theft.
- Transaction interception.
- Location tracing.

Reversing the effects of this erosion is a complex undertaking. Some elements of the environment which may be successfully addressed are listed below in Table 3:

Table 3 - Reversing the effects of erosion of privacy in online public spaces

Environmental element	**Attributes and characteristics**
Action	Active avoidance and blocking of negatively-disposed individuals and organisations.
Self-regulation	Organisations and governments very seldom willingly self-regulate. By implication, regulation is an imposed regime requiring voluntary surrender of absolute self-direction in favour of ethically appropriate conformance. Arguments for and against are discussed later in this dissertation in Chapters 6 and 7.
Infrastructural and mechanistic support	Where inherent trustability is lacking, support for protection of privacy and return of control of data to the individual is an absolute prerequisite.

The World Wide Web (WWW) has its origins in a low-bandwidth, multipathed, non-

centralised architecture intended to allow reliable, resilient sharing of distributed information. Growth in richness of content and bandwidth demand, together with the admixture of different technologies has placed demands on security not foreseen at the outset.

A significant problem in the online world is the common perception that "you can't see me, therefore I am safe". People who would blanch at the idea of scribbling their ATM card PIN number next to the ATM are often the very same ones who will happily do online banking with no antivirus software active or will wander off to make tea at the office whilst in the middle of performing an online transaction. One such example is the ABSA Bank Internet fraud scare of 2003 (Granova, 2004).

Security is perhaps the most important issue in the online world, although it has not always been considered during the development of Internet technologies. A vast amount of confidential information passes across the Internet's highways and unauthorised access to this information is a real danger.

When an individual has a need to store a digital record, an online archive of life (or any part of it), great demands are placed on the chosen architecture in terms of resilience, preservation of confidentiality, access control and ensuring unfettered and unrestricted control of *own* data from anywhere. In other words, a trustable architecture which is not limited to a single mode of access.

Access via the WWW, cell phones, Plain Old Telephone Systems (POTS), Interactive Voice Response (IVR) systems and others is required to make such a trustable architecture accessible beyond the immediate reach of those on the formally connected side of the digital divide.

2.4 Review of qualitative data usage and issues arising

Praxis with respect to reuse of qualitative research data has been, and remains, largely bound to the dissertation that going back to review 'own' qualitative data is valid, whereas reviewing third-party qualitative data is invalidated through lack of intimate familiarity with the context and subjective associations of the original researcher (Corti, 2000; Fielding, 2000; James and Sørensen, 2000; Kuula, 2000).

The model for creation of trustable privacies in public spaces aims to provide alternatives to the persistent praxis of exclusion of secondary reuse of qualitative data by establishing routes for flexibility of access and authority.

Qualitative data, whether raw or research output, whether closed or open in intended use, requires mechanisms and models to support access and use via the only universal and near-ubiquitous resource, the WWW.

2.4.1 Access, confidentiality, ownership, proxy and consent

Questions surrounding confidentiality, ownership, proxy and consent arise from the following:

- Obtaining consent for retention and archival of data.
- Anonymisation of data.
- Restricted access (controlled by the researcher).
- Restricted access (controlled by the research subject).
- Loss of contact with research subjects.

Clarke, Sandberg, Wiley and Hong (1999), Corti , Day and Backhouse (2000), Bromseth (2002), Brunk (2002), Nissenbaum (2000), Thompson (2000) and Hammersley (1997) all note concerns over confidentiality of access, ethical use of publicly accessible data and insidious erosion of individual privacy. In every case,

the primary problem of realising a universally useable model lies in maintaining control at the level of data owner, and not simply at the level of the researcher or data custodian.

2.4.2 Perceived trust, separation of control and ownership

Suryanarayana and Taylor (2004) classified trust management systems into three categories:

- Credential.
- Reputation.
- Social network.

(Suryanarayana and Taylor, 2004)

Published literature indicates required membership of at least one of these categories for any model proposing support of trust, control and ownership in the hands of the data owners. The conceptual basis which is to be developed in this dissertation is required to be able to lay claim to membership of all three of the categories, with the crucial distinction that it is able to support all three without recourse to any trusted third party requirement.

2.4.3 Trust as privacy-dependent attribute of archival models

The weight of literature indicates a basic requirement for credible privacy as a part of any model or mechanism proposed as a trustable option. Without an individual having the facility to grant or rescind access to data specific to himself or herself, privacy does not exist. Without privacy, trust has little opportunity to be established.

2.4.4.1 Trustable architectures

To establish and maintain a trustable privacy in a public space requires that private data may be left under digital lock and key so as to ensure that the owner or custodian

may be confident that the data will be where they were left and will not be accessed or used by any unauthorised parties. There are manifold contemporary architectures purporting to do just this.

Many of these are server-based, many others depend on so-called 'trusted third parties' to hold keys in escrow, others are platform-bound, and the most promising alternatives have characteristics which render them less than suitable for the purpose at hand – creation and maintenance of trustable privacies in public spaces.

There are two broad classes of trust which may be imposed on the digital world, for the purposes of this dissertation being the following:

- Third party arbitrated / adjudicated protocols, where the trusted third party either passively or actively ensures that both parties (these being the data owner/custodian and the data requester) act fairly and ethically.
- Self-directed protocols, where breaches of privacy and trust are either negated or detected and addressed through protective mechanisms.

The concept of self-directed trust protocols is central to the development of the trustable privacy model presented in this dissertation.

2.4.4.2 Peer to peer networks

These are the peer-to-peer (P2P) architectures of which the likes of *Napster* (undated) and *Gnutella* (2001) are perhaps best known. *Napster* originated as a P2P indexed music sharing service which turned each member's computer into an indexed and searchable music server for any other logged in *Napster* user. Gnutella addressed several architectural shortcomings experience by *Napster*, but remained an indexed sharing service (Eytan and Huberman, 2000), with little provision for protection of privacy. These architectures are the antithesis of privacy-enabling tools. Initiatives such as *Edutella* (Nejdl, Wolf, Staab and Tane, 2002) are significant attempts to adapt

the basic P2P concept to better suit the needs of shared online resources, though not the requirements of a diasporic privacy in public spaces.

2.4.4.3 Boundary definition, management and assertion

Boundaries are dynamic and seldom easily predictable. The significance of the online environment lies in its ability to dynamically blur and redefine boundary states. Controlling or mediating access and process where redefinition occurs is fraught with ethical and social issues. Convention holds that restriction or nondisclosure are means of limiting access to that data; of controlling at source. Palen and Dourish (2003) comment that:

> ... *one of the roles of disclosure can ironically be to limit, rather than increase, accessibility. Views of privacy that equate disclosure with accessibility fail to appreciate this necessary balance between privacy and publicity.*
>
> (Palen and Dourish, 2003:131)

The boundaries defined by conflicting requirements of privacy and accessibility overlap and create tensions in the models purporting to support creation of trustable privacies. Technology is both part of the problem and a fundamental key to the solution.

2.4.4.4 Selective access and disclosure mechanisms

The concept of selective disclosure and access is key to all privacy enhancing technology (PET) models and mechanisms. Seamons *et al.* (2002) identify two classes of credentials used in selective access and disclosure mechanisms:

- Possession-sensitive credentials.
- Attribute-sensitive credentials.

Means of avoiding information leakage and loss are discussed and strategies proposed to secure against inadvertent exposure of credential material. Seamons *et al.* (2002) continue with a consideration of measures which may be used to enhance this *ad hoc* approach used by trusted third parties such as eTrust[iii].

2.5 Information ownership and perceived control

Mutka (2003) lists a set of goals intended to promote privacy and security in a connected environment. The same goals are crucial for supporting building of perceived control on the part of the data owners. These are listed as:

- Users expose personal information prudently.
- Services only respond if user has proper credentials.
- User only supplies credentials if service is trusted.
- Without proper credentials for requested domains, devices stay silent.
- Automated supply of credentials.
- Security is automated.
- Sharing devices and revoking privileges of shared devices are easy.

The same goals may be recast in terms more closely approximating the stated goals

[iii] eTrust is a commercially-motivated peer-certification organisation purporting to provide assertive trusted 3rd party facilities for ecommerce purposes.

of this research with respect to facilitation of user-defined levels and types of privacy and trust:

- Data owners expose only that metadata necessary for resource discovery.
- Data is accessible only with appropriate credentials ("accreditation").
- Data owner grants or rescinds accreditation at own discretion.
- Security is automated on a transactional and processing basis.
- Granting and rescinding of accreditation may occur on a push (sending grant/rescind permissions to users requesting access) and a pull (republishing of data with changed accreditation requirements).

2.5.1 Other domains affected by ownership and control questions

Domains affected by ownership and control extend far beyond the initial domain considered for this research. Additional domains include children in care systems, diasporic communities, broad communities of practice and private-public interactions including, but not limited to, interactions between individuals and the state.

2.5.2 Diasporic and Information Age societies

Societies become, or are, dispersed for any number of reasons. These include political and economic reasons, and reasons of natural disaster. Possibly the best known instance of a pre-Information Age diasporic movement is the Diaspora (the global Jewish community).

Movements inevitably develop which seek to create a sense of community amongst the dispersed, the diaspora. The immediate and compelling appeal of the Internet is the opportunities afforded for creating virtual communities. These communities are able to achieve levels of cohesion generally not possible in a pretechnological age.

The facility for a community to assert itself is crucial for its survival and growth - the

Information Age brings with it this ability through persistence of information, universal accessibility and the creation of community voice for the diasporic. The Internet is perceived to be an open, yet anonymous place. This perception works both for and against the creation of trustable privacies.

A distinction between perceived personal and social anonymity, and one-way anonymity is essential, with one-way anonymity being characterised by its non-reversible nature. Perceived personal and social anonymity is merely a masking of identifying characteristics, not the obliteration of these characteristics.

2.5.3 Codifying knowledge and encapsulating societal memory

Large bodies of research exist covering Knowledge Management Systems (KMS) (Abecker, 2001; Ackerman, 2000; Tomek, 2001), access and archival models and qualitative research data encoding and usage. Eberhart (2004) and Markus (2001) comment that successful applications of KMSs tends to occur in vertical commercial and domain-limited academic arenas.

Klamma and Schlaphof (2000) make specific reference to directed and domain specific knowledge management in the commercial domain. Users in these cases are either domain experts or have domain experts as 'data intermediaries'. Klamma and Schlaphof go on to acknowledge the roles of domain experts in interfacing with the repositories. It is clear from their work that knowledge is generally perceived to be readily codifiable and is simply a tool. This perception of knowledge as transferable commodity and tool is at odds with the expressed concerns of qualitative data archivists and researchers (Bromseth, 2002; Brunk, 2002; Clarke, Sandberg, Wiley and Hong, (1999); Corti *et al.*, 2000; Fielding, 2000; Hammersley, 1997).

This philosophical standpoint is not readily transferable to the domains touched on by academic research and the needs of diasporic communities and dislocated individuals.

A great shift in priorities and problems is found with research data users and reusers who are not domain experts.

Markus (2001) comments that knowledge may be categorised as explicit (knowledge which has been recorded, articulated and codified) and as tacit (that which has been internally constructed and exists only within the minds of people). Markus further notes that it is only explicit knowledge that falls within the scope of information technology. It is this explicit knowledge which is referred to in this dissertation. Schirmer (2003) notes that

> *... knowledge management is not itself a technology, knowledge management technology solutions ... have been developed to realize that aim [of commoditising knowledge].*
>
> (Schirmer, 2003:519)

There appears to be almost universal abuse of the meaning of the term 'knowledge'. Little in the range of flexible and adaptable models have appeared in response to the question as to whether or not it is possible to represent personal and confidential knowledge fragments in public spaces. Application of expert domain knowledge and preservation of contextual prompts is the norm where KMSs are considered.

Where questions of expert and context-relevant access arise, KMSs prove inadequate for cross-domain usage (Markus, *op.cit.*; Klamma and Schlaphof, *op.cit.*). A lack of robustness and flexibility in the mechanisms often used is a parallel problem. Usual characteristics of KMSs are that domain experts are frequently required at the storage and extraction of 'knowledge' fragments. It is on this basis that KMSs are disqualified as appropriate tools for general application in the scope of this project. It is the objective of this research to add to the overall pool of knowledge with reference to establishment of privacies in public spaces and the needs of qualitative data reuse and interdisciplinary collaboration.

2.6 Models and methods in privacy in public spaces

Listed in this section are brief outlines of models addressing aspects of privacy and data security. Some of the models are outlined following:

- Augmented Social Network (Jordan, 2003). The Augmented Social Network (ASN) proposes four elements assembled into a social network:

    - Persistent identity - enabling individuals to establish and maintain a digital identity, transferable online. This identity extends beyond a simple digital profile to reflect a closer image of the individual in a holistic sense. Control over this identity is asserted by the individual; although access to the identity is not thus asserted.
    - Interoperability amongst online communities - controlled movement from community to community, requiring protocol development and acceptance. Resource discovery is required and demands universal acceptance to ensure viable community roaming options.
    - Brokered relationships - the trusted third party to enable and facilitate discovery, introduction and relationship management amongst individuals and communities. ASN aims to collapse the six degrees of freedom to no more than three degrees of freedom.
    - Public interest matching - Ontology and semantic web-related linking of individuals and communities via interests and assertions.

- MIX (Chaum, 1981). MIX functions by:
    - Data from multiple parties wishing to exchange anonymously are processed by the MIX, by hiding links and connections.
    - MIXes obscure by reordering, padding, splitting and otherwise manipulating traffic.
    - Viable MIX implementations require chains or meshes of MIX servers to achieve convincing, and hence trustable, levels of anonymity.
    - MIX does not decouple data from original locations, only obscures the links amongst inbound and outbound.
        - Developments of the MIX model include:
            - *Onion routing* (Reed, Syverson and Goldschlag., 1998) relies on anonymous and obscure routing of traffic through web MIXes and selected routing agents. It does not address the preservation of data, merely its anonymous path across the web.
            - *Crowds* (Reiter and Rubin, 1998) - Users join a metaphorical "crowd" of users for *en masse* protection by means of:
                - Web requests unable to be linked to individuals.
                - Protection from end servers, other crowd members, system administrators and unauthorised intruders.
                - Hiding traces on the WWW without depending on the services of a third party/central authority.
                - Premise of anonymity in a crowd being more readily achieved than anonymity in isolation.
                - Main distinction between *Crowds* and the underlying MIX network architecture is in path selections.

- *Eternity* (Anderson, 1996). Notable as a precursor to *Publius*. The Internet was intended to be a resilient, redundant communication environment. Resistance to denial of service and disruption of channels is a key feature of the web. *Eternity* proposes using the Internet to build a storage medium with similar attributes - inheriting these from the underlying architecture. Additionally, use of redundancy and scattering to replicate data across a very large number of points and anonymity mechanisms to make denial of service attacks prohibitively expensive (in resource terms) characterises *Eternity*. *Eternity* is a chargeable model relying on trusted third parties and therefore counter to part of the ethos of this research, which is free access to technology and ideas. Related to this is OceanStore, described by Kubiatowicz *et al.* (2000), which proposes a global architecture for persistent storage of data.
- *Publius* (Waldman, Rubin and Cranor, 2000), attempts to make censorship or modification of content by parties other than the authors/publishers difficult. It is noteworthy that work on *Publius* appeared to be abandoned after it became evident that there was little scope for revenue on a commercial basis. Additionally, the identity of the publisher is concealed and protected after data is published. Nine goals which shaped *Publius* are listed:

    - Censorship resistance.
    - Tamper resistance and evidence.
    - Source anonymous - no way to tell who published data.
    - Updateable - allow publishers to make changes to own material.
    - Deniability – carriers legitimately able to deny knowledge of data.
    - Fault tolerance.
    - Persistent - publish without fear of expiry dates or obsolescence.
    - Extensible - support addition of features and users.
    - Free - all software free and freely available.

- *Kepler* (Nelson, Liu, Maly and Zubair, 2004) is a model attempting to bridge the chasm between institutional digital repositories and researchers wanting to publish research but retain control and the advantages of OAI digital repositories. Open Archive Initiative (OAI) is an initiative to develop and promote interoperability standards that aim to facilitate the efficient dissemination of content. *Kepler* uses archivelets (self-contained, self-installing software which acts as an OAI data source) to enable this level of control and integration to be achieved.
- *Hyperion* (Kementsietsidis, Arenas and Miller, 2003) is an open P2P network where peers may exchange data and/or services. *Hyperion* aims to investigate precise definitions of P2P data management architecture, the study of data integration/exchange/mapping mechanisms in P2P networks, and the development of algorithms for the efficient search, retrieval and exchange of data among the peers.

Anonymisation and obscurity through rewebbers (services which successively obscure both the client and the server from each other via a succession of web server relays) such as the TAZ rewebber (Goldberg and Wagner, 1998) is yet another example of a workable idea premised on trusted third parties.

This is not an exhaustive list but does describe the major range of options available on an open technology basis (not necessarily on a free-to-use basis).

2.6.1 Knowledge Management as model for qualitative data access and control

The intention to preserve meaning and context has the initial consequence of proposing knowledge management (KM) as a potential approach. Literature surveyed has highlighted limited areas of application for KM in preservation of meaning and context. Equally, surveyed literature has highlighted the inability of KM to

adequately address issues around perceived privacy and retention of control in the hands of the data owners. This directly undermines the establishment of privacy.

2.6.2 Knowledge and privacy

Before elaborating on knowledge management systems and privacy, a brief expansion on knowledge is required. Aside from the debate on whether or not knowledge is, indeed, able to be stored outside of the brain, a brief excursion on alternative flavours of knowledge is required. The Greeks recognised four essential varieties of knowledge, summarised here in Table 4:

Table 4 - Four varieties of knowledge

EPISTEME	Abstraction and the fundamentals of scientific statement
TECHNE	Codified practices, detailed breakdown of how things are done
PHRONESIS	Wisdom of experience, a socially-based knowledge
METIS	That which characterises the streetwise, the cunning. Basis of claims of intuitive "knowing"

Establishing a philosophical understanding of knowledge allows subsequent consideration of social, institutional and technological aspects thereof. These are summarised as follows in Table 5:

Table 5 - Other dimensions of knowledge management and facilitation

SOCIAL	Creation of networks of knowledge-enabling communities, shared community, all based on building of layered trust, supporting individual endeavour alongside community.
INSTITUTIONAL	Collective archival, storage, backup, discovery and reuse. Seeking gain for the community good.
TECHNOLOGICAL	Archival tools, networking and databases, resilient systems.

Much of what is offered in the Knowledge Management System (KMS) tool domain

is based on institutionally-biased architectures, usually intended to facilitate some form of organisational memory / expert pooling mechanism or philosophy. No user-centric tools seeking to empower the individual in an independent manner were found during this review. This dissertation started with a user-centric perspective, and it is with this in mind that KMS approaches and tools are examined. Perceptions of knowledge inform design of KMSs and therefore affect the ability of KMSs to function as privacy protective environments.

2.6.3 Tools and methods in KMSs

Tsui (2000) expands the basis of evaluation with a taxonomy of KMS tools and methods over the years. Amongst others, the elements listed in Table 6 are extracted from Tsui:

Table 6 - Aspects of institutional knowledge management tools and philosophies

Aspects of institutional knowledge management tools and philosophies.	Organisational knowledge maps
	Taxonomies
	Search tools
	Collaboration environments
	Repositories
	Resource discovery tools
	Information portals

Tools which work and are technically resilient are invariably dependent on central deployment and control and as such, not feasible for use by individuals or by ad hoc associations of interested parties (secondary academic reusers, public enquiry).

2.6.4 Limitations of knowledge management systems

It is a primary goal of the model to be proposed in this dissertation to empower individual users and to preserve a state of constructive anarchy. Institutional KMSs do precisely the opposite of this. Individuals are subsumed into the organisational whole, losing control over their own contributions in the immediate and the future. The Semantic Web (Miller, 2001) offers a future technology platform for a non-institutional alternative to centralised control KMSs. This alternative is likely to prove crucial in the qualitative research and diasporic individual domains, including facilitation of persistent and resilient recording of life histories of asylum seekers and displaced children (Jenkins, 2004). Skuse (2000) expands UK government policy position with respect to provision of services and architecture supporting the displaced and poverty-afflicted on a global scale, encouraging initiatives to develop non-proprietary architectures for preservation of digital records.

It is hoped that the model to be proposed will find a logical and comfortable home in the future of the Semantic Web.

Bonifacio (2002) comments that the popular approach to KMSs represent an epistemologically objective view of knowledge, where meanings are not under debate and that subjective aspects may be safely disregarded in favour of objective definition and encoding. This is at odds with many of the knowledge theorists (Jonassen, 1993; Thomas, Kellogg and Erickson., 2001; Klamma and Schlaphof, 2000) who place significance on differing perspectives and syntheses from individuals and societal interactions.

KMSs treat knowledge as a unitary and transferable commodity, albeit a complex one. It is proposed that knowledge is a dynamic collection of overlapping subdomains, navigated by users. This dissertation subscribes to this more dynamic and less prescriptive view of knowledge.

2.7 Abstraction of resource from underlying architecture

All of the privacy and trust models considered are closely aligned with specific communications and network architectures. The Semantic Web offers the best future compromise which is less tightly bound to underlying architecture. The Semantic Web provides a framework allowing data to be shared across domain and community boundaries. This dissertation aims to propose a model which may be regarded as edge-member of the collection of technologies related to the Semantic Web and its philosophies.

O-Telos, as described by Nejdl, Dhraief and Wolpers (2001), uses RDF Schema (RDF-S) as described by Brickley and Guha (2000) to provide a basis for extended meta-modelling and creation of metastructures. RDF supports autonomous data graphs in its native and simple form, although with limitations on inherent security and navigability. Schema are created, adapted and used according to specific views and needs of users. Individual requirements affect both the use of and the linking of additional information to affected schemata.

There is a flavour of constructivism in that information placed may be dynamically restructured on review by user(s) who may not necessarily be the original depositors.

2.8 Standards supporting elements required in trustable privacies

Research into and development of a model supporting the creation of trustable privacies is premised on the use of published standards where possible. This section aims to set context and reference points with respect to these standards.

The objective is to develop a model on the back of what exists, is understood and defined by open access standards. Commercial interests have identified standards including XML, UDDI, WSDL and others as crucial components of the information flows required for interoperability. The Open Group comments that this information

flow requirement will not be satisfied by technology alone, "but by many technical and best practice standards". Blevins (2004) proposes the following working definition of interoperability, in a presentation entitled "Boundaryless Information Flow":

> *The ability of two or more entities or components to exchange information and to use the information that has been exchanged "to meet a defined mission or objective"*
>
> (Blevins, 2004)

2.8.1 Published standards

Standards covered are the Platform for Privacy Protection (P3P), the Resource Descriptor Framework (RDF) and eXtensible Markup Language (XML). Applicable and related standards (XHTML, amongst others) are omitted both here and in the model description in Chapter 5. Inclusion would offer no substantive support to the arguments and model presented.

2.8.1.1 Platform for Privacy Protection (P3P)

The Platform for Privacy Preferences (P3P), developed by the World Wide Web Consortium (W3C), has become accepted as a standard for providing a technically simple means for users to theoretically gain greater control over use of personal information harvested by websites. Specifications, protocols and intended applications are described on the P3P website (P3P Project, undated). Details on how a site proposes handling personal data is encapsulated in P3P policy statements which are machine-readable format by the user's computer. P3P enabled applications examine the policy snapshot and compare this with the user's stated preferences. P3P attempts to enhance user control There are fundamental issues with the premises upon which P3P is based, not least of which is the requirement for trust.

This dissertation proposes subverting the original and published use of P3P and to apply it as an accreditation and permission request component as opposed to the policy statement tool role in which it is cast by design and default. P3P defines policies defined in XML namespaces (W3C) encoding the P3P vocabulary to describe the entities and practices with respect to stated privacy policies. Data types are enumerated and data usage policies described. SCRIPSIT subverts and reuses the functions and purposes of P3P to define explicit access and rights actions. P3P convention asserts that positive statement of data usage intent (with respect to data collection) is *de facto* normative and only partly negotiable, whereas SCRIPSIT's application of P3P assigns positive and externally immutable properties to content wrapped by P3P-derived trust assertions or accreditation fragments. *De facto* application of P3P requires that policies not make false or misleading statements. SCRIPSIT application of P3P raises the level of authority of the P3P policy statement to that of a positive and assertive agent through assignment of policy protection to wrapped SCRIPSIT entity content.

P3P user agent requirements are loosely defined and would fit well implemented as Java applets, JScript or other appropriate embedded scripting option. These user agents search for P3P policy information in the exposed parts of SCRIPSIT entities. SCRIPSIT's application of P3P has the absolute requirement that the P3P statements are always local, embedded items. These may be exposed and hence subject to public examination or concealed and subject only to authorised access by a suitably accredited

The only exceptions to this are P3P fragments and aggregations submitted by secondary users as part of access request (where no access rights currently exist) and as part of the granting or revocation of access rights by data custodians, owners and secondary reusers (where such rights exist by default or have been previously

granted). Published P3P specifications assert little in terms of meaning of symbols , thus leaving open the door for SCRIPSIT requirements and implementation.

2.8.1.2 Resource Descriptor Framework (RDF)

The Resource Description Framework (RDF), described on the W3C website (*RDF Modelling for P3P,* July 2000) defines a general and abstract model for representing metadata. The RDF Schema Specification defines a schema language for describing specific RDF information models. RDF Schema are used to describe an abstraction of the information model for P3P abstract information model (policies, references and schema). RDF's usefulness is extended by, amongst others, the Dublin Core Metadata Initiative (DCMI). DCMI simply describes a range of networked metadata entities. Baker (2000) describes DCMI as follows:

> *... (a) language for making ... class of statements about resources*
>
> (Baker, 2000)

Nouns and qualifiers exist, allowing the creation of statements which determine the subjects of the language. DCMI is able but not uniquely capable of determining specific meaning or processing direction. It enhances interoperability at basic levels of understanding.

Resource Description Framework (RDF) data consists of nodes and property/value pairs describing nodes. An node is any object which can be pointed to by a URI. Properties are attributes of nodes, values are either atomic values for the attribute or other nodes. Information about a research topic (a node), may include the property owner. The value for the owner property may be a string of text, a URI pointing to another document or a persona definition.

RDF defines metadata processing frameworks and data models based on triples (subject/resource, predicate/property, object/property value). Data graphs with unique

identifiers may be formed with these data triples. RDF forms the basis of tools able to link, classify and extend data and add subjective value. An example is the aggregation of a collection of XML documents into an RDF model. Document collections may be complete and fully-formed, they may be data fragments and they may also be networks of multiply-linked XML documents. This forms the essential basis of RDF/XML used as dynamic and extensible repositories. Semantically-dependent queries against knowledge encoded in an ontology are available via RDF/XML document networks.

Resource Description Framework is simply a framework for describing and exchanging metadata. The premises on which RDF is built are:

- Resources are any things which can be located. In other words, anything that can have a URI, literally anything - logical, referential or physical.
- Properties are named resources. Properties may have properties associated with themselves. Metadata metadata (data about metadata).
- Independence - Property is defined as a resource and may be created by any agent.
- Interchange - RDF Statements may be translated into XML and therefore interchanged as needed.
- RDF statements are simple data triples (Resources, Properties and Values) and are hence straightforward to discover and reference in context of the WWW
- Finally, Statements are made up of Resources, Properties and values. Formally stated, the Subject (Resource), the Predicate (Property) and the Object (Value).

These points are illustrated in Figure 1:

```
<rdf:description xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
 rdf:about="http://www.rodda.org/aboutme">
        <projectsInProgress rdf:resource="http://www.rodda.org/aboutme/myhistory.pdf"/>
</rdf:description>

<rdf:description about="http://www.rodda.org/aboutme/myhistory.pdf">
        <name>"PoliticalAffiliations"</name>
</rdf:description>
```

is RDF declaration of the graph following:

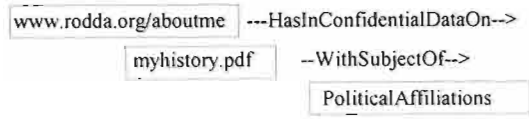


Figure 1 - Simple RDF graph representations

2.8.1.3 eXtensible Markup Language (XML)

eXtensible Markup Language (XML) is an appropriate medium for metadata because it is widely understood, extensible and processable. XML provides a facility to define tags and the structural relationships between data or metadata forming a syntactic (and later, a semantic) tree. There is no predefined tag set and no prescribed semantics. These are defined by the XML itself. Interpretive semantic operations are imposed upon the data and are not inherent in it. XML is a format for transmitting metadata independent of platform and application. Goldfarb (2000) succinctly defines XML as

> *XML data is smart data, HTML tells how the data should look, but XML tells you what it means but XML data isn't just smart data, it's also a smart document and you don't have to decide whether your information is data or documents; in XML, it is always both at once. Data processing or document processing or both may occur at the same time.*
>
> (Goldfarb, 2000)

eXtensible Markup Language is described on the official website (XML

Specifications, February 2004). The XML content in the SCRIPSIT context is limited to one or both of String or Base64. Plain text content is adequately addressed by the XML String data type and requirements of any binary content (images, audio clips) and of encrypted content (including encapsulated engines) are addressed by XML's Base64 data type (See UTF-7 and Base64 in the Glossary).

2.8.2 Standards as basis for innovation

This dissertation is built on the premise of achieving innovation and advancement in creation of trustable privacies through the development and use of existing, accepted standards. The distinguishing attribute of this work is a simple change of perspective in the application of published standards, these being P3P, RDF and XML. None of these standards are able, in isolation, to address the questions raised in this dissertation. A simple change in perspective has extended the usefulness of these standards. The Semantic Web and the Web Ontology Language (OWL) constitute a structural, though non-privacy enhancing technologies on which SCRIPSIT extensions may be considered.

2.8.3 Semantic Web and Web Ontology Language (OWL)

Both the Semantic Web and Web Ontology Language (OWL) are extensions to the set of tools available on the WWW which enable rich extension of functionality in terms of management of ontologies of domain knowledge and of resource discovery based on contextual and semantic cues.

2.8.3.1 Semantic Web

The Semantic Web is a rich extension of the WWW where information is assigned explicit meaning and context. It builds on XML's enabling of definition of tagging schema and RDF's ability to represent data explicitly and implicitly via metastructures. XML imposes no meaning on the contained data. One level of abstraction away is RDF. RDF is a data model for resources and relationships between them. It provides semantics for the data models RDF Schema is a vocabulary for describing properties and classes of RDF resources, with semantics for generalisation hierarchies of such properties and classes. OWL extends RDF by addition of vocabulary for description of properties, classes, typing and relationships. Bosak and Bray (1999) describe a richer Internet in terms of XML and Semantic Web enhancements, building on the foundation laid down by Tim Berners-Lee of the W3C.

2.8.3.2 Web Ontology Language (OWL)

An ontology defines the terms used to describe and represent an area of knowledge. Ontologies are used where there are needs to share domain information. Encoded knowledge in domains and cross-domain encoding makes knowledge more readily reusable and accessible. OWL is a Web Ontology language which is WWW-optimised and intended to work with the Semantic Web. The language started out as the "Web Ontology Language" but the Working Group disliked the acronym "WOL." The decision was justified in terms of "noted ontologist" A.A. Milne who wrote of the wise character Owl, in Winnie the Pooh:

> *He could spell his own name ... WOL, and he could spell Tuesday so that you knew it wasn't Wednesday...*
>
> (A.A.Milne, 1926)

OWL extends the reach of RDF by adding an ability to be distributed across systems, compatibility with W3C standards for accessibility, open and accessible standards

2.8 Key themes identified in privacy in public spaces

Privacy is a fraught battle, with access to useable technology generally restricted to an elite set of technologically-aware users. Creation of a trustable privacy model requires the simplification and generalisation of technology to allow it to be trustable because technologies and mechanisms are self-contained and controllable by the users. It is acknowledged (Zinnbauer, 2001) that users of the WWW are generally naïve about the risks and pitfalls involved when present online.

> *...technologies exist that allow to remain anonymous, prevent interception of email communication or route around blocked websites. ... The average Internet users does not command the technical competence and confidence to safeguard ... information privacy and anonymity in ... a technology race between an IT savvy regime and the development of subversive online tools.*
>
> (Zinnbauer, 2001:53)

Key themes identified by this literature review are:

- General requirement for technical competence to understand and to use PETs.
- Tension between institutional/governmental views of privacy as potentially subversive and individual views of privacy as providing subjective security and trustability.
- De facto dependence on trusted third party solutions for privacy problems.
- Requirement for architectural independence of trustable privacy models to enable unchallenged generalisation of principles and mechanisms.

2.10 Conclusions

A considerable body of research exists concerning privacy, archival of personal and private data. Much of this research covers privacy and trust models relying on the good offices of institutions, governments and commercial interested parties. Those models driven by mildly anarchic motives tend to excel at viral distribution of material (Napster, 2003; KaZaA, undated), but fail to present viable options for the archival and accessing of private data. Liang *et al.* (2004) expand the functions and architecture of KaZaA, highlighting KaZaA's suitability for file sharing and not for access mediation of third party access.

Literature presents models which effectively secure data or with models which are resilient, if restricted in their ability to be applied in multiple problem spaces. There is a paucity of research addressing questions of trustable privacies in public spaces from the perspective of the data owner, the source of data for all subsequent use. The failure of trusted third party negotiation protocols to adequately address privacy on the part of the individual is highlighted. Discussion occurs, in Chapter 5, of a potential third class of credential, namely user-assisted credentials in the form of accreditation fragments which are not uniquely sensitive to absolute possession nor to specific attributes.

Surveyed literature indicates a rare addressing of the personal and the individual with respect to privacy and its realisation. This translates directly into a lack of inherently trustable models for creation of privacies in public online spaces.

Amongst the problem spaces with issues of trustability, as referred to in this dissertation, are cradle-to-grave personal life records, displaced refugees, orphaned HIV children, tribal headmen interviewed and recorded *ab initio* in the written word, secondary school learners interviewed on personal topics and so on.

This research aims to add to the body of knowledge in the following areas:

- Establishment of a technically feasible architecture supporting creation of personal privacies in a multiplicity of environments and domains
- Linking of control, privacy and trust in a simple, unified architecture

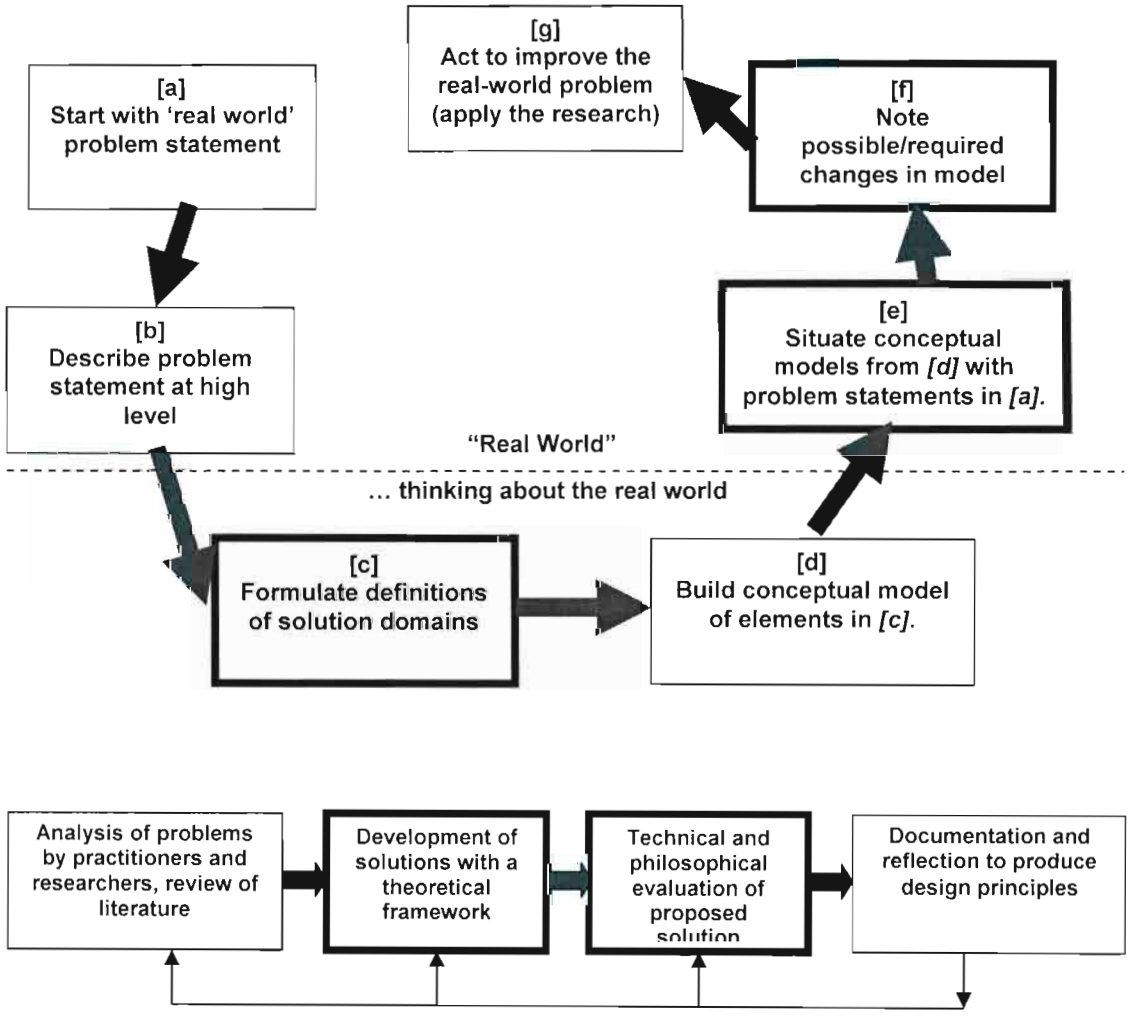
The dominant thrust in contemporary research and praxis in the digital privacy domain concentrates on trusted third party models relying on server-side and/or digital vault approaches. Here, the fundamental elements required to make a model trustable in terms of the perceptions of the individual are seldom addressed. When they are addressed, there is an almost inevitable assertion made that somewhere exists a trustable third party whose credentials do not require questioning. This research aims to propose options previously unexplored.

This review of relevant literature demonstrates the gap in approach and solution with respect to models and mechanisms supporting the creation of trustable privacies in public spaces.

Kubiatowicz (2003) observes that the behavior of distributed P2P systems parallels life, the whole being greater than the sum of its parts. Whilst there are no guarantees of anything in online public spaces, the model to be proposed is required to be simple in concept, diverse in application, and extremely tolerant of underlying network architectures. This aim is supported and validated by the body of literature reviewed.

Chapter 3

Theoretical framework and research methodologies



3.1 Introduction

A theoretical framework is outlined in this chapter for the intangible aspects of privacy and trust which define the tangible and quantifiable aspects of the model intended to support creation of trustable privacies. Research methodologies employed are discussed in this chapter. This is preceded by an outline of the original research plan.

The basis for research is a consideration of the intangible currencies of privacy and trust which underpin the requirements of mechanisms and models for creation of trustable spaces. These conflict with the accessibility and openness required of public data which is expected to be available and accessible in the public domain.

Establishing trust requires a credible and robust protection of privacy, no matter what the domain or environment. Privacy and trust in the digital world are perceptually and technically quite different from privacy and trust on a face-to-face level. There is almost always very little or no substantiation of the credentials of the 'other party' in the digital world. Protection of individual privacy is discussed in a community context, with trust as a construction based on respect for and reliability of individual privacy. Corti, Day and Backhouse (2000) discuss issues impinging directly on privacy and trust, namely confidentiality, informed consent and anonymisation of data.

The research methodology includes analysis of and commentary on real-world problems by the researcher and references to interviews and notes on meetings and email exchanges with researchers. A pilot questionnaire on perceptions was distributed to a number of researchers in different domains to help with the triangulation process. Reeve's and Hedberg's Development Research methodology (see Figure 2) was employed as the core methodology in this research. Development of a model is an iterative process and this was identified early in the design of this research project. It was necessary to adapt the Development Research methodology

through incorporation of aspects of the Soft Systems Methodology (Dick & Swepson, 1994) illustrated in Figure 3. An ability to iterate rapidly and make both detailed and high-level changes to the model was required and subsequently supported by the blended approach taken.

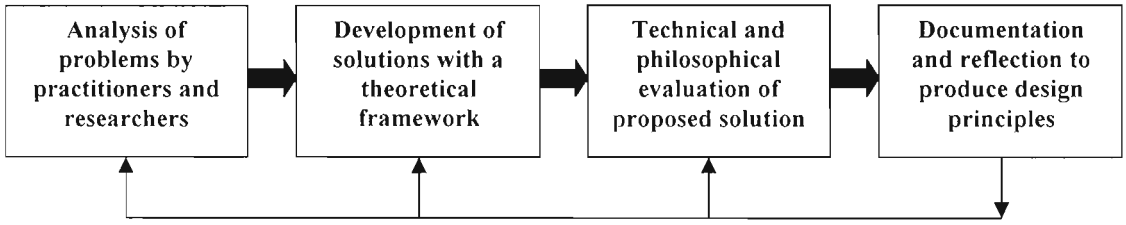


Figure 2 - Outline of development research model

Checkland (1990) describes a soft systems methodology, applicable in the Action Research[iv] domain. This is laid out diagrammatically by Dick and Swepson (1994) and succinctly illustrates the overall approach to the research component in this project. It is used in conjunction with the Development Research methodology.

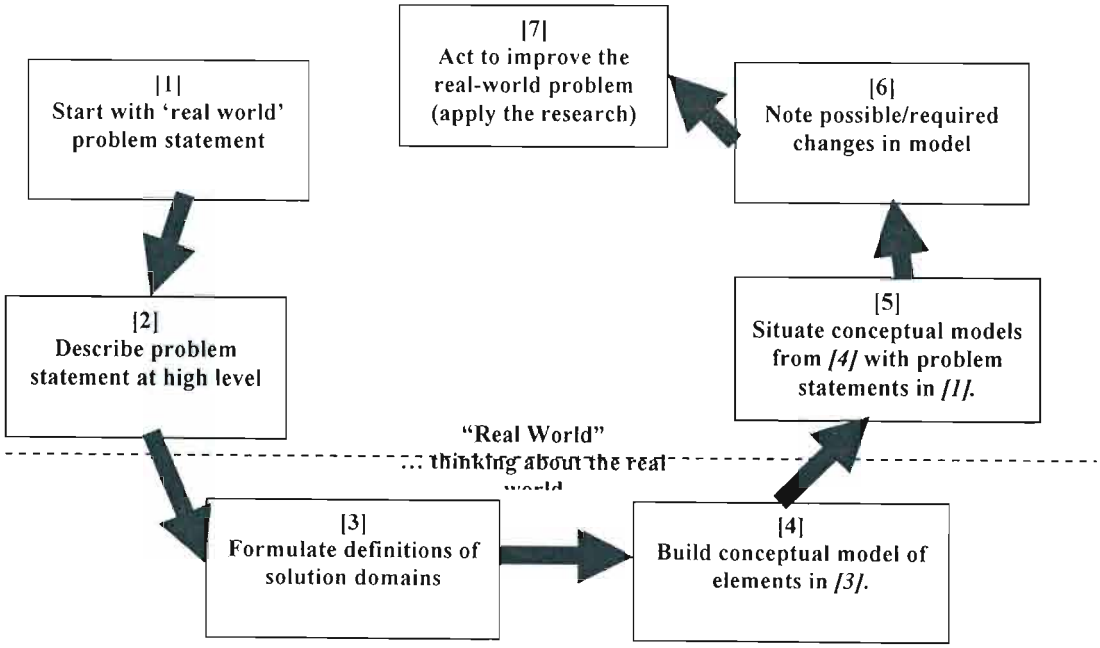


Figure 3 - Soft Systems methodology

[iv] Action Research is used primarily in the Human and Social Sciences, and is primarily a reflective, iterative process where solutions are discovered through progressive refinement. Rose (2000) provides situated examples of action research in the Information Systems domain.

According to Dick and Swepson (1994), achieving validity in real-world problem based research requires investigation into the veracity of different sources of information and perspectives on those sources.

This research established such a dialectic amongst the philosophical, ethical and technical domains, and this dialectic has informed much of the process of research and development of the model proposed. The fundamental research methodology to be employed is that of Development Research, with aspects of action research methodology incorporated. The purpose of a hybrid approach is consideration of the dual social and technical streams in this research.

Both the development research and soft systems methodologies are used as chapter prefaces throughout this dissertation to indicate chapter relevance and location within the informing methodologies. The iterative and non-linear character of the research highlighted by the prefaces entries.

3.2 Elements of research methodologies

Three categories of research epistemologies are listed in order to place the interpretivist research approach of this dissertation in perspective:

- Positivist research. Assumes reality is objective and may be described by measurable properties, independent of the researcher. Attempts to test theory to increase understanding of phenomena.
- Interpretive Research. Assumes that socially-constructed reality is through social constructions including consciousness and shared meanings. Does not define dependent and independent factors. Attempts a holistic consideration of problem domains and identifies key issues as work progresses.
- Critical Research. Assumes that social reality is historically based and perpetuated

by people within social groupings. Recognises that ability to alter conditions is constrained by social, cultural and political domination. Critical research pursues social critique where conditions of social status quo are considered.

The problem space which defined this research is clearly identified as a largely subjective space early on (Chapters 1 and 2). Control issues, shared meanings and social constructs place the approach and methods used firmly in the interpretive/critical research arena.

Further analysis of the defined problem space identified qualitative and quantitative methods, with triangulation employed as a means of testing convergence of conclusions drawn from literature reviewed and assessments of tools and methods employed. These are described in summary below:

- Qualitative methods. Methods employed included extensive interpretive review of published literature on philosophical views on privacy, trust and security in order to establish gaps, omissions and opportunities for innovation. A short perceptions and preferences questionnaire was employed (primarily) as a means of verifying conclusions drawn from the literature review and (secondarily) as tool to establish a basis for more extensive survey as a part of future work.
- Quantitative methods. Methods employed included feature and capability assessment of CAQDAS tools and information security features. This was used to inform qualitative evaluation of tools, standards and methods related to the technical requirements of privacy and trust models.
- Triangulation. Not strictly a method, more a means of using multiple methods and sources to study and test a common problem statement. Inadequacies in one method are usually highlighted and compensated for by better competencies in others. Weaknesses in any one method are compensated for by the strengths of another. Triangulation was employed in this research to test for convergent results in assessment of strengths and

weaknesses of tools, approaches and standards relating to creation of trustable privacies in public spaces.

3.2.1 Qualitative - social and technical analyses of problem

A strong reliance on the perspective of the individual, obtained by interpretation and critique of views of privacy in reviewed literature (Chapter 2), within society has formed much of the philosophical basis of this research. Myriad technological solutions proposed for solving the issues around creation of trustable privacies were noted in Chapter 2.

3.2.2 Quantitative - technical assessments of model and methods

Technical assessments of models and methods were used to elect attributes and features of alternative models which have merit. Similarly, technical assessments were used to disqualify some models and certain categories of superficially appropriate solutions for public, mediated access to private qualitative information, with KMSs referred to as the primary disqualified category of tools.

3.2.3 Triangulation

As the attribute fundamental to the establishment of trust, and the attribute most closely linked to technical support by information security considerations, privacy was considered from a number of aspects, namely

- Institutional,
- Individual,
- Information systems, and
- Social

Triangulation revealed an almost complete absence of privacy-enabling aspects in tools, systems and philosophical approaches. Closely allied to this absence is an immediate concern arising over perceptions of trust, including asserted trust, trusted third parties and a mechanistic view of trust (presuming trust to be a simple personal assessment of risk). Viewing questions around privacy and trust from a doggedly individual perspective provided strong support for the assertion of the need for a model supporting creation of trustable privacies in public spaces.

Interviews with researchers were used partly to provide an additional source of opinions to help validate conclusions drawn from literature and evaluations, and to provide an indicator as to whether or not future survey work was required. The survey performed made it apparent that there is room for more in-depth interviews and opinion surveys, extending the reach to other affected parties including asylum seekers, diasporic communities and other displaced parties.

3.3 Survey and Interviews as informers of methodological and framework

3.3.1 Unstructured interviews

Unstructured interviews were conducted in the period from September 2003 to

August 2004 with researchers, including Dr. Louise Corti of Qualidata (Essex University), Dr. Patrick Carmichael (Cambridge), Dr. Stephen Heppel (Anglia Polytechnic University) and Dr. Jenny Preece (University of Maryland in Baltimore County). The primary purposes of these interviews were:

- to test concepts underlying this research against a range of domains and acknowledged domain experts (interviewees),
- to inform the development and research processes through incorporation of comments made by interviewees, and
- to gauge potential areas of application for the model under development.

[1] Dr. Louise Corti was interviewed at Qualidata's offices at Essex University in September 2003. Dr. Corti has a professional interest in online access to archived qualitative research data, and therefore offered the promise of being able to make targeted criticisms of aspects of the trustable privacy model as it was constituted at the time of the interview. Dr. Corti's extended involvement with the UK Data Archives, with academic, government and industry gathering, analysis, archival and primary and secondary access to qualitative research data identified her as a key expert interviewee. Brief interview notes follow, with interview conclusions after.

> The embryonic trustable privacy model was discussed with Dr. Corti and situated with respect to longitudinal qualitative archives (a key example being Edwardians Online). Questions around access and control of qualitative data were discussed, with particular emphasis by the interviewer on data where the data owner is both alive and in need of a model which presents a viable possibility for retention of control and ownership to that data owner.
>
> Some skepticism was expressed when the possibility of a peer-centric trustable privacy model was discussed. There were initial questions as to why the "trusted 3rd party" option was not a considered option. This was explained by the interviewer from the perspective of the perceptually betrayed data owner.

It was commented that such a model had the potential to extend the usefulness and reach of existing qualitative data archives. Issues of resource discovery were noted as being critical to such a model succeeding.

A further note was made that extended versions of tools such as QSR International's NVivo would enrich the secondary researcher's ability to access and use archived qualitative research data.

Conclusions drawn from this interview were:

- Explanation of how the peer-centric model differed from client-server and peer-to-peer models would be required both as a differentiator and as a technical support for the claim to trustability.
- Future work would involve approaching organisations such as QSR International with proposals for extension of their products with modules supporting the trustable privacy model.
- A programme of researcher and subject education would be required for successful acceptance of such a model in research communities.
- Such a model would be required to integrate successfully with existing archives and infrastructural investments.

[2] Dr. Patrick Carmichael was interviewed at Cambridge University in September 2003, after responding to email enquiries centered on the basic premises of the research into creation of a model supporting trustable privacies in public spaces. An additional reason for interviewing Dr. Carmichael was his awareness of ICT and displaced persons in Africa, through involvement in activities around the Rwandan genocide survivors and in the Blue IQ initiative in Gauteng province, South Africa. Brief interview notes follow, with interview conclusions after.

Discussed application of peer-centric model to displaced communities

(references Rwandan genocide survivors). It was noted that the needs of displaced persons centre primarily on a basic level of survival and then only on information persistence. Once basic survival aspects are in hand, there is the requirement to be able to prove identity, title, assets and a host of other attributes and rights linked to the individual.

Considering political upheaval, economic breakdown, and the violent dispersal of families and communities, a model offering ubiquitous and multimodal access to such information was commented on as being a welcome possibility.

Concepts of trust and privacy in the above context are shown in sharp relief when the Rwandan example is considered. High level discussions and considerations of privacy and trust tend to develop sharp focus when contextualised by specific instances as described.

Technical feasibility of a model supporting flexibly defined information collections, while still supporting preservation of access, control and security of such collections and their composite elements was commented on favourably.

Conclusions drawn from this interview were:

- Access to personal (private, confidential) information is not a need limited to those in First-world, industrialised nations. It is of particular importance to those who find themselves in undeveloped and developing nations with a history of upheaval and politically-inspired violence.
- Such a model is feasible, given a technically independent architecture
- Privacy must be considered from the perspective of the individual who has experienced betrayal of trust.

[3] Dr. Stephen Heppel, head of ULTRALAB at Anglia Polytechnic University, was interviewed at the ED-Media 2004 conference in Lugano, Switzerland. This interview was brief, following Dr. Heppell's keynote address. Interview notes follow, with conclusions after.

- The NotSchool (http://www.notschool.net) initiative was discussed, with the aim being aimed "...the reengagement of children into learning" (Stephen Heppell, interview, 24 June 2004). Targeted children have been denied access for a number of reasons, and thoughts around a model supporting trustable privacies were discussed in this context.
- Key issues discussed were preservation of perceived controls of personal data and interactions in the NotSchool context – these were held to be crucial by Dr. Heppell.
- Multimodal aspects covered as extension of Dr. Heppell's discussion of cell phone-based assessments of pupils in schools, and with reference to NotSchool and its objectives.

Conclusions drawn from this interview were:

- Applications for trustable privacies may be found in domains such as that of http://www.notschool.net, an environment for displaced and otherwise disturbed children – truants, those in social care systems, socially maladjusted children and others.
- Any model proposed would need to show a fundamental ability to integrate with existing initiatives and use cross-platform technologies.

[4] Dr. Jenny Preece, of the Information Systems Department, University of Maryland in Baltimore County, was interviewed at the ED-Media 2004 conference in Lugano, Switzerland. The interview was conducted after her keynote address.

Dr. Preece commented on the lack of consideration given to both social and technology questions where online communities are concerned. The interviewer asked of Dr. Preece's concerns regarding privacy and trust questions in online communities and the response received indicated that the initial comment on lack of consideration of social and technology questions was probably key, and ought to form a mainstay of the research into creation of such a model.

Discussion around the design of a trustable privacy model was centred on the social and community consideration aspects, many of which have subsequently fed into aspects of the control and privacy mechanisms developed.

The primary conclusion drawn from this interview was:

- The premise that the perceptions and needs of the individual were paramount and were superior to the technology employed. In essence, the technology must follow the perceptions of the individual and of the affected community.

3.3.2 Pilot survey on perception and opinion

A perception/opinion survey was conducted in order to provide a means of testing convergence of conclusions from literature reviewed, conclusions and assessments drawn from evaluation of tools and standards, and initial assumptions made during the design of the original project proposal. An additional aim of this survey is to establish a basis for more detailed and extensive survey as a part of expected future work.

This survey was targeted at a diverse academic research audience. Respondents included those from media and communications, information systems, anthropology, sociology, life sciences and educational technology. Geographic diversity was evident in respondents' locations in South Africa, the United States and Europe.

Criteria for selection for participation in this survey were:

- Professional interest in creation of archived/published qualitative research data, and in secondary use of existing qualitative research data (with publications indicating such interest as an additional qualifier)
- Diverse professional domains
- Diverse institutional associations

The survey was kept short and uncomplicated to elicit direct opinion and commentary from participants unaware of the specific or detailed nature of the research being conducted. Following is a narrative presentation and discussion of responses received. A need for more a detailed survey as part of future work is apparent from the respondents' comments.

Following is a summarised account of the responses received. Respondents included academics involved in communications research, new media/sociology, information systems, library and archival services and the natural sciences:.

A prose discussion of responses received may be found on the page following. A copy of the original questionnaire may be found in Appendix A.

Questions 1 and **2** (professional/private, respectively): Please indicate your THREE primary professional/private interests in qualitative data archives (in descending order of importance). Replies are open-ended:

Replies for this question included the following interest domains and more general notes:

- Media studies
- South African Studies
- Political ideology
- Social movements
- Social activism
- Reuse of data for alternative research questions
- Facilitation of longitudinal studies
- More effective use of limited research resources
- Retention of integrity and privacy of data
- Archival and library services (digital and paper)

The questions (1 and 2) were open-ended with the intention of eliciting responses which would inform the design of questionnaires for surveys in any work arising from this research.

Question 3: Please indicate your perception of the degree of control appropriate to place in the hands of research subjects, with respect to access to data specific to individual subjects. Check ONE BOX only to indicate.

[ ]None,
[ ]Request via researcher,
[ ]Mediated by trusted 3rd party,
[ ]Full (mediated) and
[ ]Full (autonomous).

In all responses received, the indicated option was Full (mediated) or Full (autonomous). The uniform non-selection of the Trusted Third Party (TTP) option was noteworthy as a response to institutional assertions of acceptance and trustability of third parties.

Question 4: Please indicate your perception of the degree of control which you feel appropriate for your own control over your own private (non public, non research-related) data. Check ONE BOX only to indicate.

[ ]None,
[ ]Request via researcher,
[ ]Mediated by trusted 3rd party,
[ ]Full (mediated) and
[ ]Full (autonomous).

Again, in all responses received, the indicated option was Full (mediated) or Full (autonomous). A repeat of the uniform non-selection of TTP as an option is noted again. Of interest is the correspondence in clustering of responses in this question (private data) with that of question 3 (archived / publicly accessible data).

Question 5 : Please describe briefly your views on the general usefulness of a model supporting distributed, user-controllable granting and rescinding of access rights in

[a] the qualitative research arena

- Will empower learners in the use and integration of subject specific information into the academic study environment, if necessary control and acknowledgement is built into the system. [Communications/media researcher]
- ...concept you're working on - full control over all information that can then allow access to be granted or rescinded as appropriate - is a great one and equally applicable (though obviously different) to both qualitative research data and personal privacy. [New Media/sociologist researcher]
- Has great potential once its ability to stand up to hacking has been demonstrated... [Business Information Systems researcher]
- From a repository perspective, this is a policy decision, and network controlled [Archivist/Librarian]

[b] in terms of preservation of personal privacies in public spaces

- Will ensure peer accreditation and acknowledgement within the qualitative research arena. [Communications/media researcher]
- ...concept you're working on - full control over all information that can then allow access to be granted or rescinded as appropriate - is a great one and equally applicable (though obviously different) to both qualitative research data and personal privacy. [New Media/sociologist researcher] *This response was identical for both questions.*
- Has great potential once its ability to stand up to hacking has been demonstrated... [Business Information Systems researcher]

One respondent identified a link to constructivist learning environments which was not identified in the original investigation into potential areas of application of a

model for trustable privacies in public spaces. This respondent additionally comments that such a model would "ensure peer accreditation". This comment is of considerable interest in that it identifies one of the key attributes envisaged as validating the model proposed in this dissertation. Another respondent's comment that such a model has potential, once an ability to resist attack ("hacking") has been demonstrated, is supportive of the concept of an encrypted, embedded engine to reduce vulnerability to attack/unauthorised accessing of contents.

This survey served to confirm the assumptions made regarding the necessary attributes of a model for creation of trustable privacies in public spaces. Additionally, the survey results were convergent with the interpretations of literature reviewed in Chapter 2.

3.4 Research activities

Research activities occurred in an iterative and incremental fashion, subject to reflection and review at each step. Steps were laid down in a linear fashion, but visited iteratively throughout the process of this research. Most of the activities listed below in Table 7 correspond with steps in the soft systems methodology (Dick and Swepson, 1994):

Table 7 - Research and reflective activities

Activities	Chapters
• Analysis of practical problem statements • Contextualisation problem statements to existing models and contemporary praxis[v].	1,2,3,6
• Considered the development of solutions incorporating requirements of multiple problem domains (philosophical, ethical, technical, practical) into a theoretical framework. • Identification of the key attributes of privacy-enabling models in order to inform the conceptualisation process.	2,4,5
• Technical and philosophical evaluation of tools, models • Identification of gaps and weaknesses in tools and models used to further inform the refinement of the model development process. • Identification of appropriate standards and platforms as components of the model.	2,3,4,5
Model conceptualisation in terms of: • Function • Standards • Structure • Process support for privacy and trust-building requirements	3,4,5,6
• Analysis of, and commentary on, actual problems and scenarios. • Reflection on interviews, meeting notes and email exchanges.	2,5,6
• Conclusions and discussion on future work	6

[v] Reeves and Hedberg, (2003)

3.5 Theoretical framework

It is not sufficient to take a simple view of mechanistic privacy, otherwise known as information security, and combine this with guarantees and assertions.

Trust forms when communities of practice mutually negotiate terms of engagement and exchange. To this end, brief notes on communities of practice, social constructivism and aspects of privacy and trust are presented as a part of the notes on the selected theoretical framework.

3.5.1 Constructivism and social networks

Social constructivism forms the basis of the emergent position that information and context is added to mental scaffolds in a uniquely individual manner, and that knowledge is not inherent in the information available, but rather in the manner in which that information is added to the internal mental scaffold. Social constructivism holds that members negotiate meanings, set contexts and form consensual pools of accepted practices.

The relationship between social constructivism and the creation of trustable privacies in public spaces is that which links the development of individual trust in a model to that model's ability to simultaneously keep control over personal data in the hands of the individual and to allow that same personal data to be accessible in communities of practice.

Communities of Practice may be informal or formal, consist of one member or of a very large number of members, exist in a single location or scattered around the globe in a diasporic manner. Wenger (2001) comments in depth on the operation and function of communities where members form social networks. Communities of practice are based on the assumption that these communities arrive at negotiated agreements on the goals and purposes of the community. Members further review

their interaction and participation with reference to the guidance provided by the selected CoP.

3.5.2 Privacy and trust effects on methodologies employed

Brunk (2002) takes a view of privacy as an example of an applied human value (or set of values). Privacy is presented as a primary human value, an exemplar, of the basic desire to control personal information flows. Privacy and related ethical considerations emerged as strong indicators in the design of privacy and trust systems. Brunk, in particular, had a significant influence on aspects of the model development process. The notable point is made that systems may be designed with integrity and privacy in mind, or may be retrofitted with privacy-enhancing features. A definition of online privacy is offered:

> *... we define "online privacy" as having the ability to control information leaving you while online, and being able to exercise that control consistent with your values. In a passive sense, privacy is also about being able to control unwanted intrusions. We claim that people seek designs that provide easy and effective ways to achieve online privacy, verify that they have done so, and monitor effectiveness.*
>
> (Brunk, 2002)

This emphasis on being able to exercise control consistent with personal values corresponded with initial assumptions made when defining the problem statement. Further, this has influenced the nature of the privacy and trust model developed. Following on from this argument is the established premise of trust being built on credible privacy. This emerged during the course of the literature review as a requirement of the framework used to develop the model for trustable privacies. Seamons *et al.* (2002) propose negotiated trust of a variety not unlike that of the negotiated trust experienced when negotiating with a vendor over price and conditions of sale. A cautionary note was observed with this proposal of negotiated trust. This is

an instance of a flavour of trust incompatible with the concept of trustable privacies in public spaces. Trustable privacies are not about exchange of value-bearing tokens, nor about the extension of trust credit or mutual deterrence with respect to unauthorised release of information.

3.5.3 Standards and theoretical extensions

The decision to base the proposed model on published standards allows the model to be developed with a set of known frames of reference[vi] in terms mechanisms at its core (XML, 2004; RDF, 2000; P3P,undated). An additional and desirable consequence is that the model proposed became elegant and simple in conception and design.

3.6 Conclusions

Chapter 3 situates the research linked to the development of SCRIPSIT in terms of the use of multiple perspectives of the problem space. It is acknowledged and readily apparent that creation of trustable privacies in public spaces is as much about technological ingenuity as it is about the intangible nature of individual perceptions. For these reasons, this dissertation considers both technological and perceptual issues. The use of questionnaires in this research was aimed at confirming and extending the opinions and positions stated in the referenced research and publications of researchers and of end users of qualitative data.

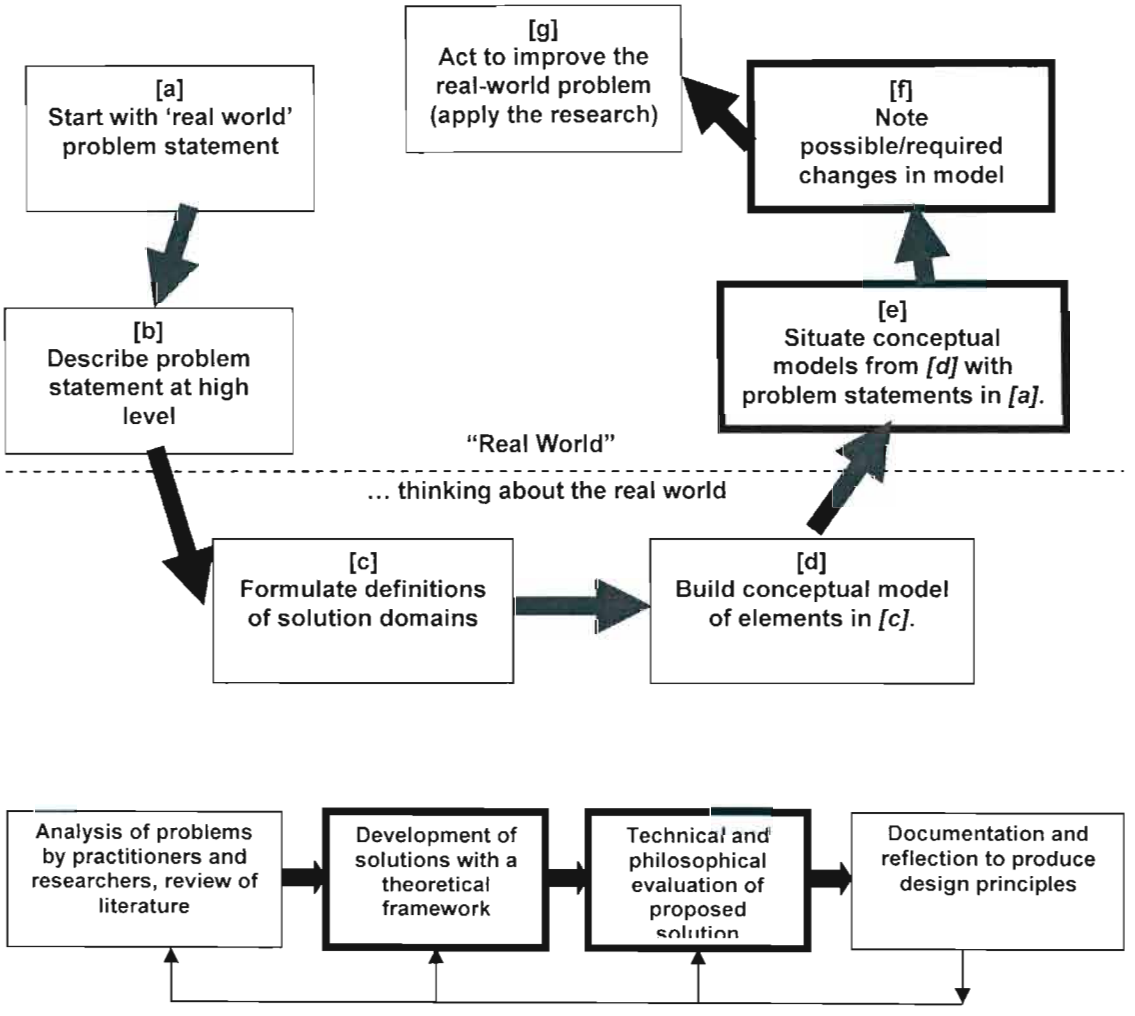
Triangulation of the proposed model provided a degree of initial validation of philosophy and approach. This is performed from a qualitative view of attributes and perceptions of individual privacies, a qualitative and technical view of CAQDAS tools and their application, and from an assessment of fit and application of the proposed SCRIPSIT model.

[vi] The World Wide Web Consortium (W3C) has a globally-based peer-review and publishing process for standards and working groups.

Chapter 4 builds a review basis of existing qualitative data tools and applications, with the intention of proposing means of extending the usefulness of the tools through functional and data integration with the model proposed in Chapter 5.

Chapter 4

A view of qualitative data analysis tools and Knowledge Management Systems



4.1 Tools for access to, and manipulation of, qualitative research data

The very best tools and mechanisms for data analysis and reuse are those which integrate functionality within applications with external resources. Allowing a multiplicity of access and analysis options as an inherent feature is the ideal which is sought. Qualitative data analysis tools are discussed with reference to two points:

- Use as tools for access, management, analysis, communication and dissemination of qualitative data.
- Suitability for extension of functionality through integration with a trustable privacy model.

Muhr (2000) comments on the advantages of common data formats where output from transcription systems is available as source for analysis systems. The model being developed seeks to abstract this common format and provide a catholic environment in which security, privacy and inherent trustability are native. Parallel to this lies the resource linking to the concept of the Semantic Web. This chapter touches on tools for access to and manipulation of qualitative research data, Knowledge Management Systems (KMSs), and publicly accessible data archives.

Babble (Thomas, Kellogg and Erickson, 2001), a social knowledge building tool, is discussed briefly to highlight conceptual elements supporting the need for persistence and resilience. Babble addresses aspects of the social nature of stored information and knowledge; it does not consider the tension between community and personal privacy. The proposed model will endeavour to accommodate both community and personal (individual) requirements. *Babble* shares conceptual elements with *Publius* (Waldman, Rubin and Cranor, 2000), a censorship and tamper-resistant web publishing environment which met a developmental dead end with the realisation that it did not a commercially-sustainable charge model, intended to levy a charge per use to generate an ongoing income stream. Another example of a social tool, albeit with

no pretence of social knowledge building, is *Crowds* (Reiter and Rubin, 1998). *Crowds* attempts to provide security and privacy enabled infrastructural elements, though without any fundamental support for semantic and resource discovery requirements.

The purpose of this overview chapter is to link common themes and requirements as informers of the development of the proposed model.

This section provides a summary of the attributes of some of the better known Computer Aided Qualitative Data Analysis Software (CAQDAS) tools. It is intended to inform applications and aims of the SCRIPSIT model later in this dissertation (Chapters 5 and 6).

A representative sample of mainstream CAQDAS tools is illustrated in Table 8 following.

Atlas.ti, NVivo and N6 (QSR International, 2004b), of the CAQDAS tools listed, support XML as data format. Observations on the sample listed are:

- Atlas.ti's *Networking tool* allows links to be created amongst quotations, codes, documents and memos, and its *Object crawler* allows searching for strings, keywords, phrase across entire project.
- QSR's N6 and NVivo allow qualitative cross-tabulations (matrix searches).
- Of the sample, only QSR N6, QSR NVivo and Atlas.ti support XML in any meaningful manner.
- Those supporting structured, open data formats and having proprietary differentiating features appear to lend themselves to extension through support of the trustable privacy model being developed.

Table 8 - CAQDAS tools

QSR NVivo	NVivo is aimed at annotation and organisation of qualitative research data. It is particularly suited to structuring and organising data, allowing multiple perspectives of data. (QSR, 2004a)
QSR N6	N6 is designed to code text and to facilitate search and navigation of locally stored research data. (QSR, 2004b)
Atlas.ti	ATLAS.ti supports annotation of text, video and audio. There are tools for assisting in the categorisation of data. Support for causal networks is built in. XML, raw text and SPSS data formats are supported. (Atlas.ti, 2004)
The Ethnograph	The Ethnograph is one of the earliest popular CAQDAS applications, though it has not been updated since the late 1990s. It supports hierarchical coding, text annotations, and advanced data search strategies. (Qualis Research, 2004)
MAXqda	MAXqda (successor to winMAX) a less widely known functional analogue of QSR NVivo and ATLAS.ti. (MAXqda, 2004)
Kwalitan	Kwalitan is aimed at development of grounded theory analyses of qualitative data. Uses hierarchical coding and boolean searches of data. Less well known, with little in the range of options for data import and export. (Kwalitan, 2004)

4.2 Tools and resources for mediation of access to qualitative research data

Carmichael (2002) identified a failing of many CAQDAS tools as being a lack of extensibility and operating system. Carmichael (2002), Kuula (2000) and Fielding (2000) noted a need for a platform-independent, network aware and enabled, groupware-oriented application and/or set of services for qualitative data access purposes. Carmichael further (2002) comments that there is a universal acknowledgement that eXtensible Markup Language (XML) is a common format denominator at the data presentation level.

Privacy-enhancing technologies (PETs) are protocols, standards, and tools which directly help with protection of privacy. This is done by eliminating or minimising collection of personally identifiable information. Phillips (2001) observes that

entrepreneurial players have started to offer PETs as the issues around preservation and creation of online privacies have gained visibility. The Electronic Privacy Information Center (EPIC) has a comprehensive list of PETs on its website (www.epic.org).

None of these tools and technologies is sufficient in isolation, nor are any able to provide a comprehensive, open technology solution to the requirements of qualitative research data reuse and the creation of trustable privacies in public spaces.

This dissertation aims to provide a model which establishes an open technology basis for existing CAQDAS access and annotation tools to extend their reach without compromising the need for confidentiality and security of data. It is a *sin qua non* of an open technology access mediation and encapsulation model that it has application in the wider social context, and that this includes support for the creation of trustable privacies in public spaces. Examples of tools supporting dispersed use of qualitative data are listed in Table 9 below:

Table 9 - Examples of access and usage mediation initiatives

Basic Support for Collaborative Work (BCSW)	Basic Support for Collaborative Work (BCSW) is a free groupware application which enables network users to distribute, share and discuss documents. These may be text based, HTML, images and Microsoft Office documents (Word, Excel). BCSW is an example of a conceptually sound application which is ultimately not able to provide guarantees of access or control to the original research subjects, nor is it able to move away from the need for a server resource in order to exist.
Escalate	The Escalate project (Grey, 2004) is aimed at the deployment and evaluation of a web-based collaborative tool intended to foster discussion and development of ideas. Escalate provides the means for users to contribute to an online database of material available to other users. Annotation and comments are visible to others. The project has aspirations of wider, cross institutional application. A possible extension of Escalate is into secondary use of qualitative research data.

4.3 Concerns around mediation of access to qualitative data

Amongst the fundamental concerns around access to qualitative data are confidentiality and lack of trust based on fear, insecurity, ignorance or arrogance. The broadening of the qualitative research data user community is described by Williams as follows:

> *... not just social scientists require research training but also G.P.s [medical doctors], nurses, midwives and health policy analysts are encouraged to become at least research literate.*
>
> (Williams 2000: cited in Fielding 2000)

This expanding collection of user communities is direct motivation for investigation of alternatives for the protection and management of qualitative data with ethics and confidentiality considerations. There are manifold archives of data, many purporting to be public data archives. It is apparent that the term *public data archives* refers, in the majority of cases, to public data which has been archived. It does not expose many archives of data (public and otherwise) which are necessarily *publicly accessible or controllable*. Listed in Table 10 are examples of qualitative data archives.

Table 10 - Public qualitative data archives

Qualidata / ESRC	The ESRC Qualitative Data Archival Resource Centre (better known as Qualidata) is supported by the Economic and Social Research Council (ESRC). ESRC has a Datasets Policy which offers data generated from ESRC-funded projects for archiving. ESRC has a mix of digital and paper-based archive data. There is an ongoing initiative to digitise paper, audio and video material and mediate access to these data.
VERBATIM	French electricity company, Electricité de France, has a social science research group tasked with qualitative surveys to better understand customer requirements and problems encountered by employees. VERBATIM aims to archive surveys and studies and to facilitate reuse of the same. The archives are closed access but public in data collecting extent.
FSD (Finnish Social Science Data Archive)	FSD is an independent unit at the University of Tampere. The primary task of FSD is to promote use of existing social science data in Finland. Functions include acquisition, archival and dissemination of data for reuse (primarily in secondary research). There is ongoing state support for reuse of research data, especially in support of the intention to create an information society (sic.) in Finland.
Murray Center (Canada)	The Murray Research Center is a Canadian national repository of social and behavioural science data on human development and social change. There is specific emphasis data on the lives of American women. Access is limited to qualified scholars and researchers for reuse, secondary analysis, and follow up (longitudinal) studies. There circa 300 data sets including detailed in-depth interviews and open-ended surveys. Whilst it is claimed as a public data archive, there is no facility supporting insertion, amendment or deletion of data by research subjects.
SADA (South African Data Archives)	Serves as a broker amongst a range of data providers, including statistical agencies, government departments, NGOs and academic users.

It is characteristic of all of the public data archives encountered that there is no independent access facility for research subjects to access data specific to themselves in any way. Pratt (1978) describes Hegel's transference of the concept of personal character to institutions, cultures and nations. It is further noted that perceptions of the individual tend to be subsumed by the stated interests of greater society. Highlighted here are aspects of difficulties encountered in moving from the specific case of the individual to the general case of a culture or greater society.

4.4 Knowledge Management Systems (KMS)

For the purposes of this dissertation, knowledge is defined in the broad context as that experiential and factual resource which is synthesised by an individual with reference to an internally constructed framework. A social constructivist perspective further reinforces the proposition that knowledge (in a generalised sense) does not exist outside of the consciousness of the individual. This does not preclude the exchange of knowledge through information transmission and synthesis on an individual level. In other words, a personal understanding of a common pool of contextualised information and semantic situation. Knowledge Management Systems (KMS) are, of necessity, usually domain-specific and intended as repositories of focussed information with specific and tightly defined user communities. Jonassen, Beissner and Yacci (1993) comment that the degree of integration of domain knowledge is best described as structural knowledge. This is explained as the knowledge of interrelationships of concepts within a domain.

From this, it may be postulated that culture and context are fundamental to the construction of knowledge in a societal context. This position is supported by Bandura's social cognitive theory (1986). A general statement made by Fitzgibbon and Reiter (2003) is that the wider challenge in information systems is the extraction of knowledge from data, and the subsequent use of this extracted knowledge in the creation of further tools. Inherent in this statement is the flawed reasoning that it is possible to generate knowledge from a broad pool of data which may or may not have

sufficiently large contextual and semantic payloads to enable such synthesis of knowledge. Fitzgibbon and Reiter (*op cit.*) note that there is discordance between the technology available in terms of the building of diagnostic and expert systems and the available data required to successfully build autonomous knowledge-driven systems.

There is much to suggest that confusing KMSs with machine extraction of contextually-relevant information (and, indeed, "knowledge" as a broad construct) is a persistent danger. This dissertation suggests that KMSs are invalid as management options for creation of trustable privacies in public spaces.

A summary of the positive and negative attributes of knowledge management systems is found in Table 11 below:

Table 11 - What Knowledge Management Systems do more and less well

What KMSs do well	What KMSs do less well
Answer domain-specific queries	Generalise responses across domains without loss of context
Provide a common interface to varied information repositories	Incorporate cultural and other social cues in execution of queries and mapping of information
Extend the reach of social networks	Preserve confidentiality, and hence, privacy of information
Enhance the accessibility of domain experts	Maintain accessibility across domains

This summary provides some justification for the development of a more general model for the encoding of contextually-linked information. KMSs tend towards providing a general and broad access to any prequalified member of a defined user community. Implicit in this access is a general acceptance of specific data queries. This contrasts with the philosophical and perceptual requirements for mediation of access to trustable privacies which are specific and qualified acceptance of both specific and general data queries. KMSs are broadly disqualified as appropriate models with respect to representation of meaning and context. It is superficially evident that KMSs and trustable privacies are at least partly incompatible. What is

less clear is the degree of incompatibility between KMSs and the broad category of public data archives.

4.5 Mechanisms of access mediation and qualitative data management

Mechanisms of access mediation are conventionally reliant on retention of access control by the researchers, usually as a part of the duty of ethical responsibility borne by the researchers on behalf of the subjects (Corti, 2000; Fielding, 2000; Roberts and Wilson, 2002).

A substantive change in the tools and mechanisms available to the research community, subjects, and would-be casual browsers of qualitative research data is required to extend the functionality and flexibility of the qualitative data analysis domain. Separating the duty of ethical responsibility from the archived data itself is the challenge to be partly addressed by the trustable privacy model proposed. Management of archived qualitative data would benefit in kind from such separation of the duties of care from the data,

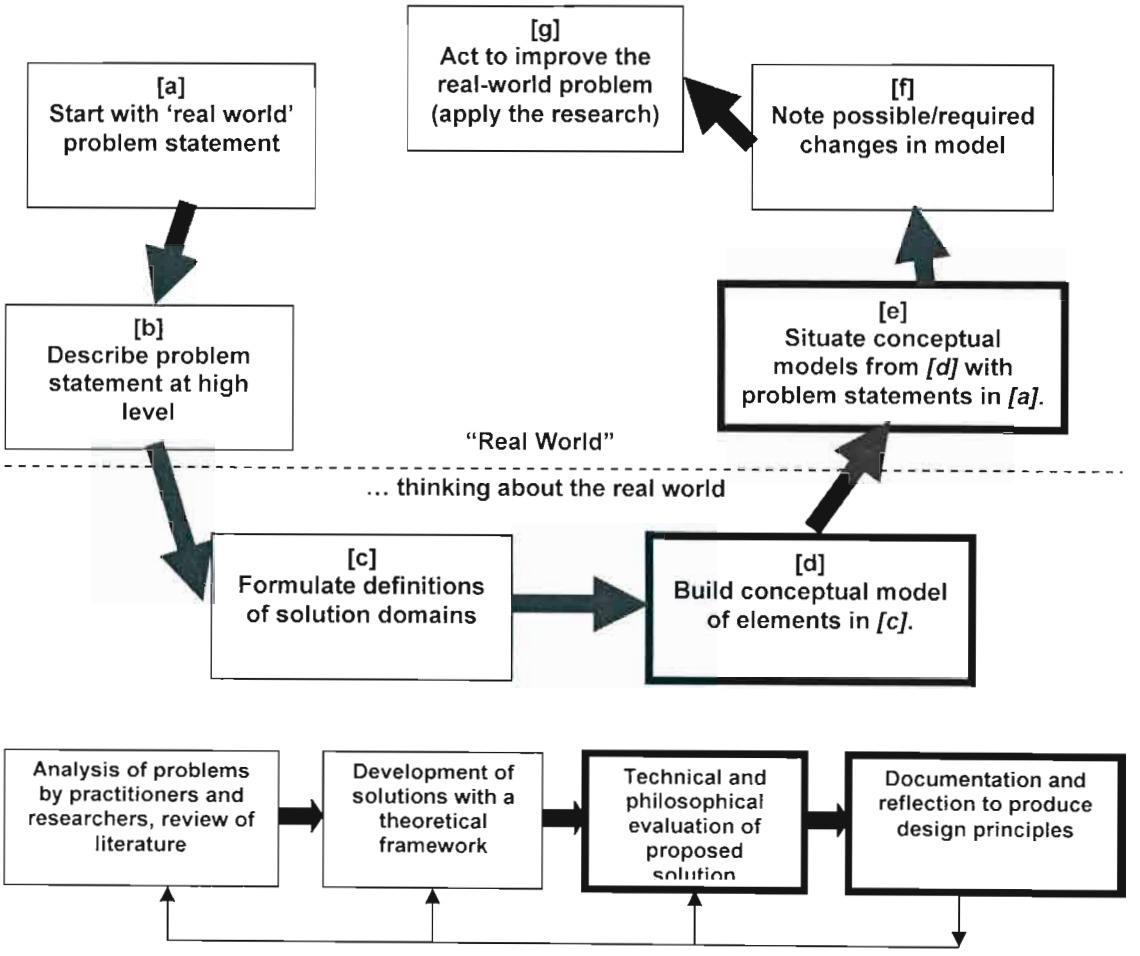
4.6 Conclusions

Knowledge management systems, Computer Aided Qualitative Data Analysis Software tools and file sharing peer-to-peer networking paradigms all address specific aspects of access, security, privacy and related domains. None of these support user-centric control and management of data. For this, a model is required which is able to translate into multiple and diverse application domains. Such a model is proposed, with situated examples of applications, in Chapter 5.

Chapter 5

Model and situated application frameworks for encapsulated peer nodes

Exposition of the SCRIPSIT model in structural, functional and process paradigms. This chapter presents the model and lays the foundation for implementation and extension of this research work.



5.1 Introduction to proposed SCRIPSIT model

> *A good theory of trust cannot be complete without a theory of control.*
>
> (Castelfranchi and Falcone, 2000)

Castelfranchi and Falcone succinctly sum up the argument in favour of the user-centric privacy and trust model as proposed in this dissertation.

Chapter 1 of this dissertation outlines the argument for architectures and models supporting the creation of trustable privacies in public spaces. Reviews of literature and contemporary proposals exposed many architectures, tools and modes of practice. Philosophical aspects of privacy and trust, including discussions on the nature of the proposed trustable architectures and models, have been addressed in Chapters 2, 3 and 4.

In this chapter, the following aspects are covered:

- Description of a model for encapsulation of data and security at an entity (atomic node) level which is wholly independent of server-side control and security considerations
- Situation of the proposed model in a variety of usage scenarios
- Testing of hypotheses and discussion on future work emerging from the dissertation.

Self-**C**ontained **R**epos**I**tory **P**roces**SI**ng **T**emplate (SCRIPSIT) describes a model intended to support trustable, resilient, persistent, peer-centric and serverless meshes or networks of encapsulated nodes. SCRIPSIT is built on existing standards including XML, RDF, P3P and portable scripting technology.

It must be noted that the choice of scripting environment is not limited to Java or its

derivatives. Time and resource limitations precluded detailed evaluations of the full range of scripting environments. For the purposes of this dissertation, the specifics of the scripting environment are not critical, hence a decision was made to choose a scripting platform supported across a variety of computing platforms. Java's platform and vendor ubiquity qualifies it on the basis of wide and competent technical and procedural support.

Published standards are used for the following reasons:

- Unless there are compelling technical or other reasons for doing so, there is little to be gained by creation of further standards or pseudo-standards. Use what there is, where feasible.
- Evolution of published and accepted standards feeds directly into SCRIPSIT in terms of extension of functionality and application. This applies at the external (exposed) level and also at the internal (functional) level.

SCRIPSIT aims to provide a basis for platform and environment independent trustable entities, able to persist and be useable on anything from a PC to a Java-enabled cellular phone to a Pocket PC. There are many possibilities in addition to those listed. More flexible modes of access are discussed under the scenarios later in this chapter. Architectural and processing problems are often solved using real world metaphors subsequently abstracted into models. Such conceptual models are then translated as implementations.

A real-world denial of possibility lead to the thinking which preceded SCRIPSIT's conception and formalisation. A limited set of scenarios formed the basis of questions posed in this dissertation and SCRIPSIT. The essential aspects of these initial scenarios were used to derive basic attributes and characteristics defining a model which could support the concept of a trustable privacy in a public online space.

Implementation and deployment of models are processes that take place in the reality and not in the virtual world. It is therefore essential that models, mechanisms and architectures proposed are described in situated frameworks. Some examples of these are described later in this chapter.

5.1.1 Constructing the Framework

An implementation framework is described to provide a basis for real world implementation of the SCRIPSIT model. The framework described is a simple implementation of the model.

The primary goal is a statement of the framework and a subset of mechanisms attached thereto. A basis for future research and implementation of this framework is expected as an additional outcome. Expected result is an extensible framework with multiple possible uses and variations in functionality and complexity.

Essential elements of the framework are:

- A self-contained repository entity architecture, described in terms of published standards.
- Resource concealment and exposure mechanisms.
- Mechanism used to create collections of SCRIPSIT entities.
- Encryption engine embedding and execution.
- Peer-local handling of decrypted data. The term *peer-local* is expanded further on in this chapter in Figure 9.

Motivation for creating SCRIPSIT is to provide a robust and practical option for the creation of trustable and secure peer-centric information collectives (or meshes). The terms *collective* and *mesh* are used to distinguish the result from the familiar and well-understood concepts of networks and webs. These terms are used interchangeably. In truth, the architectural outcome is not far from either, but a clear distinction is

necessary for purposes of clarity.

Peer-to-peer (P2P) models abound and are formally and soundly established. *Napster*, *Gnutella*, *Freenet* and *Edutella* are well-known examples of P2P networks. All rely to greater and lesser extents on serverside processing (even where the server is constituted on the peer node with which the communication is established) and on less lightweight and secure local applications which exist distinct from the data.

This chapter demonstrates the rationale behind the structure of SCRIPSIT entities. SCRIPSIT entities are subsequently presented in a linked or meshed view, as part of the description of the model itself.

This is followed with a series of application scenarios for SCRIPSIT. SCRIPSIT is, not uncoincidentally, Latin for '*he or she wrote*' – this is intended to indicate the user-centric nature of envisaged application areas. The user is intended to be owner, custodian, arbiter, censor, manipulator and mentor to their own data. Roles are expanded on in the scenarios outlined later in this chapter. Before outlining the proposed model, the concept of encapsulated peer entities is described in the context of the World Wide Web, independence at entity level and peer centricity/server independence are discussed.

5.1.2 Positioning SCRIPSIT in the context of the World Wide Web

At a conceptual level, the World Wide Web (WWW) is not dependent on any single set of servers or network paths. This very independence forms the basis of the resilience and near-indestructible attribute of the WWW. Within this decentralised and resilient model, it is possible to create client-server, peer-to-peer and any number of variations on these themes. Typical use of the WWW is in the context of a thin-client client-server model. There has been a dramatic surge in popularity of peer-to-peer (P2P) models in recent years, with the legally questionable sharing of online

music via *Napster* being one of the most publicly visible examples. Subsequent developments have included *Gnutella*, which is one of the purer P2P networking models to be made publicly available. Resources in a P2P model exist across many peer nodes in a network and are usually interrogated via an overlaid query mechanism.

The Internet was an association of computers, almost all of which both served data to others and requested data from the same. In other words, a resource sharing paradigm where contribution and consumption tended to balance out.

A significant consequence of the World Wide Web was the broadening of the data consumer base. Consequent to this has been a deviation from the load-balanced and resource-fair early Internet. A great reliance has been placed on serverside data storage, querying and script execution. Consider the back-end processing on almost any website – the client simply submits a page request to the server and all of the data processing functions occur on the server end. A necessary flaw in the public access client-server model is that the privacy and security of served data is specifically compromised at one stage or another in the process, even where a secure (HTTPS) link is used.

Even when considering a secure (HTTPS or SSL) link, the flaw lies in the fact that data are served and, in an ultimately decrypted form, sent across a hopefully secure link to requesting clients.

Where trustable privacies are considered, the fact is that unencrypted data flow across links (encrypted or otherwise) raises the level of perceived risk on the part of the data owner. Asking the bank to open your sealed envelope containing the PIN number for your bank card and then asking an allegedly a trustworthy messenger to bring it to you rather negates the point of the sealed envelope. In much the same manner, the serverside processing of meaningful data is unhelpful where perceived risk and hence trustability are concerned.

The data owner's concern does not lie with the messenger, but with the broken seal on the envelope. Breaking the seal dramatically raises the data owner's perception of risk associated with the action. Raising perceived risk has the corollary of reducing the user's ability to trust the mechanism.

The client-server model asserts resources existing on servers (or server-referenced repositories). In order for a requester to gain access to these resources, the client (or requester) requests resources from or through the server. Peer-to-peer models assume that all nodes in the network are able to be both client and server. In describing the basic architecture of *Freenet*, Clarke, Sandberg, Wiley and Hong (1999) argue strongly in favour of equally capable and authorised nodes in a P2P network.

SCRIPSIT neither knows about, nor is affected by, the particular network topology on top of which its entities and collections of entities reside. In the SCRIPSIT model, a client only ever serves unwrapped or decrypted entity content to itself. Any inter-peer traffic is at the level of an unintelligent transfer of an entity. Zhao *et al.*(2001) propose overlaid mechanisms for fault-tolerant routing and resource discovery. A self-organising and tolerant model built with existing web components and technologies offers an open source alternative to proprietary architectures.

SCRIPSIT's mission imperative is to make available and accessible securely encrypted packaged entities which bring the entire processing operation back to the client (the requesting agent). The accreditation and decryption process occurs entirely on the client side. A complete and untampered entity is served by whatever server node on the underlying network responded to the request. The ability of a node in a P2P network to be self-descriptive, via metadata, is crucial to its ability to be found and used appropriately.

Distinguishing data and metadata is problematic in some instances. Metadata may be used to select entities when discovery is in progress; the data within an entity may be

used refererentially by the SCRIPSIT engine when navigating an entity collection. In this instance, the data and metadata are difficult to distinguish. An example of a problematic distinction may be classification of the originating research institution. This may be data, as it is aggregated with the actual research data and this constitutes a definable part of the whole. It may also be interpreted as metadata as it describes the institution responsible for the research referred to in the entity or entity collection. Megginson (1999) comments that the distinction is created primarily by the application of the data.

It is axiomatic that metadata may have metadata. Taken a step further, metadata metadata provide a basis for alternative and richer ways in which metadata may be discovered. Berners-Lee (1999) notes that "*metadata is found when it is looked up in another document*". This statement succinctly describes the self-descriptive character of metadata and, consequently, of the Semantic Web and of the fundamentals of resource discovery within loose associations of metadata-enhanced data nodes, such as would be found in collections of SCRIPSIT entities. Taking a working definition of resource[vii] to be "anything that has identity", SCRIPSIT is required to be able to accommodate constant identity with changing context and location.

SCRIPSIT asserts a requirement for combining data and service provision in the entity structure and function. The only services which are exposed (i.e. are external to an entity) are discovery services and literal serving of complete SCRIPSIT entities. Ahlborn, Nejdl and Siberski (2002) propose a P2P open archive model (OAI-P2P) which stresses the need to separate data and service provision in order to simplify the model and its implementation. Ahlborn, Nejdl and Siberski (*op.cit.*) crucially introduce a limited hierarchy with the provision for provider peers to aid in the discovery process. SCRIPSIT embeds resource descriptor information and pointers at

[vii] Examples include documents, images, services (news reports, weather information) Not all resources are retrievable (people, institutions and printed papers) across the WWW. The resource is the conceptual mapping to an entity or set of entities, and not necessarily the entity corresponding to the mapping at a specific time. A resource can therefore remain constant whether or not the entities to which it corresponds change over time. This is predicated on conceptual mapping remaining constant.

both the exposed and concealed levels of the entity.

It is the intention that a SCRIPSIT entity may be discovered by almost any search-capable WWW tool. Much of the ease of discovery depends upon what is exposed and on how (if at all) entities are indexed or referenced by external agencies. Nejdl *et al.* (2000) comment that "metadata are useful and important, for Peer-to-Peer (P2P) environments metadata are absolutely crucial".

The lack of simple navigability of P2P information networks is highlighted by Nejdl *et al.*(2000), and underpins the argument presented in favour of externally supported and mediated search and discovery mechanisms. SCRIPSIT relies on two classes of metadata. The first is metadata which is unencrypted and publicly visible and accessible. This is essential for primary access to a SCRIPSIT collection – if data cannot be found, then data cannot be used. The second class of metadata is that which exists on an encrypted level within a SCRIPSIT entity. Purposes of the encrypted metadata are specific to the nature and application of the data carried within the entities and entity collections being addressed.

Without metadata, P2P networks are unable to function at any level beyond that of random connections and messaging. Without metadata contained in RDF triples SCRIPSIT is little more than a novel variation on an encrypted data element, existing within an arbitrarily selected network environment.

The primary goal of this research is the presentation of a model which not only includes appropriate metadata and support mechanisms at network node (or entity) level, but which encapsulates security and accreditation support at the same time.

The WWW is an inherently untrustworthy place. In this vein, Metzger (2004) notes that trust is

> *...the degree to which an organisation is perceived to be reliable, competent, benevolent, and to have integrity.*
>
> (Metzger, 2004)

The same attributes and metrics referred to by Metzger (*op.cit.*) apply to all models and mechanisms which are asserted as being trustable. The WWW is reliable in that it is resilient and persistent. Reliability at the level of individual messages and transactions is not at all guaranteed. For perceived and realised reliability to be possible, a model or mechanism must acknowledge and use the existing architecture and characteristics of the WWW.

Individual and institutional benevolence, integrity and competence are applicable at the level of any overlaid applications and mechanisms in the WWW. SCRIPSIT removes the requirement for dependence on organisational benevolence and integrity through its peer-centricity and independence from institutional support for preservation of integrity and its resilient nature. SCRIPSIT is an integrity enabler.

Organisations perceive a lack of control to be threatening and untenable for organisational survival, perceived or real as this lack of control may be. Consequential to this is the organisational imperative to regulate, to police and ultimately to interfere in the unfettered exchange of information. On a governmental level, this is realised in the form of imposition of ideological frameworks on repositories and data exchanges. It is inevitable that ideologies shift and that what was acceptable or protected under one ideological environment is no longer acceptable under another. The ability for the vulnerable individual to determine levels of acceptable risk in, and hence the inherent trustability of, data storage facilities is therefore severely compromised in conventional P2P solutions.

The concept of trusted third parties is flawed when issues of ideological shifts come into play. On a personal level, there is only one trustable party, and that party is self. Given this cynical view of the risk of trusting on an organisational level, it is essential that the control of risks associated with storage of confidential data is placed directly in the hands of the data owner.

An understandable if limiting attribute of P2P network implementations tends to be that of field or domain specificity. This is largely with respect to the model's zone of effectiveness. This zone of effectiveness may be expressed as the potential breadth of effective applicability of the architecture achievable without substantial evolution or modification of the original architecture. One example of a mutated model is *Edutella* (Edutella, 2003), a development of *Gnutella* (Gnutella, undated). The narrower the zone of effectiveness, the greater the tendency to lean towards a plethora of P2P models and variations on these models, each one suited to a narrow domain, a limited zone of effectiveness.

Whilst accepting the statement that there is no universally complete and suitable architecture, the search must be for a model which allows both great flexibility of application and infrastructural portability.

The philosophical core of SCRIPSIT is built around the concept of trustable privacies. There is a layered pyramid which starts with the actual resources at the URI level and is capped by the intangible stone of privacy and trust. It is not the focus of this chapter to concentrate on definitions of privacy and trust; suffice to say that building a solid foundation for trustable privacies requires a number of areas to be addressed.

Figure 4 illustrates the layers in the trust pyramid. For a document to be trustable, the requirement exists for it to be self-describing and able to withstand scrutiny according to archival criteria.

In other words, the following questions apply:

- Is the document inviolate? Can the discoverer be certain that the document returned is as originally posted?
- Is the document universally available (if not universally readable without appropriate accreditation)?
- Is the document originator/source verifiable? How do I verify that the document in front of me is the same document posted by myself?

Starting at the URI and XML layers, no attributable trust elements are apparent. There is no verifiability other than the literal face value of the structures and references visible to the discoverer. Moving onto the metadata and ontological layers, a degree of relevance and resource verification starts to become apparent. There is still no strong and direct indicator of reason to trust the document 'as discovered'. Arriving at the policy and accreditation layers, the option for the discoverer to attribute reduced or no risk to the authenticity of the document allows a degree of trust to be established.

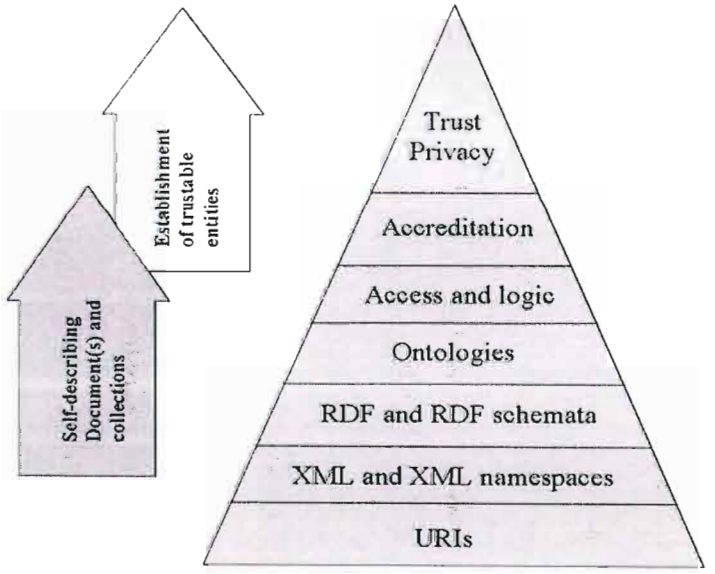


Figure 4 - Building logical layers of privacy and trust support

It is apparent from Figure 4 that the self-describing nature of the document becomes more complete the higher the examined layer in the model pyramid. And hence the closer the document moves to being able to build and claim trust in the perceptions of the discoverer. Note that the discoverer is any user who may find the document, ranging from the originator to archivist (if any), accredited user and ending with the least accredited discoverer, the casual browser.

5.1.3 Independence at entity level

SCRIPSIT independence exists at a unitary level. This is a logical independence simply because an entity requires a host on which the self-contained engine may execute.

In the SCRIPSIT domain, the host is always the client which requested the entity. SCRIPSIT itself is an equal peer P2P architecture. Traditional servers have no specific place or function in this environment other than as unintelligent repositories or staging posts for transitory SCRIPSIT nodes.

Any network nodes identified as servers function purely as holding points from where SCRIPSIT entities may be requested. No intelligence is attributed or allowed to the serving host in respect of entity processing.

5.1.4 Peer-centricity and server independence

SCRIPSIT's architecture is peer-centric. An important distinction is drawn between peer-to-peer (P2P) and peer centricity. The function of serving requested SCRIPSIT entities is essentially supported by any appropriate underlying P2P arrangement. Considered in isolation, SCRIPSIT entities may exist within non-P2P environments with degraded ability for resource discovery and URI-following functionality.

SCRIPSIT entities, in contrast, are all equal in ability, in security options, and in mobility. These entities demonstrate many of the attributes of nodes in an equal peer P2P network, without being bound by topological and architectural dogma. Server independence is supported by simple diminution of the role of the (network) server to unintelligent staging area.

5.2 Outline of proposed model

A high level outline of SCRIPSIT, the required functionality, and the amalgam of published standards is described as the first part of this section A SCRIPSIT entity is built up as a compound element from XML, P3P, RDF and secure script engine components.

Compound entities of arbitrary complexity may be built up from any number of SCRIPSIT entities. These arbitrarily complex structures serve the purpose of building complexes of SCRIPSIT entities. SCRIPSIT is a self-contained repository entity architecture, described in terms of existing published standards.

5.2.1 Standards and components

Standards used and referred to are described briefly in this section. The reader is directed to the relevant published standards websites for further detail and expansion. P3P, RDF and XML are described in the Chapter 2 (Literature Review).

- P3P Platform for Privacy Protection.
- RDF Resource Description Framework.
- XML eXtensible Markup Language

Additional notes on the Semantic Web[viii] and OWL (Web Ontology Language) are also to be found in Chapter 2.

Further to formally described standards are references the embedded engine and link models used by SCRIPSIT are the embedded script engine and links to other SCRIPSIT elements:

- ENGINE The embedded script engine key to SCRIPSIT's peer-centric model
- links Links to any resource, usually a SCRIPSIT entity but may be any legitimately referenceable resource. These may be logical (at a metadata level) and/or literal (at URI level).

5.2.1.1 Encapsulated/embedded engine (SCRIPSIT)

The encapsulated engine (s) present in every SCRIPSIT entity and wholly responsible for formal and trustable processing of all requests and presentation of unwrapped data to the user or his/her agent applications.

5.2.1.2 Resource linking

Links are Universal Resource Information (URI) pointers to any resource internal to or external to the SCRIPSIT entity. A distinction is drawn between *Plain Text Links* and *Secured Links*. *Plain Text Links* are as described, a link which is not encrypted nor hidden from direct viewing in any application capable of displaying the raw contents of a SCRIPSIT entity. *Secured Links* are those URIs which are contained within an encrypted portion of a SCRIPSIT entity and which therefore require

[viii] Semantic Web and OWL references are not expanded in this chapter as they are not fundamental to the structure of the model.

permission to be viewed or used as navigational links. Both variants are literal links as they point directly to a resource or resource location. Implied resource links exist at the metadata level, at both exposed and concealed levels. Implied linking of resources is possible at a metadata level. Debate as to whether or not this is to be included in the embedded functionality of the SCRIPSIT engine is an area for future research, not covered in this dissertation. It is arguably a function which may exist outside of the entity (partly or completely) and hence require exposed metadata to make this possible.

5.2.2 SCRIPSIT described at entity and collection levels

The simplest structural description of a SCRIPSIT entity is:

- An XML document core, limited to plain text and Base64 data, which is enhanced through metadata tagging and embedded URI data. This corresponds to a 'well formed document' at the RDF syntactic and structural levels. This document core also contains concealed (except from accredited users requesting access) P3P assertions and concealed URI data.
- Wrapped around this is a publicly-exposed layer containing P3P-derived access and accreditation assertions, and RDF content. This is the publicly visible face of a SCRIPIT entity and, as such, is the only entry point into any SCRIPSIT collection of entities.
- The outermost wrapper of a SCRIPSIT entity is the (optional) HTML layer, allowing presentation of publicly-visible parts of an entity as a simple Web page. This layer is not required if the entity is a proper well-formed XML document and is compatible with appropriate resource discovery tools.
- Embedded in the well-formed XML document is an encrypted engine which provides the key to concealed data, metadata, access mediation mechanisms and URIs within the document.

A SCRIPSIT entity, simply stated, is an encapsulated and layered assembly of permissions and accreditations, semantically explicit resource descriptors, tagged and structured documents and embedded security/access engine. A limited example of an expansion of the SCRIPSIT entity structures in this chapter is shown in Appendix C.

All of the preceding exists within the confines of an HTML document or well-formed XML document requiring no more than an unsecured HTTP link for all operations legitimately allowed to requestors and custodians. A collection of SCRIPSIT entities may consist of one or both of:

- Compound entities which are two or more SCRIPSIT entities wrapped by a single SCRIPSIT entity.
- Linked entities which are two or more SCRIPSIT entities linked either by URI or by metadata. The URI link is self-explanatory. Metadata links are logical links only, where metadata existing in one entity are used to facilitate discovery of one or more other SCRIPSIT entities. The facility for a properly diasporic collection of related SCRIPSIT nodes exists as a consequence of the metadata logical links. Metadata links are explained diagrammatically in Appendix B with reference to edge-directed RDF graphs. See Figure 20 in Appendix B.

5.2.3 Simple SCRIPSIT entity

A simple SCRIPSIT entity has a plain HTML outer wrapper as its basis. Unencrypted P3P-based accreditation fragments provide a public confirmation of access and simple trust-based assertion of right to access and propagate by the discovering agent(s). In a similar manner, public assertion of RDF metadata at this unencrypted level aids discovery and discriminatory selection of SCRIPSIT entities.

The XML document embedded within the HTML wrapper constitutes the functional

core of the SCRIPSIT entity and is expanded in the subsection following. Conceptual rooting of the SCRIPSIT entity in unique data triples (see B.2 in Appendix B) is illustrated in Figure 5 following:

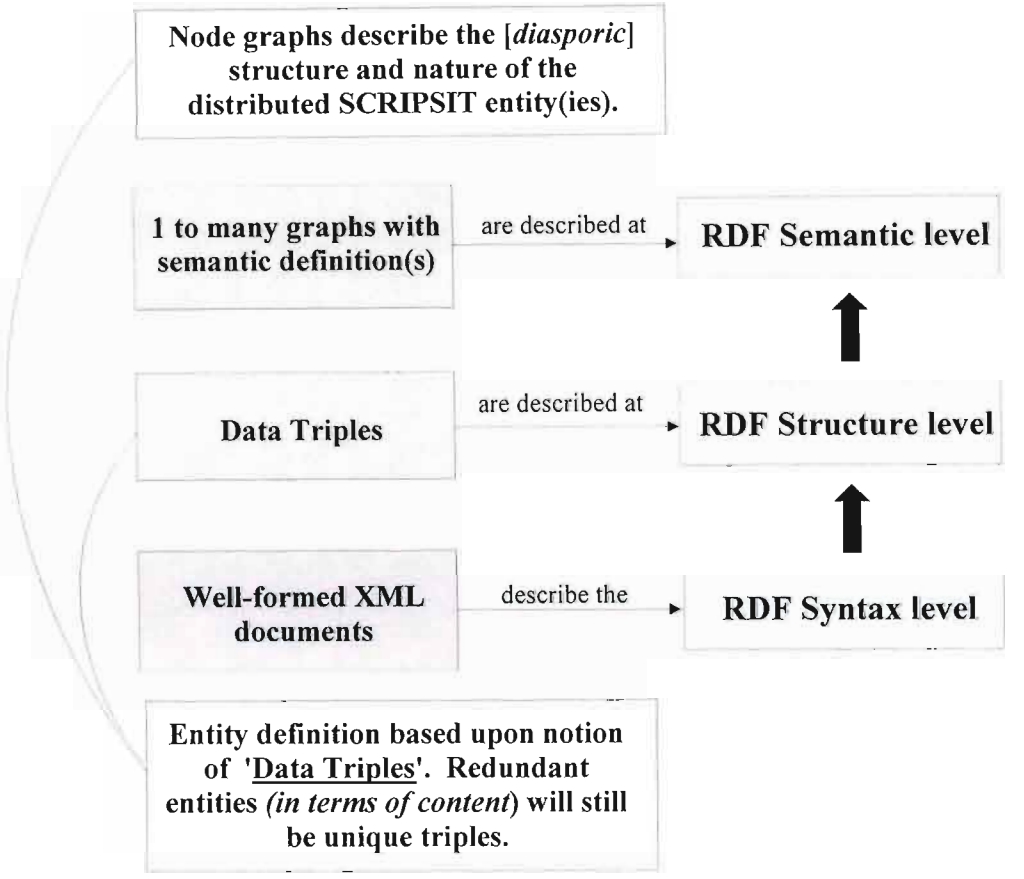


Figure 5- Conceptual basis of SCRIPSIT entity definition

5.2.3.1 Structure, function and instantiation

The SCRIPSIT entity structure is straightforward and is described diagrammatically in Figure 6 below. The HTML outer is the completely exposed and inert base which carries accreditation and permission elements (P3P-based), resource description and location attributes (RDF) and actual SCRIPSIT document structure (XML) and content. The embedded SCRIPSIT engine exists within the confines of the XML document. The XML document data types used are limited to String and Base64 (*XML, undated*) for the purposes of this initial SCRIPSIT proposal. It is not the

intention at this point to expand upon or argue the case for extension or adaptation of XML data types.

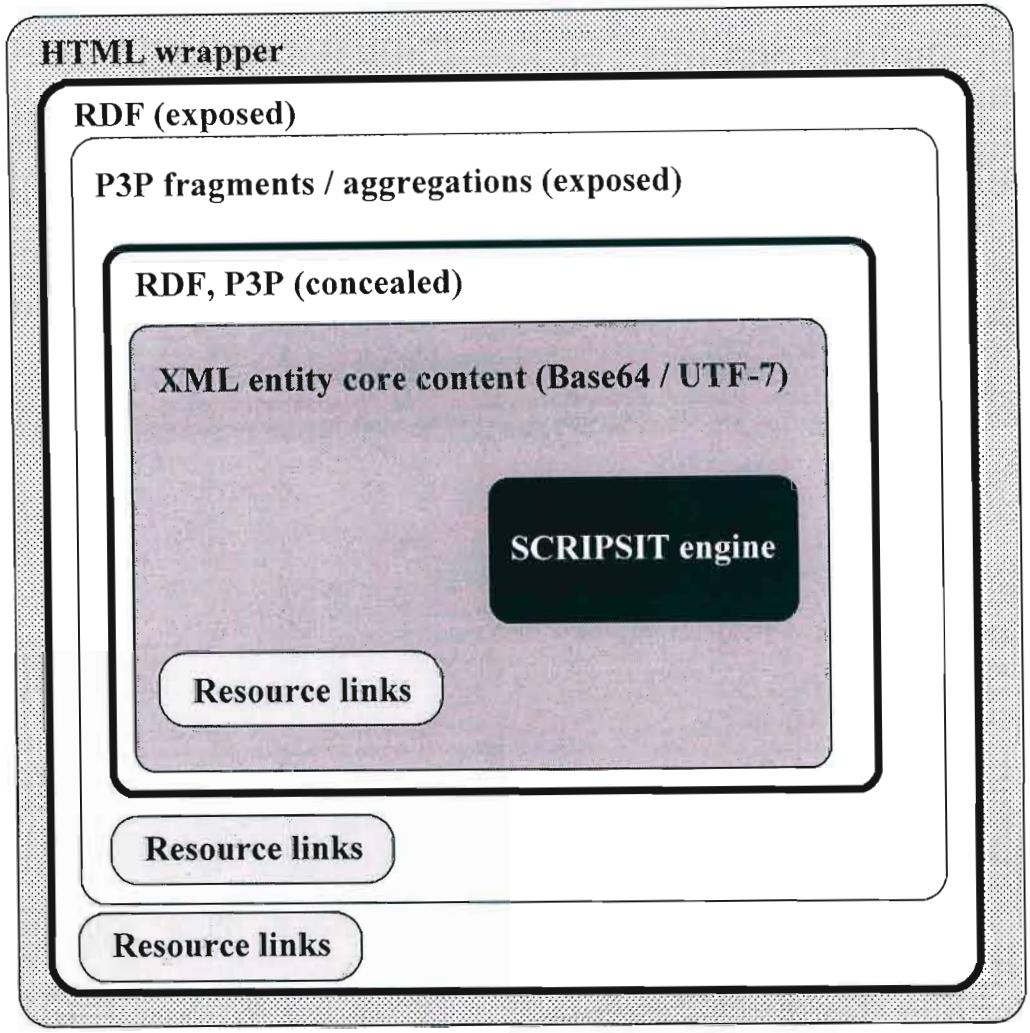


Figure 6 - SCRIPSIT entity structure

A SCRIPSIT entity is only able to function upon instantiation of a virtual machine (VM) on the requestor. The embedded engine referred to in the entity structure description requires a host on the requestor (a simple example would be a Java-enabling browser plugin or similar) in order to examine accreditation profiles, assemble or solicit decryption keys, display entity contents or perform any other processing. The implied hierarchical layering of RDF within the P3P accreditation layer above is not absolute. Both the P3P and RDF layers are exposed and are

therefore valid resource discovery routes. It is conceivable that an entity may be discovered on the basis of data content fit (as described in exposed metadata) or on the basis of accreditation fit (as described in exposed P3P accreditation fragments). A mixture of both discovery options is not precluded. Moving into the encrypted contents of the XML core document in an entity, all of the preceding points are repeated, albeit with restrictions applicable to the level(s) of accreditation existing on the requestor host.

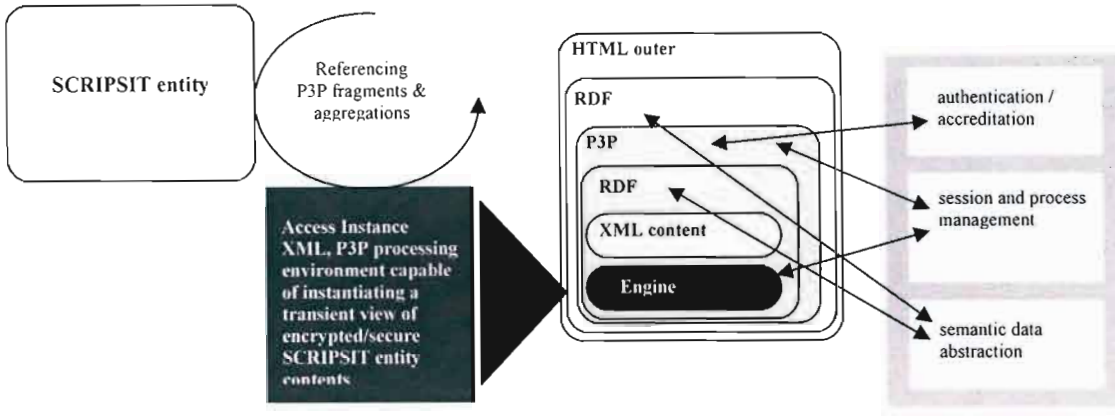


Figure 7 - SCRIPSIT processing and functional instantiation

The access instance referred to in Figure 7 is primarily concerned with being able to generate or access keys for the encrypted data contained in the XML entity core document. Along with this, the facility exists to manage the display of session-relevant data based on asserted accreditation fragments and aggregations present at a P3P level. Further to this, semantic data mapping and abstraction is the primary responsibility of the embedded engine.

A more general expansion of a SCRIPSIT access instantiation is shown in Figure 8 following:

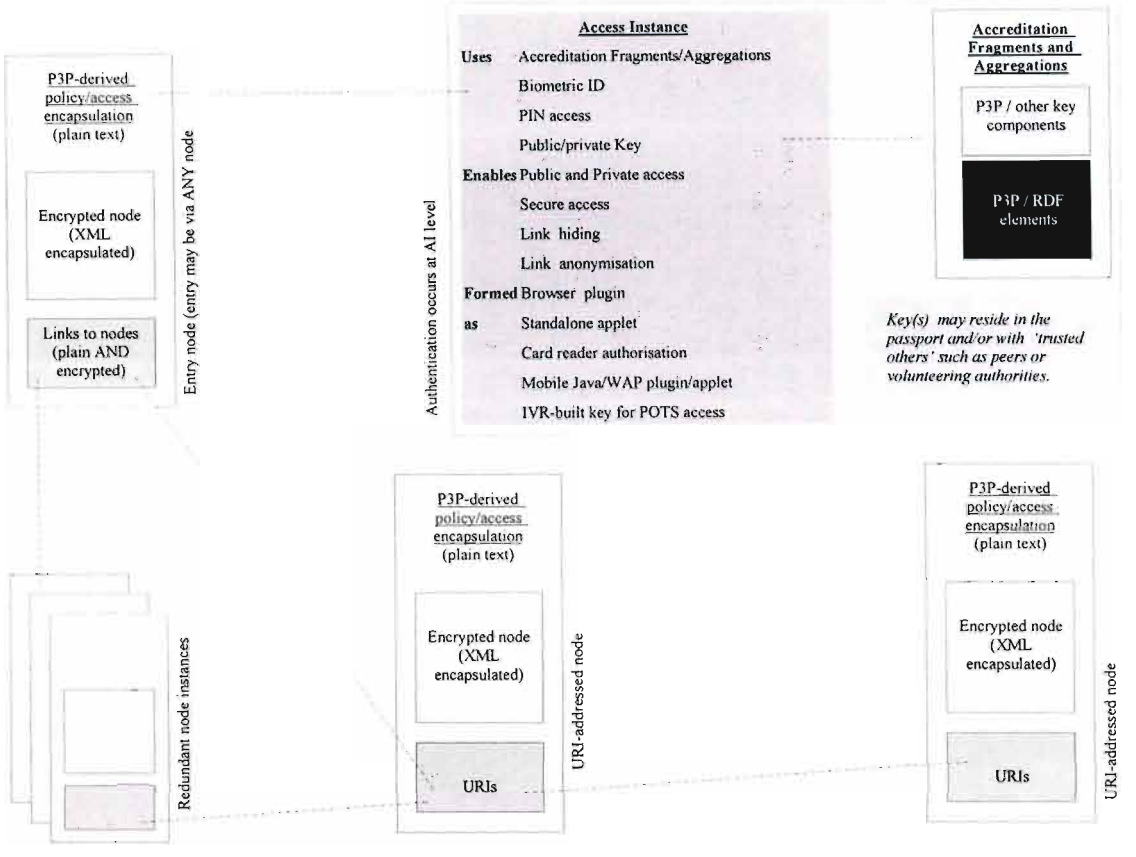


Figure 8 - Generalised attributes of SCRIPSIT access instantiation

Access instantiation via the office of the embedded engine is the only legitimate route available for SCRIPSIT entity access.

5.2.3.2 Communication and security for simple and compound entities

A partial handshake sequence is described in Figure 9:

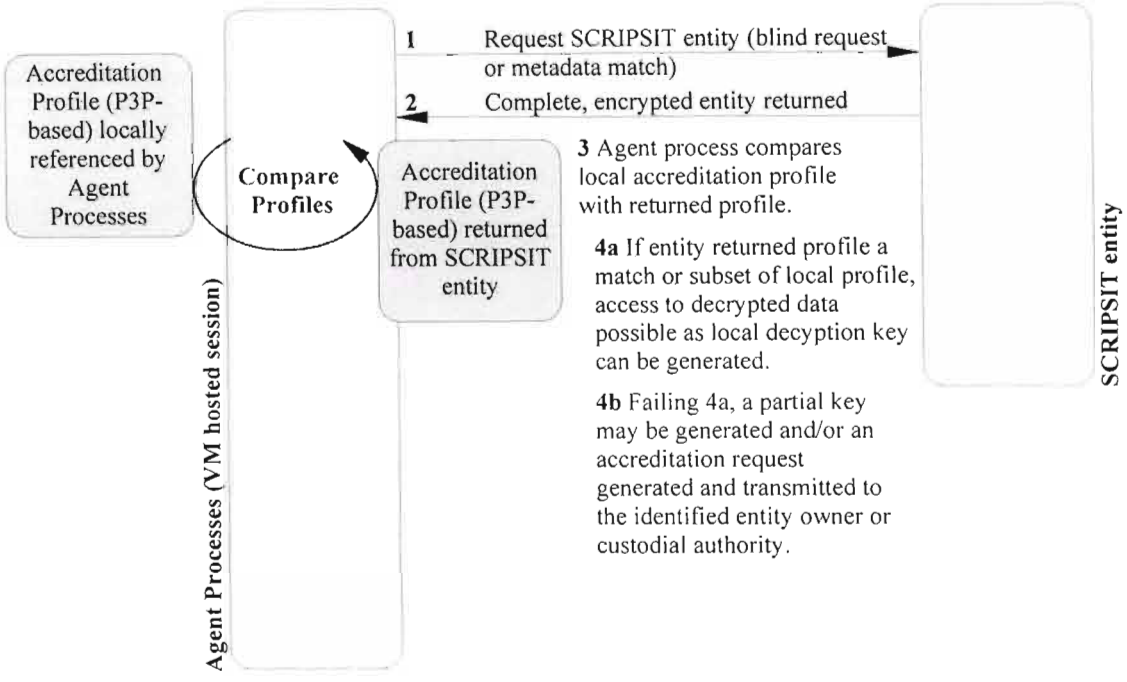


Figure 9 - Peer-local level comparing entity and hosted fragment profiles

The SCRIPSIT entity, requested by the Agent Process (see Figure 9Figure 9) in step 1, is returned in step 2. The P3P-based accreditation profile returned with the SCRIPSIT entity is compared with locally-referenced accreditation profiles. If there is a successful match (steps 3 and 4a) then a local decryption key can be generated and used to access the encrypted contents of the SCRIPSIT entity. Where this is not achieved (step 4b), a partial key is generated and the option of generating an accreditation request is opened. This request may be transmitted directly or via a custodial authority to the entity's data owner.

5.2.4 Compound and linked SCRIPSIT collections

Moving a step beyond simple SCRIPSIT entities, there are two varieties of SCRIPSIT aggregations:

- Compound SCRIPSIT entities – consist of two or more fully-formed SCRIPSIT entities wrapped in SCRIPSIT-compliant P3P and RDF outer shells. The outer shells define the outermost layers of accreditation requirement and discovery.
- Collections of SCRIPSIT entities – these consist of two or more fully-formed SCRIPSIT entities (which may be simple or compound) and are associated via any number of embedded links. A collection may also be defined purely through common discovery-enabling elements, avoiding the (assumed) requirement that a network or collection of related entities explicitly reference one another.

All have plain HTML outer wrappers, or may be discovered within a more complex XML document. Unencrypted P3P-based accreditation fragments provide public confirmation of access and simple trust-based assertion of right to access and propagate by the discovering agent(s). In a similar manner, public assertion of RDF metadata at this unencrypted level aids discovery and discriminatory selection of SCRIPSIT entities. Alternate discovery routes apply equally to compound entities and entity collections as described for simple entities.

5.2.4.1 Structure of compound entity

SCRIPSIT compound entity structure is a little more complex than the simple case and is described diagrammatically below. The compound entity's HTML outer is the completely exposed and inert base which carries accreditation and permission elements (P3P-based) applying to the whole publicly visible part of the compound entity. Also at this level are the resource description elements. Below this are found the SCRIPSIT entities making up the compound element itself. Note that the encapsulated SCRIPSIT entities may legitimately enjoy differing levels of exposure

and concealment. They may also provide links to each other within the compound SCRIPSIT entity. This is illustrated in Figure 10 following:

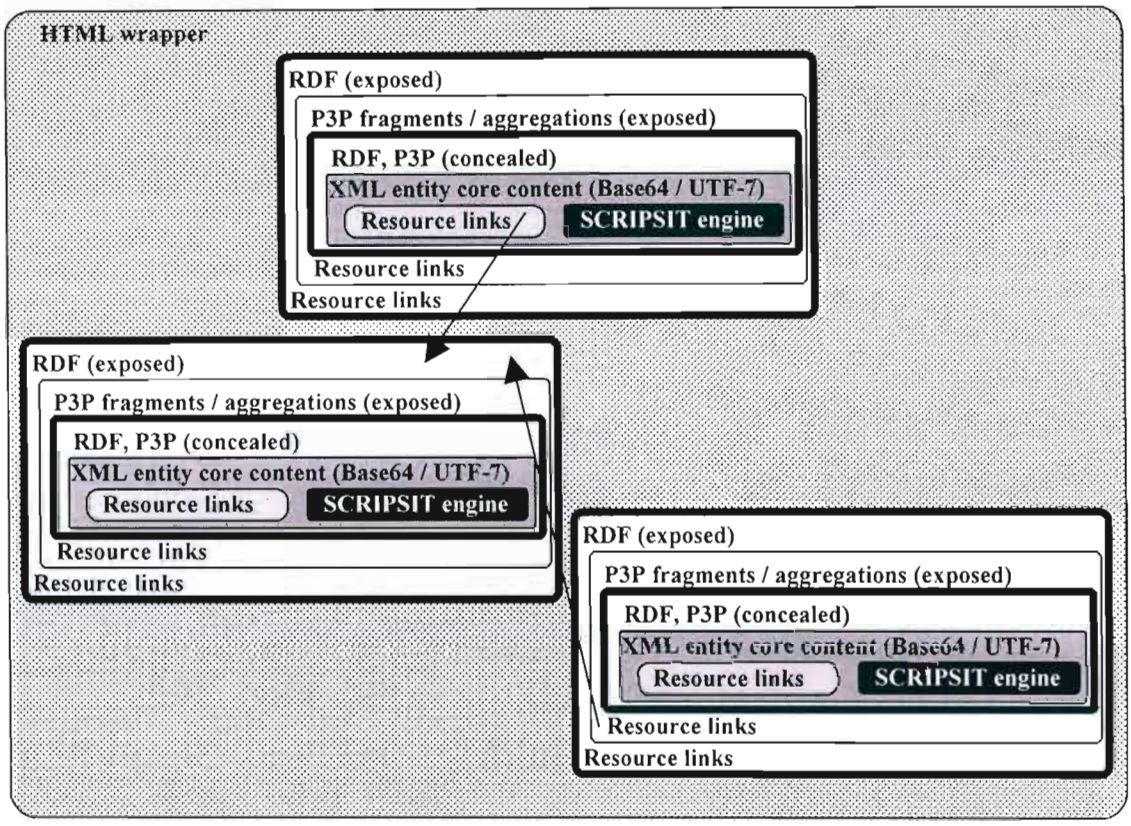


Figure 10 - Simple example of a compound SCRIPSIT entity

5.2.4.2 Structure of entity collection

A given SCRIPSIT entity collection may be structured in any manner which conforms to the basic requirements of the WWW. There are no limits or restrictions on the number or types of links included, save those imposed by the WWW itself. The following illustrates a simple SCRIPSIT entity collection. Figure 11 illustrates the structure of a compound entity.

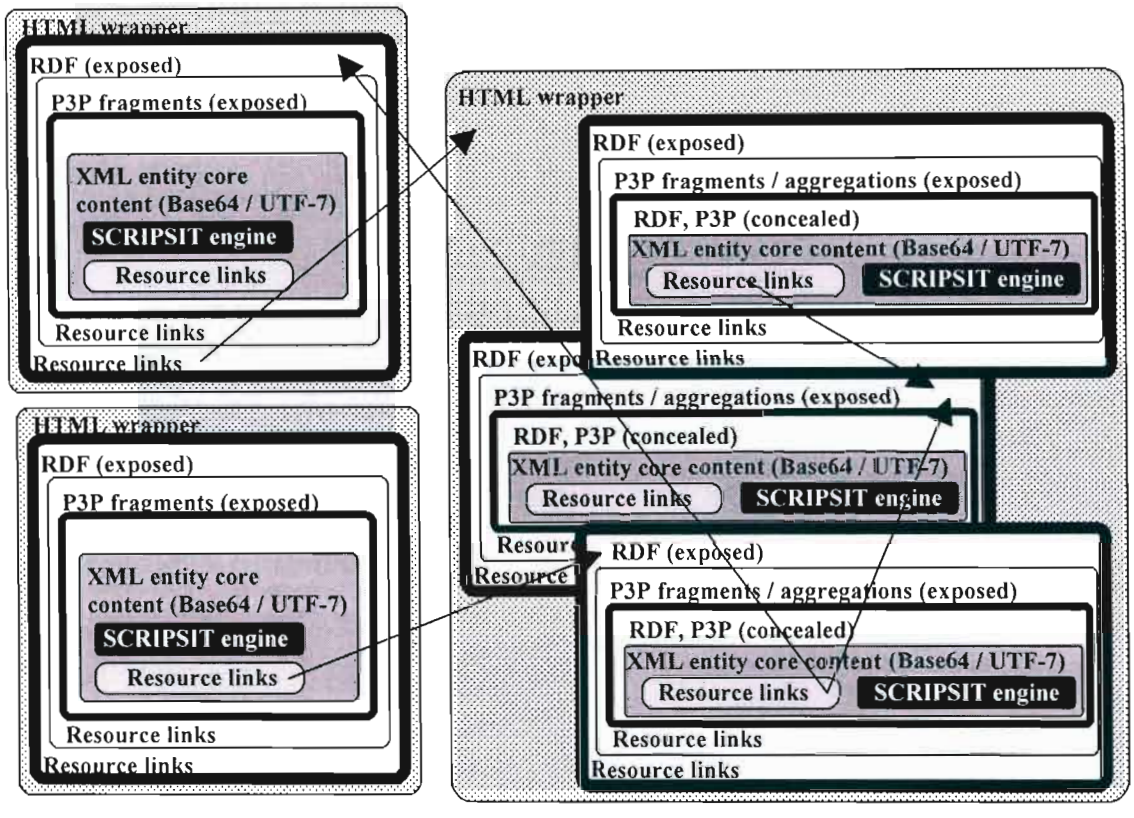


Figure 11 - Example of a SCRIPSIT entity collection

5.2.4.3 Function and instantiation

Refer section 5.2.3.2. Requirements are as for a single SCRIPSIT entity.

5.2.5 Topological and architectural considerations

Peer-to-peer models tend to be best suited to specific varieties of application. The major architectural variants and typical uses are briefly described here, and SCRIPSIT's ability to be used in the given context is addressed.

5.2.5.1 Centralised access

Of the many flavours of P2P networking, some are only loosely P2P in nature and rely on a single point of routing and access. Distributed computing is an instance of this where the peer aspect is realised as a large, dynamically accessed array of voluntary computing resources. A central server parcels out processing work and the peers have no actual communication with one another, save for that routed or facilitated by the server. All routing and allocation is via the hub. Here a high degree of control exists, although the central point remains a vulnerable single point of failure. SCRIPSIT entities have no functional or data restriction with respect to existing in such a context.

5.2.5.2 3rd Party routing (or brokering) access

There may be one or more servers which are tasked with managing the peer links. Peers rely on having to identify servers and themselves in order to successfully link to other peers. Once P2P links are established, servers are functionally disengaged and no longer required by the peers. SCRIPSIT entities may successfully make use of such a link brokering service without compromising their ability to be found (RDF as an element of a discovery mechanism for the entities).

5.2.5.3 Decentralised and Equal Peer access

Decentralised and equal peer access is defined by a serverless environment where peers actively engage directly with each other. The only server involvement in this instance is as pure entity routing points. *Edutella* (a conceptual development of *Gnutella*) is one example of a decentralised access P2P network. The notion of equal peers is a step beyond a decentralised P2P architecture. In this instance, any node (or SCRIPSIT entity, in this case) may be used as a point of entry to a particular SCRIPSIT collection.

5.2.5.4 Hybrid access (Super Peer)

Mixing the attributes of 3rd party routing and decentralised equal peer architectures results in an architecture which allows some SCRIPSIT entities to become super nodes through which a smaller SCRIPSIT collection may be accessed. Without the super node, the subordinate SCRIPSIT entities within the collection will be almost or completely impossible to find via most resource discovery mechanisms. This is subverting somewhat the pure intent and application of SCRIPSIT but remains a valid option.

5.2.6 SCRIPSIT accreditation mechanism

Three basic flavours of SCRIPSIT accreditation request mechanism are outlined sections 5.2.6.1 through 5.2.6.3. These are the simple accreditation case (5.2.6.1) where the requesting agent possesses at least the minimum required accreditation authority needed to generate or acquire decryption keys required. Following on from the simple case is the accreditation case where the requesting agent does not have sufficient accreditation (5.2.6.1).

In this case, a request for additional accreditation (via appropriate P3P profile fragments – for both the request and the accreditation) is generated and routed to the data owner or issuing authority identified by the requesting agent. The last case presented is the passive/active accreditation issuing and retraction (5.2.5.3). A previously valid level of accreditation may prove invalid on subsequent access by a requesting agent. In this instance, the data owner/custodian has amended the accreditation profile and/or permissions encapsulated in the SCRIPSIT entity. This is in effect an extension of the case presented in 5.2.6.2.

5.2.6.1 Simple accreditation verification case

The elementary case presented for accreditation is the static instance. A SCRIPSIT entity or agent places a request (Step 1.) with a SCRIPSIT serving instance. A copy of the requested entity is returned (Step 2.). P3P accreditation fragments (P3P-based profiles) and collections are checked against each other and, in this instance, sufficient accreditation is established. The required decryption keys are generated by code executed on the local virtual machine (VM) and then the decrypted data are pushed to the presentation layer (Step 3.), managed by the VM engine. This is an uncached, transient instance of the data with the aim of preserving the trustability of the entity's encrypted content. Figure 12 illustrates the process.

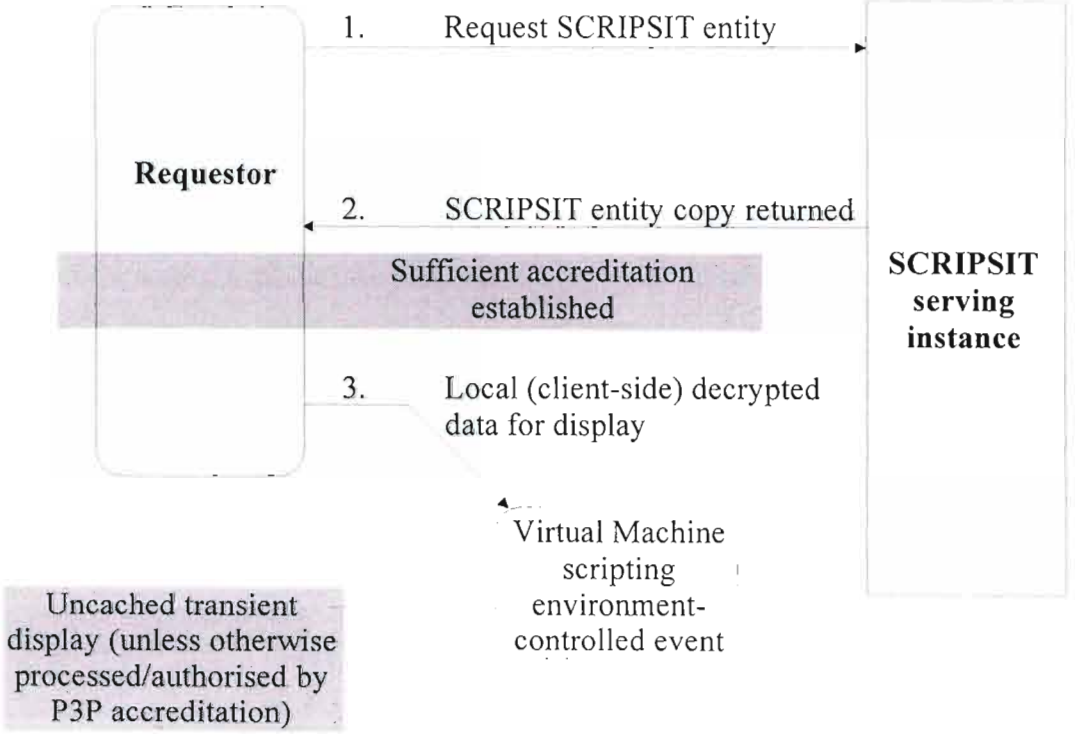


Figure 12 - Sufficient accreditation available to process SCRIPSIT entity

5.2.6.2 Accreditation request case with optional outcomes

The simplest accreditation case presented is the static instance. A SCRIPSIT entity or agent places a request (Step 1.) with a SCRIPSIT serving instance. A copy of the requested entity is returned (Step 2.). P3P accreditation fragments (P3P-based profiles) and collections are checked against each other and a state of insufficient accreditation is established. Here (Step 3.), a P3P fragment containing an accreditation request is generated by the requesting agent and passed to the issuing authority for the requested entity. This authority may be an automated agent or a human agent, and may do one of the following (Step 4.):

- Issue a P3P accreditation fragment (Step 4a.), actively routed to the requesting agent
- Issue an explicit P3P denial (Step 4b.), actively routed to the requesting agent
- Remain inert and issue no response (Step 4c.)

The response (Steps 4a & b.) may be returned to the requestor via any route or mechanism able to convey a P3P-based accreditation fragment back to the requestor. Refer to Figure 13 following for a diagrammatic expansion.

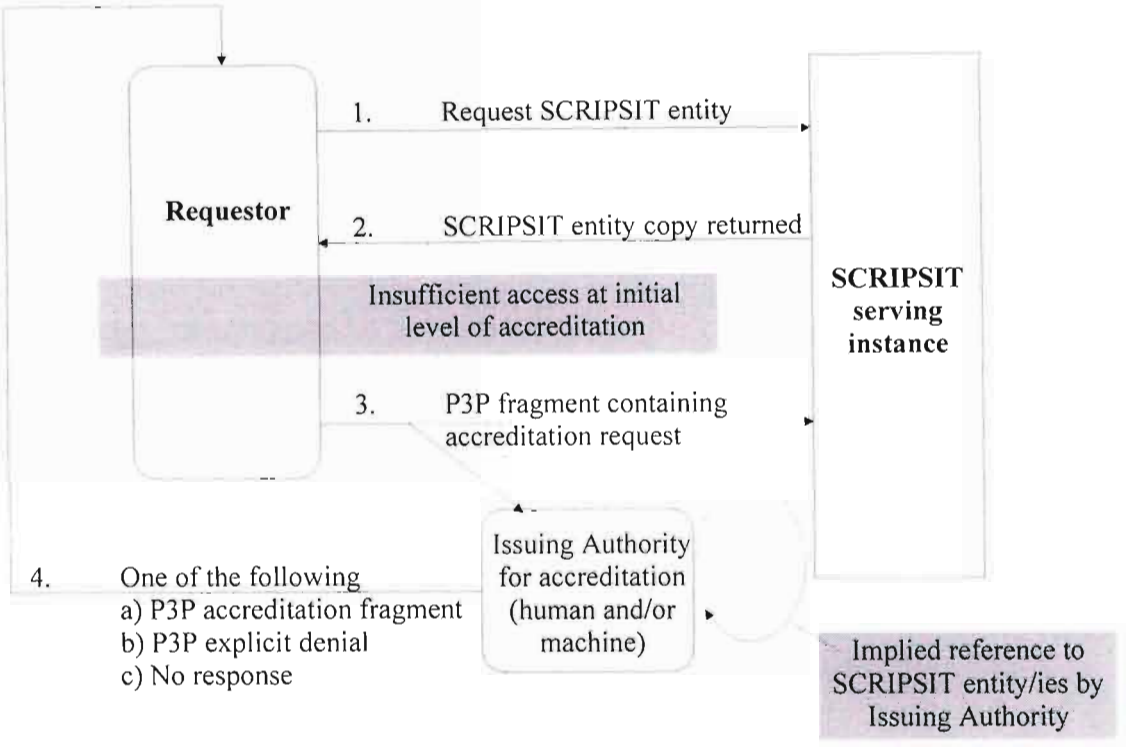


Figure 13 - Accreditation request with optional outcomes

5.2.6..3 Passive and active forms of accreditation issue and retraction

A SCRIPSIT entity is requested in Step 1 in Figure 14 by an agent previously possessing adequate accreditation. In the time between the last request event and the current request, the accreditation profile of the entity has been changed by the entity data owner/custodian. The returned SCRIPSIT entity brings P3P fragments which modify accreditation profiles on the requestor. At Step 3, the requestor may issue a request for rescinding or reinstating of the new or superseded profiles.

Step 4a may result in an explicit granting/pending/refused response from the issuing authority. A passive result (Step 4b.) is obtained where the issuing authority updates P3P accreditation fragments included with the originally requested entity. These are cases of passive and active issuing and retraction of accreditation.

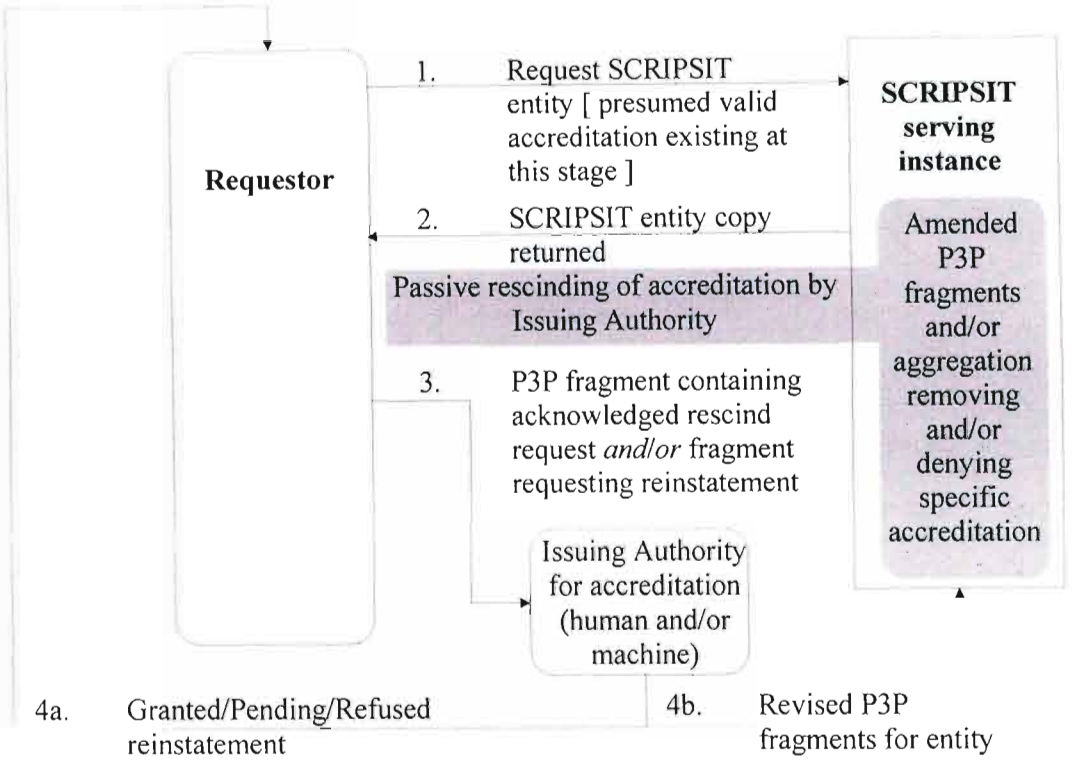


Figure 14 - Passive and active accreditation issue and retraction

5.3 Resource concealment and exposure

Resource concealment and exposure mechanisms are the composite of SCRIPSIT's P3P-based accreditation/token mechanism and the embedded encryption engine. Exposure is extant at the level of publicly discoverable information at the P3P level (in terms of accreditation), at the RDF level (in terms of discovery of relevant resources and at the unencrypted XML content level.

Concealment in SCRIPSIT terms is simply the negation of all or any part of the exposure options. At the very minimum, a SCRIPSIT entity will identify itself as such and expose no additional information to public view.

A crude process of wide, shallow resource (or entity) discovery constitutes the most elementary level of discovery and access. More realistically, a minimum level of

discovery data (RDF) is expected to be visible and useable as basic relevance selection criterion.

SCRIPSIT requestor accreditation – a SCRIPSIT accreditation profile consists of one to many P3P-based accreditation fragments. These may be assembled in any P3P-valid construction. A valid RDF triple or set of triples may attach to any properly formed accreditation fragment. At this point the basis of a discovery and access mechanism is established.

There are variations on this basic requestor accreditation structure:

- Cookie or transient accreditation – the SCRIPSIT entity engine may be instructed to create a transient accreditation profile instance on the requestor. As with website cookies, this accreditation profile may have a validity period ranging from a limited session period to permanent or indefinite duration. Risks associated with this accreditation strategy are minimal. The transient accreditation profile is of little discernable value without a SCRIPSIT entity against which to perform profile validation.
- URI-based accreditation – access to the resource(s) referenced by the URI is limited to accredited agents and/or classes of agent. The agent requests access to a URI-referenced resource and requires access permission to be granted not only by successful accreditation profile matching but also by an agent nominated by the URI-referenced resource.

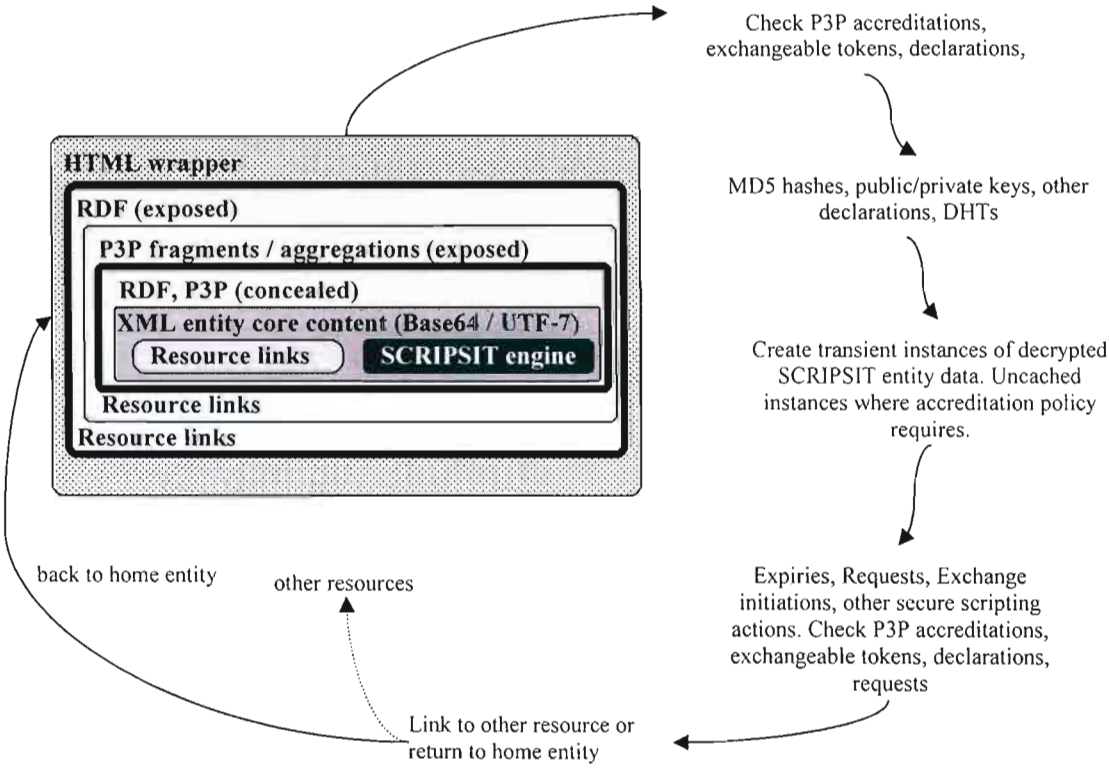


Figure 15 - Exposure and concealment - elements and processes

Flow through the exposure and concealment mechanism is described in Figure 15 above.

5.4 Creation of collections of SCRIPSIT entities

Creating a collection of SCRIPSIT entities may be done via a number of routes, including but not limited to:

- SCRIPSIT-enabled authoring or data publishing tools, combine one or more of the following:
    - Web publishing
    - Data annotation and markup
    - Accreditation policy wizards
    - SCRIPSIT engine publishing (the ability to insert a validated SCRIPSIT engine into the published SCRIPSIT entity)
    - Entity validation
    - Resource discovery wizards or markup tools
    - Graphical layout tools for indicating entity links and relationships

- Simple SCRIPSIT entity creation, requiring the author to manually perform many of the operations. This simple level of entity creation may be achieved via script-based tools or SCRIPSIT-enabled browsers and/or word processing tools. In both of the preceding cases, SCRIPSIT authoring functionality may be achieved via plugins which enable an author to create entities in a familiar software environment, though without the facility for creating complex meshed collections of entities.

5.5 SCRIPSIT engine embedding and execution

Fundamental to SCRIPSIT is an ability for the peer-local processing to support a sandbox-style[ix] Java VM or similar environment. The embedded, encrypted script is concealed from the public interface of the SCRIPSIT entity. Access to the embedded engine is possible only when the locally held accreditation profile at least matches that required to generate a key to access the concealed document content's outermost secure layer. There is scope for complex partial access mechanisms which may be further developed in future work.

A trivial piece of script is required to be embedded at the publicly visible interface of the SCRIPSIT entity in order to process the accreditation fragments and generate the access profiles required to execute the embedded SCRIPSIT engine.

An alternative approach required no trivial script at the public interface, but requires this processing logic to be built into SCRIPSIT-enabled tools and applications. Included in such applications would be CAQDAS tools (e.g. *QSR NVivo, Atlas.ti*) and SCRIPSIT plugins for web browsers and email clients.

[ix] To realise the sandbox model, Java applets are controlled by three consecutive processes:

- Byte Code Verifier. This checks if the code presented fits the rules.
- Applet Class Loader. Ensures that important parts of the Java runtime environment are not replaced by code masquerading as legitimate code, thereby preventing class spoofing.
- Security Manager. Performs runtime checks on suspect or dangerous methods written into the presented code.

The sandbox model history runs thus:

- JDK 1.0. (Local) Java applications are completely trusted, with Java applets running in a sandbox.
- JDK 1.1 added the concept of code signing. This is to say that if an applet is digitally signed, and if the signer is trusted, the applet is treated as a local application and does not run in the sandbox. This is not a desired state for SCRIPSIT.
- JDK 1.2 no longer distinguishes system code, applications and applets. All code runs in controlled environments. More importantly, there is the facility for several customised sandboxes where privileges each piece of code gets may be configured according to differing security policies. Location and/or signer of the code hereby identify what security policy should be used. Thus, Java 2 gives the ability to grant exactly these privileges that are needed, and only when they are needed. This converges very well with SCRIPSIT's requirements.

5.6 Peer-local handling of data

The handling of data embodied in a SCRIPSIT entity is always at a *peer-local* level. *Peer-local* refers to the fact that decrypted data only ever exists in a transient form on the requesting client. Fundamental to this is an embedded engine in the SCRIPSIT entity which decrypts and displays the data at client level (see Figure 16). This happens out of sight and out of reach of any browser caching operations to avoid leaving any unintentional trace of data.

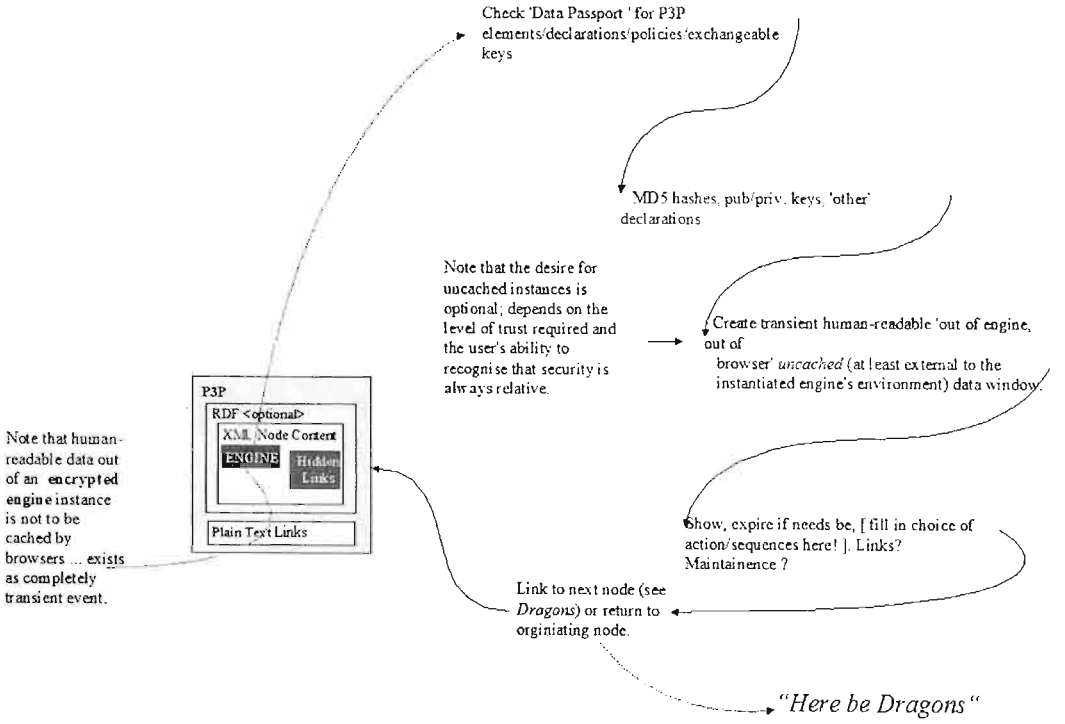


Figure 16 - Overview of peer-local data and process handling in an entity

5.7 Proposed model in situated contexts

Setting the background for SCRIPSIT's subversion of P3P's intended application, a common usage scenario is outlined, based around an imaginary company's website. Considering the proposed subversion of P3P in the context of SCRIPSIT, a simple online shopping scenario is described below. This is not a SCRIPSIT-appropriate application. It serves only to illustrate the switch in usage of P3P from policy

statement to accreditation facility. From passive statement to assertion of right to data and its usage.

> An online shopper seeks a discount notebook sales site. A site is found, http://www.cheapknockoffs.co.za, and our shopper visits this site. CheapKnockOffs have graciously published P3P policies on its site, on each page in the website. It is a prerequisite that our shopper has a browser which is P3P-aware.
>
> With the shopper's P3P-enabled browser, any policies published on the site are automatically fetched and checked against the preferences which our shopper has set up before starting to browse. CheapKnockoffs has a policy statement that data found in HTTP access logs will be collected, and placement of an unlimited duration cookie on the shopper's computer will be requested. The shopper's browser checks these stated policies against existing preferences and finds that there is no objection to interrogation of standard HTTP logs. It also finds that the shopper is absolutely unwilling to allow placement of unlimited duration cookies containing sensitive personal information.
>
> The site's P3P policy states that the requirements of CheapKnockOffs corporate rules are that all customers and potential customers surrender data about themselves to allow the company to engage in consumer profiling. This falls outside of our shopper's default preferences and triggers a warning from the browser. The shopper responds by allowing this data collection to take place as a one-off exception for this site. Having passed this user hurdle, the website entices the shopper to proceed to the purchasing page of the website. After selecting a suitable KnockOff Notebook, the payment section of the website is visited. More sensitive personal information is gathered here, including biographic data, details on the means of payment and delivery address.

> All of this falls outside of the shopper's usual comfort zone and a warning is flagged by the browser. CheapKnockOffs promise that they will only use the data supplied to confirm and complete the online order. Here the shopper is faced with a major trust hurdle. There are absolutely no guarantees that CheapKnockOffs will adhere to their stated data retention and discard policy with respect to the personal data recently harvested.
>
> Assuming that our shopper assesses the risk of personal data being retained and used for unauthorised purposes, the transaction may be confirmed. Failing an assessment of acceptable risk, our shopper decides not to trust the website and cancels the transaction.

It is important to note that this scenario relies heavily on levels of assumed honesty on the part of the policy publisher. Other than checking stated preferences against stated data policies, the user has no guarantees whatsoever. Coyle (1999) comments that P3P

> *... is an engineer's vision of how humans decide who to trust and what to tell about themselves.*
>
> (Coyle, 1999)

Coyle goes on to conclude the point by noting that P3P

> *... has a set of data elements and neatly interlocking parts, and it has nothing at all to do with how people establish trust.*
>
> (Coyle, 1999)

There are numerous critiques of P3P, and Coyle succinctly addresses the primary concern held with respect to establishing trust. Thibadeau (2000) observes that P3P "lacks the ability to negotiate with the Web Server on a contract, and to make a contract with the Web Server", effectively dismissing P3P as a toothless privacy tool. Thibadeau continues and comments that P3P fails to account for or provide any remedy for the transitivity and openness of data on the WWW. The transitivity issue,

according to Thibadeau (*ibid.*), is that of how to assert protection and control over data after it has been put out into a public space.

SCRIPSIT attempts to address this by removing the promissory assertion from the hands of the data harvesters and placing it in the hands of data owners. The proposed model's use of P3P is an inversion of the original, published form. This is crucial for the mobile and assertive control over personal data which forms the basis of SCRIPSIT. Other usage scenarios will now be outlined to situate SCRIPSIT in the mind of the reader.

5.7.1 Secondary reuse of archived qualitative research data

A researcher (academic, governmental, public naive user) has a requirement to access archived qualitative research data stored as a collective (or mesh) of SCRIPSIT entities. The researcher has a basic, predefined, set of academic research SCRIPSIT credentials which allow a level of access commensurate with being able to determine the likelihood of the SCRIPSIT entity mesh under interrogation being suitable for the asserted academic research requirements. Should the researcher find that access to potentially useful data is needed, SCRIPSIT allows for a P3P fragment representing the requested access and the researcher's specific credentials to be submitted to the controlling (custodial/ owner/ archival) entity for consideration. Should the request be granted, a P3P fragment containing the appropriate partial key information (and any other relevant elements) allowing the requested access is returned to the requester.

The requester must then incorporate the received P3P fragment into their own credential/profile for use in any further requests for access to data. A facility for reciprocal credential exchange is also supported by this mechanism. In this sub-case, researcher *A* requests access to data custodially managed by researcher *B*. *A* submits a request to *B* and proposes a credential exchange at a nominated level which, if granted, will allow *A* and *B* mutual access to a larger, aggregated pool of data on a

trust basis. Revocation may be unilaterally pushed from either side and may not be challenged at a SCRIPSIT level.

5.7.2 Context-sensitive enrichment of publicly available third party data

SCRIPSIT is perhaps less effective in the area of context-sensitive enrichment or marking up of publicly available third party data. Issues of fundamental privacy are seldom an issue with data which have been explicitly released in the broadest possible domain. Here questions of negotiated access are not points of debate with respect to the publicly-available data.

Negotiated and/or requested access to data may flow from the publicly available (unencrypted) data where a data user discovers reference to data which has not been made publicly available. In such instances, the data user may be presented with the option to request access to further data. Such requests need to satisfy the following criteria:

- The data user has SCRIPSIT-aware/enabled software tools at his or her disposal.
- Publicly-available data initially investigated is SCRIPSIT-enabled, at least as far as possessing a SCRIPSIT (P3P-based) accreditation request fragment which may be submitted by the SCRIPSIT-aware/enabled software.

Such requests fall under the general scope of information/data discovery and as such will not be expanded upon here.

5.7.3 Creation of dispersed trustable personal data archives

Trust cannot be bought, legislated or demanded. Being an intangible, which can only be given by an individual, trust is dramatically abstracted and removed from the realm

of the specifiable. Trust is perhaps the single aspect of human relationships which cannot be commoditised.

A child removed from a family environment and placed in a social care system finds the establishment of trust particularly challenging. Loss of the natural human right to care and nurturing within a stable family and broader social environment severely damages a child's ability to trust individuals or institutions.

Children in care experience a great number of adult contacts, most of which are transient and/or sparsely episodic. Very little rich continuity is experienced and therefore a great many opportunities to build relationships layer by layer are missed. Trust and the ability to trust are amongst the first and most lasting of the casualties.

In many First World settings, there is the reasonable chance that personal history and narrative information (or digital memories) will be preserved and made accessible to a child in the care system. This is frequently true in the present and in the future tenses. Third World settings cast an altogether more dismal light on the opportunities for and trustability of preservation and access to personal history and narrative information.

Child *A* is in a care system and starts with a fundamental inability to trust as a direct consequence of own experiences.

Central to the reseeding of the ability to trust is the placing of actual control and final authority over personal history and narrative *in the hands of the child* and not in the hands of any other institutions or individuals who claim (however genuinely) to be 'trustable third parties'. Without lending intentional support to the conspiracy theorists of the world, it might be said that a child whose trust has been broken is quite right to 'trust no-one'.

5.7.4 Incorporating P3P profiles and requestors in open access applications

A derivation of the scenario in section 5.7.1 considers the inclusion of SCRIPSIT accreditation fragments and aggregations from the perspective of enabled application tools as used by academic researchers. The application is licenced for academic use and comes with a basic, predefined, set of academic research SCRIPSIT credentials allowing a level of access commensurate with being able to determine the likelihood of the SCRIPSIT entity or collective mesh under interrogation being suitable for the asserted academic research requirements.

Should the application find that access to potentially useful data is needed, SCRIPSIT allows for accreditation fragment(s) representing the requested access and the researcher's specific credentials to be submitted to the controlling (custodial/ owner/ archival) entity for consideration.

On granting of the request, an accreditation fragment containing the appropriate partial key information (and any other relevant elements) may be returned to the application and then added to the application's SCRIPSIT accreditation aggregation, thus allowing the requested access.

The reciprocal exchange of credentials may be supported programmatically. Here application instance *A* requests access to data custodially managed by researcher *B*. *A* submits a request to *B* and proposes a credential exchange at a nominated level which, if granted, will allow *A* and *B* mutual access to a larger, aggregated pool of data on a trust basis.

This potential application of the SCRIPSIT model centres around the inclusion of P3P fragments and aggregations in compliant applications. An instance of this might be academic use of a qualitative data analysis tool such as *QSR NVivo*. In short, this mini-scenario considers research data which has been appropriately tagged. This data may be made available or removed from accessibility to a wider and appropriately

accredited community of researchers. This access need not only be on a simple data access basis, but also from within academically licenced applications which grant a basic level of access to research data by default. Greater access would of necessity be by agreement/application with the data owners and custodians.

This is a special case version of the browser-based local access to a SCRIPSIT entity. Further to this, the facility for submission of access and accreditation requests may be built into the application thereby providing a formalised route for access, accreditation and annotation outside of the simple, general case of browser-based access.

Research data (SCRIPSIT entities and collectives) may be assigned unique identifiers (see *RDF triples* in Glossary) in order to allow the enabled application(s) to target/browse/graze specific SCRIPSIT collections and/or research domains in a semi-automated or scripted manner.

5.7.5 Community memories in public spaces

Publishing on the Web carries the persistent notion of a small voice having global presence and reach. Preservation of individual and community memories in a public space such as the Web carries enormous benefit and equally great risk. There are many threats to the preservation and privacy of elements and entire collections of published memories where these are placed in a public space. All of the solutions proposed to date rely on one or the other trusted third party, be this for cryptographic purposes, secure storage or resilience. Existence on the Web is at a number of levels - personal, community, institutional and/or national. Each of these is characterised by different degrees of risk and exposure.

Privacy and trust issues are intimately involved with all. Community memories all have one attribute in common. This attribute is that contributions and access to the community memory are not time limited. There is no fixed endpoint after which a memory (shared) ceases to exist or becomes invalid. A community is defined by a

shared sense of purpose, location or interests. In the case of a dispersed community, members may not even be aware of one another's existence. Here SCRIPSIT becomes of arguably significant value. A community may establish many initially unlinked online presences and, through resource discovery and a propagation of granted accreditations, gradually coalesce into a complex and self-selecting aggregation of community resources. This is not a variation on the search engine theme but something rather more powerful and universally applicable, which does not rely on any specific service or ideological support.

5.8 Testing of hypotheses

This chapter described the model and frameworks within which the model and its objectives may be achieved. The central hypothesis presented is that a model enabling the creation of trustable privacies in online public spaces is architecturally feasible and practically implementable. This has been done by:

- Describing the social and perceptual aspects of privacy and trust.
- Critiquing past and current praxis with respect to archival and retrieval of qualitative (research) data.
- Considering ethical and social consequences of qualitative research from the perspective of the research subject.
- Considering the requirements of individuals in the care of social enterprises.
- Building a model centered around the privacy and trust requirements of the user/research subject firstly, and secondly around existing technical and computing platforms.
- Considering practical applications of the proposed model in a number of situated scenarios.

The secondary and supporting hypothesis is that it is possible to abstract the model from the constraints and impositions of the communications networks (WWW) and make technically feasible the construction of logical and actual collections of data

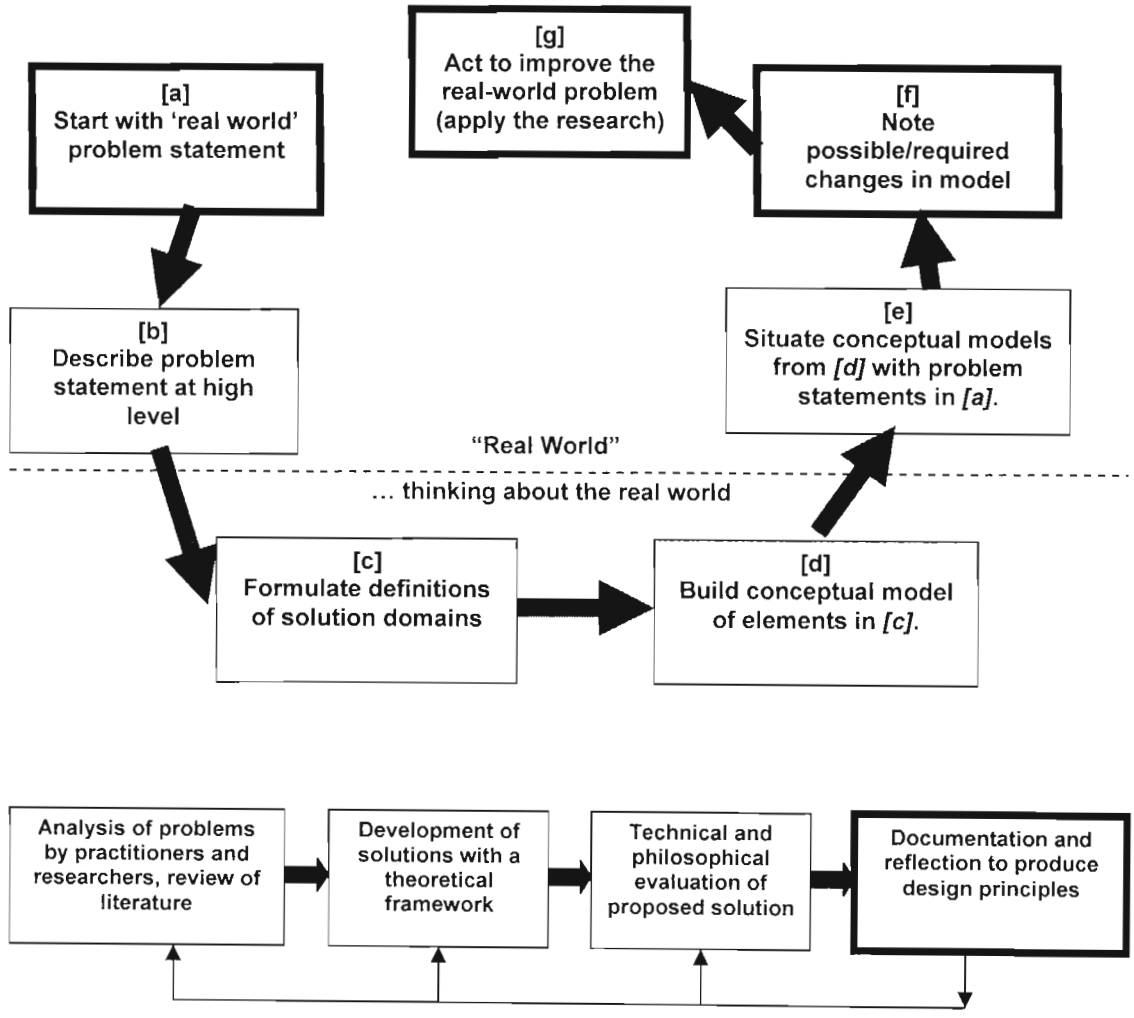
(personal and research data aggregations) which are not dependent on serverside processing or any trusted 3rd parties for any purpose other than simple storing of the proposed encapsulated data entities.

Situation of the model in scenario-based real world applications completes this statement and discussion of the initial hypothesis. Subsequent to this, potential gaps in the model provided the basis for additional hypotheses concerning trustable privacies in public spaces.

Chapter 6

Discussion, conclusions, observations and future work

Conclusions and recommendations with respect to the adequacy of existing models and praxis, apparent fitness of the proposed models and mechanisms and potential for further research in this direction.



6.1 SCRIPSIT and the Web

Berners-Lee (2001) observed in Scientific American that:

> *Knowledge representation ... is currently in a state comparable to that of hypertext before the advent of the Web: it is clearly a good idea, and some very nice demonstrations exist, but it has not yet changed the world. It contains the seeds of important applications, but to realize its full potential it must be linked into a single global system.*
>
> (Berners-Lee, 2001)

Extend this statement and SCRIPSIT might contribute a step towards realising the benefits of the Semantic Web on a level which is accessible to the individual and to primary and secondary qualitative researchers and users.

This chapter summarises issues addressed, key questions asked, conclusions drawn, and potential for future research.

The results of this dissertation lie in the proposed model supporting the creation of trustable privacies in public spaces. Conclusions drawn from the research and development are discussed in this chapter. The initial scope of the research was defined by broadly perceived requirements of, and solutions to, issues surrounding secondary reuse of qualitative research data. A broader set of applications emerged during the literature review and model development process. These are discussed later in this chapter under sections 6.4 and 6.5.

6.2 Review of original aims and rationale

This dissertation presented results of research that aimed at models for the support of privacy, ownership, trust and anonymity, along with context and intended meaning, in archived or published qualitative data.

6.2.1 Original aims

Considerations of privacy, of trust and the mechanisms which might support these will pivot on the central question asked in this dissertation, derived from Bromseth's (2002) posing of the question of who ought to be responsible for the protection of an individual's privacy.

The central question asked, therefore, is who has the right and responsibility of protecting the privacy of the individual. Following on from this question are those which arise from investigations into mechanisms and models which might support such user-centric rights and responsibilities. These questions guided selection of much of the supporting literature reviewed.

6.2.2 Review of rationale

The rationale and motivation for this research was based on the following considerations:

- Addressing fundamental requirements of secondary access to qualitative research data, especially in the social sciences. These requirements were identified as centering on issues around privacy, trust, confidentiality and ethical use. Out of this requirements list emerged a requirement for perceived support of privacy and trust from the perspective of the individual.
- Apparent unsuitability of knowledge management systems and existing peer-to-peer models as models addressing the requirements identified during the exploratory enquiry into secondary access to qualitative research data.
- Identification of domains and applications outside of the initially stated. Regional research interests, displaced persons and societies (including refugees and war and HIV orphans) were amongst those identified. This was done in order to better inform the model development process and thereby result in a model with sound general applicability.

These considerations were further broken down into:

- Challenging situations in the social sciences with respect to data access and reuse (Corti, 2000).
- Issues around collaborative access to and use of qualitative research data.
- A requirement for a self-contained and robust means of mediating, controlling and managing access to qualitative data.
- Addressing of issues around expert mediation required by KMSs in general.
- Challenging of the status of trusted third parties (TTPs) in existing models addressing some of the issues around trust and privacy.
- Applicability on regional and global levels.

The early rationale for this research was expanded to identify a wide range of application domains, and subsequently, to use these domains to test aspects of the model as development proceeded.

6.3 Reflections on the research

The research methodologies employed in this dissertation were selected as being best suited both to the model sought as research product and to the requirements and approach of the author. Arriving at the initial model concept was key to achieving subsequent progress in the iterative model development process. This initial concept relied on a description of the perceived problem domain and identification of technological supports required for realisation of a model supporting both the initial and generalised requirements.

A drawback of the iterative model development process is a lack of certainty over where limits are set with respect to the iterative process. A persistent challenge was keeping to the scope of research and development defined in Chapters 1 and 2.

6.3.1 Limitations encountered

- A model supporting trustable privacies in online public spaces was not achievable within the time and cost constraints of this research project.
- Difficulty in restricting scope of investigation and work
- Time constraints precluded development of prototype implementations of SCRIPSIT.

6.3.2 Empirical reflections

Scientific method was used in the development of SCRIPSIT. The specific problem statement described at the start of the dissertation evolved into a generalised statement and model. The process of iterative development and reflection resulted in a model which fits both the original specific problem statement and the generalised case.

6.3.3 Theoretical reflections

This dissertation considered technological and perceptual questions arising from the specific and general problem statements. Interview and questionnaire techniques were introduced to test convergence of philosophical and technical threads of enquiry which formed the large part of the research. Synthesis of theoretical approach in this dissertation was strongly influenced by the constructivist school, asserting that knowledge, privacy, trust, and many other areas of enquiry are individual experiences and constructs affected by shared experiences and information.

6.4 SCRIPSIT as contribution to privacy and trust tools

SCRIPSIT contributes to the public collection of privacy and trust tools through:

- Returning elective control over access and availability of qualitative data to the original data owner.
- Proposing a peer-centric alternative to the mainstream and well-known models based on peer-to-peer and client-server technologies.
- A novel combination of existing, published standards already in the public domain.
- Demonstrating catholic applicability across domains and user communities.

6.5 Recommendations and future research

- Further studies in this direction pursue lines of enquiry based on the perceptions of the individual. The reference to "individual" is at the broadest level to be inclusive in terms of community and purpose of study.
- Further study using assessments of diverse targeted user communities is required to further refine statements of individual perception and need. Examples of targeted communities include academic researchers, research data users from NGOs and commercial entities, displaced persons and casual users including "bloggers" (Web log users).
- A set of working prototypes of SCRIPSIT be developed and used to refine the model through empirical testing and evaluation by a selection of user communities. This has strong potential to provide the basis for a significant amount of further investigation.
- The results of this dissertation be placed in the public domain, with a GPL-style statement of open technology. The model is intended to place enabling technologies in the hands of data owners. Hence the technology and concepts in SCRIPSIT cannot be licenced nor be subject to any proprietary control.

6.6 Conclusions

The research questions posed at the start of this dissertation were the following:

- *Where is the most appropriate place for vesting of control over private data?* This question was answered progressively more certainly over the course of the research and model development. The point at which it was clear that the model was required to support retention of user control marked a defining moment for the conceptual basis of SCRIPSIT. This clear definition of requirement allowed investigation around the question following to proceed in a directed manner, yielding results which strongly supported the assertion inherent (in the question).
- *Is it feasible to develop a model supporting creation of trustable privacies, with application across multiple domains?* Pursuing this line of enquiry resulted in an early emphasis on literature around issues of privacy and trust. Concentration on these was key to extending the specific question into the general and, therefore, developing the conceptual basis for the resulting model.

This dissertation has demonstrated the initial feasibility of a model supporting the creation of trustable privacies in online public spaces. SCRIPSIT requires extension and testing through prototyping and implementation to prove its suitability in a practical manner. It is hoped that funding may be found to support further research in this direction.

- o -

References

Abecker A., Bernardi A., Hinkelmann K., Kuhn O. and Sintek M. (1998) Toward a technology for organisational memories, *IEEE Intelligent Systems and Their Applications*, May/June 1998, 13(3), 40-48.

Ackerman M.S. and McDonald D.W. (2000*) Collaborative Support for informal information in collective memory systems*, Information Systems Frontiers 2(3/4), 333-347.

Acquisti A. (2002) A *User-centric MIX-net Protocol to Protect Privacy*. UC Berkeley, Workshop on Privacy in Digital Environments: Empowering Users, CSCW 2002, New Orleans, November 2002.

Ahlborn B., Nejdl W. and Siberski W. (2002) *OAI-P2P: A Peer-to-Peer Network for Open Archives, Proceedings of 31st International Conference on Parallel Processing Workshops (ICPP 2002 Workshops), 20-23 August 2002, Vancouver, BC, Canada.*. Online at http://www.kbs.uni-hannover.de/Arbeiten/Publikationen/2002/oaip2p.pdf, Accessed 10 July 2004

Altman I. (1977) Privacy Regulation: Culturally Universal or Culturally Specific? *Journal of Social Issues*, 33 (3), 68-74.

Anderson R.J. (1996) The Eternity service, *In proceedings of 1st International Conference on the Theory and Applications of Cryptology (PRAGOCRYPT '96)*, Prague, Czech Republic

Atlas.ti (2004), *Atlas.ti*, Online at http://www.atlasti.com/, Accessed 3 November 2004

Baker T. (October 2000) A Grammar of Dublin Core, *D-Lib Magazine*, October 2000, Volume 6 Number 10. Online at http://www.dlib.org/dlib/october00/baker/10baker.html, Accessed 12 June 2004

Bandura A. (1986) *Social Foundations of Thought and Action: A Social Cognitive Theory*, Prentice-Hall, 1986 (Englewood Cliffs, New Jersey)

Barber B. (1983) *The logic and limits of trust*. 1st edition. Rutgers University Press (New Brunswick, N.J.)

Berners-Lee T. (1999) *Metadata architecture*. Online at http://www.w3.org/designissues/metadata.html, Accessed November 10th, 2003

Berners-Lee T., Hendler A. and Lassila O. (2001) The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities, *Scientific American*, April 2001. Also online at http://www.cs.nyu.edu/rgrimm/teaching/readings/semantic-web.pdf, Accessed 3 November 2004

Blevins T. (2004) *What is boundaryless information flow?* Presented at Integration Methodology Workshop (The Open Group), 3 February 2004 Online at http://www.opengroup.org/conference-live/uploads/40/4511/EAIIC-TOG-Blevins.pdf, Accessed 1 December 2004

Bonifacio M. (2002) *A Peer-to-Peer Architecture for Distributed Knowledge Management, In proceedings of Net Object Days 2002*. Online at http://www.netobjectdays.org/pdf/02/papers/malceb/0481.pdf, Accessed 4 May 2004

Bosak J. and Bray T. (1999) XML and the Second-Generation Web, *Scientific American*, 5(1999). Online at http://www.sciam.com/1999/0599issue/0599bosak.html, Accessed 3 June 2004

Brandeis L.D. and Warren S.D. (1890) The Right to Privacy. *Harvard Law Review*, 4(5):193-220, 1890

Brase T. (1999), The government is intruding on patients' right to privacy, *M.D. Confidential, Minnesota Physician Magazine*, Online at http://www.cchconline.org/publications/mpppriv.php3. Accessed 1 November 2004

Brickley D. and Guha R.V. (2000) *Resource Description Framework (RDF) Schema Specification 1.0. Technical report*, World Wide Web Consortium (W3C). Online at http://www.w3.org/TR/rdf-schema, Accessed 30 May 2004

Bromseth J. (2002) Public Places, Public Activities?, *Presentation at SKIKT 2002 Conference*. Online at http://www.intermedia.uio.no/konferanser/skikt-02/docs/Bromseth.ppt, Accessed 8 August 2004

Brunk B.D. (October 2002) Understanding the Privacy Space. *First Monday, Online Journal*, Volume 7, number 10. Online at http://firstmonday.org/issues/issue7_10/brunk/index.html, Accessed 19 June 2004

Carmichael P. (May 2002). Extensible Markup Language and Qualitative Data Analysis. Forum: Qualitative Social Research, 3(2). Online at http://www.qualitative-research.net/fqs-texte/2-02/2-02carmichael-e.htm, Accessed 3 March 2004

Castelfranchi C. and Falcone R. (2000) Trust and Control: A Dialectic Link, *Applied Artificial Intelligence Journal*, 799-823. Also in *proceedings of the 2nd Internal iTrust Workshop On Trust Management In Dynamic Open Systems (EU funded workshop at University of Strathclyde, Sept.2002)*

Checkland P. and Scholes J. (1999) *Soft systems methodology in action*. 2nd edition. Chichester, Wiley

Chaum D. (1981) Untraceable Electronic Mail, Return Addresses, and Digital Pseudonyms, *Communications of the ACM*, 24(2), February 1981. Online at http://www.inf.tu-dresden.de/~hf2/anon/Chaum1981/Chaum1981.pdf, Accessed 7 December 2004.

Clarke I., Sandberg O., Wiley B. and Hong T.W. (1999), *Freenet: A distributed anonymous information storage and retrieval system* (White paper), Division of Informatics, University of Edinburgh. Online at http://freenetproject.org/freenet.pdf, Accessed 8 May 2004

Clarke R. (1997), Introduction to Dataveillance and Information Privacy, and Definitions of Terms, Online at http://www.anu.edu.au/people/Roger.Clarke/DV/Intro.html, Accessed 19 June 2004

Corti L., Day A. and Backhouse G. (December 2000) Confidentiality and Informed Consent: Issues for consideration in the preservation of and provision of access to qualitative data archives. *Online peer-reviewed journal: Qualitative Social Research, 1(3)*. Online at http://qualitative-research.net/fqs/fqs-eng.htm, Accessed 28 January 2004

Corti L. (2001) *Qualitative Archives: Short Descriptions*, Online at http://www.qualitative-research.net/fqs-texte/3-00/archive/qualidata-e.htm. Accessed 8 December 2004

Coyle K. (June 1999) *Pretty Poor Privacy*, Online at http://www.kcoyle.net/p3p.html, Accessed 3 August 2004

Cranor L.F. and Reagle J. (1997) Designing a Social Protocol: Lessons Learned from the Platform for Privacy Preferences. *Proceedings of Telecommunications Policy Research Conference. Alexandria, VA.*

Dick B. and Swepson P. (1994) *Appropriate validity and its attainment within action research: an illustration using soft systems methodology*. Online at http://www.uq.net.au/action_research/arp/sofsys2.html, Accessed 7 February 2004

Eberhart A. (2004) *Ontology-based Infrastructure for Intelligent Applications* (Thesis zur Erlangung des Grades Doktor der Ingenieurwissenschaften (Dr.-Ing.) der Naturwissenschaftlich-Technischen Fakultat Ider Universitat des Saarlandes Saarbrucken), unpublished thesis.

Edutella Project. (2003) Online at http://edutella.jxta.org, Accessed 19 July 2004

EPIC, Electronic Privacy Information Center website (undated), http://www.epic.org\index.html, Accessed 3 November 2004

Eytan A. and Huberman B.A. (2000) Free Riding on Gnutella. *First Monday, Online Journal*, Online at http://www.firstmonday.dk/issues/issue5 10/-adar/index.html, September 2000, Accessed 30 May 2004

Fielding N. (2000) The Shared Fate of Two Innovations in Qualitative Methodology: The Relationship of Qualitative Software and Secondary Analysis of Archived Qualitative Data. *Online peer-reviewed journal: Qualitative Social Research 1(3).* Online at http://qualitativeresearch.net/fqs/fqs-eng.htm, Accessed 10 July 2004

Fitzgibbon A. and Reiter E. (2003) *Memories for life: Managing information over a human lifetime. Grand Challenge proposal*, published by UK Computing Research Committee (UKCRC). Online at http://www.csd.abdn.ac.uk/~ereiter/papers/memories.pdf, Accessed 30 January 2004

Freenet Project. (undated) Online at http://freenet.sourceforge.net, Accessed 30 May 2004

Froomkin M. (2000). The Death of Privacy?. *In Stanford Law Review* (May 2000) 1501, Online at http://personal.law.miami.edu/~froomkin/articles/privacy-deathof.pdf, Accessed 1 October 2004

Gavison R. (1980) Privacy and the limits of law. *Yale Law Journal* p.421, 1980

Goldfarb C.F. (2000) *XML in an Instant: A Non-geeky Introduction*, Online at http://www.xmlhandbook.com/press/nongeeky.htm, Accessed 11 September 2003

Goldberg I. and Wagner D. (1998*)* TAZ servers and the rewebber network: Enabling anonymous publishing on the WWW. *First Monday, Online Journal*, (3) 1998. http://www.firstmonday.dk/issues/issue3 4/ goldberg/index.html., Accessed 10 May 2004

Gnutella Project. (undated) Online at http://www.gnutella.com, Accessed 20 January 2004

Gnutella (2001) *Gnutella Protocol Specification*, Online at http://www9.limewire.com/developer/gnutella protocol 0.4.pdf, Accessed 30 April 2004

Graham I. (1999) Putting privacy into context: An overview of the Concept of Privacy and of Current Technologies. *Presented at the Insight Information conference: How to Ensure Customer Privacy in E-Commerce Transactions*. Online at http://www.utoronto.ca/ian/privacy/privacy.doc. Accessed 18 October 2004

Granova A. and Eloff J.H.P. (2004) South African Online Banking: Who carries the risk? *In proceedings of InfosecSA 2004*. Online at http://www.infosecsa.co.za/proceedings2004/081.pdf. Accessed 6 December 2004

Hammersley M. (1997) Qualitative data archiving: Some reflections on its prospects and problems. *Sociology*, 31(1), 131-136.

Hendler J. (2003) *Frequently Asked Questions on W3C's Web Ontology Language (OWL)*, Online at http://www.w3.org/TR/2003/CR-webont-req-20030818/index.html. Accessed 19 December 2003

Hoffman D.L, Novak T.P. and Peralta M.A. (1997) *Information Privacy in the Marketspace: Implications for the Commercial Uses of Anonymity on the Web, Workshop on Anonymous Communications on the Internet: Uses and Abuses*, University of California at Irvine, November 21-23, 1997. Available at http://www2000.ogsm.vanderbilt.edu/papers/anonymity/anonymity2_nov10.htm, Accessed 28 July 2004

Human Genome Information Project (2003), Genome Glossary, Online at http://www.ornl.gov/sci/techresources/Human_Genome/glossary/glossary.shtml, Accessed 4 November 2004

James J.B. and Sørensen A. (2000). Archiving Longitudinal Data for Future Research: Why Qualitative Data Add to a Study's Usefulness *Online peer-reviewed journal: Qualitative Social Research, 1(3)*. Online at http://qualitative-research.net/fqs/fqs-eng.htm, Accessed 27 October 2004

Jenkins M. (2004) *CareZone Africa*: Draft Proposal in response to a request for proposal from DFID. Unpublished proposal © metaLAB, Sussex University Innovation Centre. Submission in progress, confidential draft. Text available from Paul Rodda (roddap@ukzn.ac.za) upon request.

Jonassen D.H., Beissner K. and Yacci M. (1993) *Structural knowledge: Techniques for assessing, conveying, and acquiring structural knowledge*. Hillsdale, New Jersey. Erlbaum

Jordan K. (2003) The Augmented Social Network: Building identity and trust into the next-generation Internet. *First Monday*, 8(8), August 2003, Online at URL: http://firstmonday.org/issues/issue8_8/jordan/index.html. Accessed 1 September 2004

Kang J. (1998) Information Privacy in Cyberspace Transactions, *Stanford Law Review* Vol. 50, p1262

KaZaA Homepage (undated) http://www.kazaa.com, Accessed 10 January 2004

Liang J., Kumar R. and Ross K.W. (2004 *Understanding KaZaA*, Online at http://cis.poly.edu/~ross/papers/UnderstandingKaZaA.pdf, Accessed 6 August 2004

Kementsietsidis A., Arenas M. and Miller R.J. (2003) Mapping Data in Peer-to-Peer Systems: Semantics and Algorithmic Issues. *In proceedings of the ACM SIGMOD International Conference on Management of Data*, June 2003, pages 325-336.

Klamma R. and Schlaphof S. (2000) Rapid Knowledge Deployment in an Organisational-Memory-Based Workflow Environment, *In proceedings of the 8th European Conference on Information Systems (ECIS 2000)*, Vienna, 364-371.

Kubiatowicz J., Bindel D., Chen Y., Czerwinski S., Eaton P., Geels D., Gummadi R., Rhea S., Weatherspoon H., Weimer W., Wells C. and Zhao B. (November 2000) OceanStore: An Architecture for Global-Scale Persistent Storage, *Proceedings of 9th International Conference on Architectural Support for Programming Languages and Operating Systems*. Online at http://oceanstore.cs.berkeley.edu/publications/papers/pdf/asplos00.pdf, Accessed 8 January 2004

Kubiatowicz J. (2003) Extracting Guarantees from Chaos, *Communications of the ACM*, Vol 46, No. 2, February 2003

Kuula A. (December 2000) Making Qualitative Data Fit the "Data Documentation Initiative" or Vice Versa? *Online peer-reviewed journal: Qualitative Social Research, 1(3)*. Online at http://qualitativeresearch.net/fqs/fqs-eng.htm, Accessed 16 March 2004

Kwalitan (2004), Kwalitan, Online at http://www.kwalitan.net/engels/, Accessed 3 November 2004

Lessig L. (1998) *The Architecture of Privacy, 2nd Draft*. Presented at Taiwan Net '98 Conference, Taipei. Online at http://www.lessig.org/content/articles/works/architecture_priv.pdf, Accessed 20 February 2004

Lewins A. and Silver C. (2004) *Choosing a CAQDAS Package*, CAQDAS Networking Project, Surrey University, Online at http://caqdas.soc.surrey.ac.uk/. Accessed 8 December 2004

Margulis S.T. (1977) Conceptions of privacy: current status and next steps. *Journal of Social Issues* 33(3):5-21, p.10.

Markus M.L. (2001) Toward a Theory of Knowledge Reuse: Types of Knowledge Reuse Situations and Factors in Reuse Success. *Journal of Management Information Systems*, 18, 1 (Summer): p.57-93

MAXqda (2004), *MAXqda*, Online at http://www.maxqda.com/maxqda-eng/start.htm, Accessed 3 November 2004

Metzger M. (July 2004) Privacy, Trust, and Disclosure: Exploring Barriers to Electronic Commerce, Online journal at http://www.ascusc.org/jcmc/vol9/issue4/metzger.html, Accessed 3 August 2004

Microsoft (2003) .*NET Passport*: *Balanced Authentication Solutions*, Microsoft Corporation, Online at http://download.microsoft.com/download/a/f/4/af49b391-086e-4aa2-a84b-ef6d916b2f08/net_passport.doc, Accessed 3 November 2004

Miller E., Swick R., Brickley D., McBride B., Hendler J., Schreiber G. and Connolly D. (2001), *Semantic Web*, Online at http://www.w3.org/2001/sw/, Accessed 3 March 2004

Milne A.A. (1926), *The House at Pooh Corner*, ISBN 0416789005, Heinemann Young Books (1974)

Napster Project. (undated) Online at http://www.napster.com, Accessed 3 March 2003.

Nejdl W., Wolf B., Changtao Q., Decker S., Sintek M., Naeve A., Nilsson M., Palmér M. and Risch T. (2000) *EDUTELLA: P2P Networking for the Semantic Web*, Technical Information Library Hannover Learning Lab Lower Saxony. Online at http://www.kbs.uni-hannover.de/Arbeiten/Publikationen/2002/p597-nejdl.pdf, Accessed 17 June 2004

Nejdl W., Dhraief H. and Wolpers M. (November 2001) *O-Telos-RDF: An Extension of RDF with Enhanced Meta-Modeling and Reification Functionalities*, University of Hannover, Germany. Online at http://www.kbs.uni-hannover.de/Arbeiten/Publikationen/2001/kcap01-workshop.pdf, Accessed 14 August 2004

Nejdl W., Wolf B., Staab S. and Tane J. (2002) *EDUTELLA: Searching and Annotating Resources within an RDF-based P2P Network*, presented at the Eleventh International World Wide Web Conference, 2002. Online at http://edutella.jxta.org/reports/edutella_p2p.pdf, Accessed 3 August 2004

Nelson M., Liu X., Maly K. and Zubair M. (2004) *Enhancing Kepler Usability and Performance, In proceedings of the 8th ECDL 2004, Bath, United Kingdom, Sept 2004*. Online at http://kepler.cs.odu.edu/publications/kepler.pdf. Accessed 3 October 2004

Nissenbaum H. (2000) Protecting Privacy in an Information Age: The Problem of Privacy in Public, *Law and Philosophy*, 17: p559-596, 1998.

P3P Project. (undated). *World Wide Web Consortium (W3C),* Online at http://www.w3.org\P3P\index.html, Accessed 12 June 2004

Palacio M. (February 2002) *Investigations on metadata encoding using XML and RDF*, Department of Computer and System Sciences, Stockholm University and the Royal Institute of Technology. Dissertation submitted February 2002. Online at http://www.dsv.su.se/~johank/publications/others/ManuelPalacio/palaciomsc.pdf, Accessed 28 July 2004

Palen L. and Dourish P. (2003) Unpacking Privacy for a Networked World, *In proceedings of Conference on human factors in computing systems*, (129-136), ISBN 1-58113-630-7, 5 April 2004

Phillips D.J. (2001) The Influence of Policy Regimes on the Development and Social Implications of Privacy Enhancing Technologies, *In proceedings of 29th Research Conference on Communication, Information and Internet Policy (Telecommunication Policy Research Council)*, October 27-29, 2001, Alexandria, VA.

Pratt V. (1978) *The Philosophy of the Social Sciences*, ISBN 0-416-76370-7, Methuen & Co.

Qualis Research (2004) *The Ethnograph*, Online at http://www.qualisresearch.com/, Accessed 3 November 2004

QSR International (2004a), *NVivo*, Online at http://www.qsrinternational.com/products/productoverview/product_overview.htm, Accessed 3 November 2004

QSR International (2004b), *N6*, Online at http://www.qsrinternational.com/products/productoverview/product_overview.htm, Accessed 3 November 2004

RDF Modelling for P3P. (July 2000). *World Wide Web Consortium (W3C)*. Online at http://www.w3.org\2000\07\p3pmodel\index.html, Accessed 30 June 2004

Reed M.G., Syverson P.F. and Goldschlag D.M. (1998) Anonymous Connections and Onion Routing, *IEEE Journal on Selected Areas in Communications*, 16(4), 482-494.

Reeves T. and Hedberg J., (2003) *Interactive Learning Systems Evaluation.* Educational Technology Publications, Englewood Cliffs, New Jersey

Reiter M.K. and Rubin A.D. (1998) Crowds: Anonymity for web transactions. *ACM Transactions on Information System Security*, 1(1), April 1998, Online at http://avirubin.com/crowds.pdf, Accessed 11 September 2004

Rhea S. (2003) Pond: the OceanStore Prototype, *In proceedings of 2nd USENIX Conference on File and Storage Technologies (FAST '03)*, March 2003

Rheingold H. (1993) *The Virtual Community: Homesteading on the Electronic Frontier*. Reading, Addison-Wesley, Massachusetts. Published online at http://www.rheingold.com/vc/book/. Accessed 19 September 2004

Roberts K.A, Wilson R.W. (May 2002). ICT and the Research Process: Issues Around the Compatibility of Technology with Qualitative Data Analysis. Forum: Qualitative Social Research [On-line Journal], 3(2). Online at http://www.qualitative-research.net/fqs-texte/2-02/2-02robertswilson-e.htm, Accessed 10 October 2004

Rodda P.T. (September 2003) Preserving context and intended meaning in archived qualitative research data. *In proceedings of the 5th Annual Conference on World Wide Web Applications at the University of Durban-Westville*. See Appendix D. Online at http://general.rau.ac.za/infosci/www2003/program/WWW2003_Abstracts_POSTERS.htm, Accessed 16 June 2004

Rodda P.T. (January 2004) Open Sesame: A draft mechanism for controlling access and content visibility in archived Qualitative Research Data, *Abstract in conference proceedings*. See Appendix E. Online at http://idlelo.uwc.ac.za, Accessed 18 February 2004

Rose J. (2000) *Information systems development as action research – soft systems methodology and structuration theory*, Online at http://www.cs.auc.dk/~jeremy/pdf files/thesis.pdf. Unpublished thesis.

Rotenberg M. (2001) Fair Information Practices and the Architecture of Privacy, *Stanford Technology Law Review*, Rev.1. Online at http://stlr.stanford.edu/STLR/Articles/01_STLR_1 Accessed March 13th, 2004

Scheeres J. (June 2001) Asia. My shoe size? It'll cost you. *Wired News,* Online at http://www.wired.com/news/business/0,1367,44278,00.html, Accessed 14 December 2004

Schirmer A.L. (September 2003) Privacy and knowledge management: challenges in the design of the Lotus Discovery Server, *IBM Systems Journal*, September.2003, 519-531, Online at http://www.research.ibm.com/journal/sj/423/schirmer.pdf, Accessed 17 March 2004

Seamons K.E., Winslett M., Yu T., Smith B., Child E., Jacobson J., Mills H. and Yu L. (2002) Requirements for Policy Languages for Trust Negotiation, *In proceedings of the 3rd International Workshop on Policies for Distributed Systems and Networks (POLICY 2002)*, Monterey, California.

Seamons K.E., Winslett M., Yu T., Yu L. and Jarvis R. (2003) Protecting Privacy During On-line Trust Negotiation, *Lecture Notes in Computer Science*, Volume 2482(2003), 129-143, Springer-Verlag

Skuse A. (June 2000) Information communication technologies, poverty and empowerment, *Social Development Department, DFID (UK Government white paper)*. Online at http://dfid.gov.uk/Pubs/files/sdd_dn3.pdf, Accessed 13 April 2004

Suryanarayana G. and Taylor R.N. (2004) *A Survey of Trust Management and Resource Discovery Technologies in Peer-to-Peer Applications*, Institute for Software Research University of California, Irvine, ISR Technical Report # UCI-ISR-04-6, Online at http://www.isr.uci.edu/tech_reports/UCI-ISR-04-6.pdf, Accessed 1 November 2004

Thibadeau R. (2000) *A Critique of P3P: Privacy on the Web*, Online at http://dollar.ecom.cmu.edu/p3pcritique/, Accessed 6 June 2004

Thomas J.C., Kellogg W.A. and Erickson T. (2001) The knowledge management puzzle: Human and social factors in knowledge management, *IBM Systems Journal*, Vol.40, No.4, 2001

Thompson P. (2000) Re-using Qualitative Research Data: a Personal Account, *Online peer-reviewed journal: Qualitative Social Research, 1(3)*. Online at http://qualitativeresearch.net/fqs/fqs-eng.htm, Accessed 11 November 2003

Tomek I. (2001) Knowledge Management and collaborative virtual environments. *Journal of Universal Computer Science*, 7(6), 458-471.

Tsui E. (2000) Exploring the KM Toolbox, *Knowledge Management*, 4(2), October 2000, 11-14. Online at http://www.kmmagazine.com/Articles/DisplayArticle.asp?PageID=86406992, Accessed 1 November 2004

United Nations (1948) *Universal Declaration of Human Rights*, adopted and proclaimed by General Assembly resolution 217 A (III) of 10 December 1948, Online at http://www.un.org/Overview/rights.html, Accessed 3 October 2004

van den Akker J. (1999). Principles and methods of development research. *In* van den Akker J., Nieveen N., Branch R.M., Gustafson K.L. and Plomp T. (eds.), *Design methodology and developmental research in education and training,* 1-14. Kluwer Academic Publishers, Amsterdam

Waldman M., Rubin A.D. and Cranor L.F. (2000) *Publius: A robust, tamper-evident, censorship-resistant web publishing system*: Proceedings of the 9th USENIX Security Symposium, Denver, Colorado, USA, 59-72.

Walkerdine J., Melville L. and Sommerville I. (2002). Dependability Properties of P2P Architectures, *In proceedings of the IEEE Computer Society 2nd International Conference on Peer-to-Peer Computing (P2P'02): IEEE*

Wenger E.C. and Snyder W.M. (2000) Communities of Practice: The Organisational Frontier, *Harvard Business Review,* 78(1), 139-145.

Wenger E.C. (2001) *Supporting communities of practice: a survey of community-oriented technologies*, Online at http://www.ewenger.com/tech/index.htm, Accessed 15 August 2004

Westin A.F. (1967) *Privacy and Freedom*, Bodley Head, 1970

Williams M. (2000). Social research—the emergence of a discipline? International *Journal of Social Research Methodology*, 3(2), 157-66. Cited in: Fielding N. (2000) The Shared Fate of Two Innovations in Qualitative Methodology: The Relationship of Qualitative Software and Secondary Analysis of Archived Qualitative Data, Qualitative *Social Research* 1(3).

XML Specifications. (February 2004) *W3C Recommendation, World Wide Web Consortium (W3C)*. Online at http://www.w3.org/TR/2004/REC-xml-20040204, Accessed 8 August 2004

Zhao B.Y., Kubiatowicz J.D. and Joseph A.D. (April 2001) *Tapestry: An infrastructure for fault-tolerant wide-area location and routing. Technical Report UCB/CSD-01-1141*, Computer Science Division, U. C. Berkeley. Online at http://www.cs.berkeley.edu/ravenben/publications/CSD-01-1141.pdf, Accessed 20 November 2003

Zinnbauer D. (2001), Internet, civil society and global governance: the neglected political dimension of the digital divide, Development Studies Institute, London School of Economics, *Information & Security.* Vol. 7 (2001), 45-64.

Appendix A - Questionnaire

Access to qualitative data archives and repositories

(Short pilot questionnaire on access, control and privacy)

PLEASE NOTE THAT THIS IS A PILOT QUESTIONNAIRE AND AS SUCH, IS NOT A FULL DATA GATHERING EXERCISE.

This questionnaire is part of a pilot assessment of perceptions on access to qualitative data. This data may be primary or secondary research data, personal and private data with a restricted audience, or community memories held in public spaces. Please email me (Paul Rodda) on roddap@ukzn.ac.za, should you have any queries or concerns. There is ONE PAGE only to complete. Thank you for your time

Durban, 25th of November

NAME: ..

INSTITUTION: ..

CONTACT DETAILS: ..

Question 1 - Please indicate your THREE primary professional interests in qualitative data archives (in descending order of importance). Replies are open-ended:

1.	2.	3.

Question 2 - Please indicate THREE primary personal interests in qualitative data archives (in descending order of importance). Replies are open-ended:

1.	2.	3.

Question 3 – Please indicate your perception of the degree of control appropriate to place in the hands of research subjects, with respect to access to data specific to individual subjects. Check ONE BOX only to indicate.

None	Request via researcher	Mediated by trusted 3rd party	Full (mediated)	Full (autonomous)

Question 4 – Please indicate your perception of the degree of control which you feel appropriate for your own control over your own private (non public, non research-related) data. Check ONE BOX only to indicate.

None	Request to agency holding data personal to you	Mediated by trusted 3rd party	Full (mediated)	Full (autonomous)

Question 5 – Please describe briefly your views on the general usefulness of a model supporting distributed, user-controllable granting and rescinding of access rights in [a] the qualitative research arena and [b] in terms of preservation of personal privacies in public spaces.

[a] Qualitative Research	
[b] Personal Privacies	

Appendix B – RDF and data triples

Resource Descriptor Format (RDF) in more detail

B.1 RDF characterised and defined

- RDF is machine-understandable information
- Describes properties of resources on the Web. Uses include resource discovery, sorting and categorisation, management of library resources.
- RDF statements specify both properties and values of Web resources
- RDF is defined in terms of XML

See Figure 17 and Figure 18.

Resources pointed to by RDF may be real or virtual. These may be defined as:

- Real resources - anything named by an URL (Web pages (*.htm and so forth), email, server locations) *and*
- Virtual resources: references (electronic) to representations or links to real world resources including books, journals, paper records, people, places and so on.

B.2 RDF statements / RDF data triples

RDF statements are otherwise known as RDF triples (see Glossary) which uniquely define a subject-predicate-object data triple. Examples are illustrated below and in:

Resource / Subject	Property / Predicate	Value / Object
"Dissertation"	About	"Privacy and Trust"
http://ukzn.ac.za	MIME type	"text/HTML"
"Dissertation"	Author	P.Rodda

Properties are simple defined as relationships between Web resources and values. Values may be strings of characters or another resource reference. Properties of a

- person include age, height, mass, hair colour, gender, ...
- journal article include word count, author, topic, journal issue, ...

B.3 XML Namespaces

An XML namespace is a collection of names, identified by a URI reference, used in XML documents to specify element types and attribute names. XML namespaces are distinguished from namespaces used elsewhere computing disciplines by the fact that the XML versions have internal structure and are not sets in mathematical terms.

- RDF namespaces use "rdf" prefix
- RDF Schema namespaces use "rdfs" prefix
- Properties are declared in other namespaces and are hence unique in web terms
- Namespaces changes are driven by rule changes

Simple RDF-only sample

```
<rdf:RDF xmlns:rdf="..." xmlns:roddans="...reference to ns definition..." >
        <rdf:Description about="Draft v26.doc">
                <roddans:title>Privacy and Trust</roddans:title>
                <roddans:author>Rodda</roddans:author>
        </rdf:Description>

        <rdf:Description about="...">
                ...
        </rdf:Description>
</rdf:RDF>
```

Proxy Resource reference in RDF:

```
<rdf:RDF xmlns:rdf="..." xmlns:roddans="...reference to ns definition..." >
        <rdf:Description ID="Rodda">
                <roddans:identitynr>151204 0001 081</roddans: identitynr >
                <roddans:gender>M</roddans:gender>
                <roddans:public resource=http://me.org/rod99</roddans:public>
        </rdf:Description>
</rdf:RDF>
```

Nested descriptions – if a value is a virtual resource, it may exist inside the property element, thus:

```
<rdf:Description about="...">
        <roddans:gender>
                <rdf:Description ID="G">
                        ...
                </rdf:Description>
        </roddans:gender>
</rdf:Description>
```

B.4 RDF Containers

RDF defines containers which are virtual resources including one or more web resource references or values. The common containers defined are:

- Bag – an unsorted collection (citizens of Utopia)
- Seq – a collection with implied ordering (chapters of a book).
- Alt – a set of alternative resources for addressing one requirement, with the preferred resource listed first (UKZN e-library location of a journal, followed by alternative sources for the journal)
- Elements – named properties

Examples of bags follow:

A bag which refers to an external type resource:

```
<rdf:Description ID="committee">
        <rdf:type resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Bag"/>
                <rdf:_1>Rodda</rdf:_1>
                <rdf:_2>Clarke</rdf:_2>
                <rdf:_3>Urquhart</rdf:_3>
</rdf:Description>
```

A better implementation of a bag is as a list:

```
<rdf:Bag ID="committee">
        <rdf:li>Rodda</rdf:li>
        <rdf:li>Clarke</rdf:li>
        <rdf:li>Urquhart</rdf:li>
</rdf:Bag>
```

It is implicit in the use of "rdf:Bag" that the value of "type" is **http://www.w3.org/1999/02/22-rdf-syntax-ns#Bag**

A property can appear more than once with different values – "age" is a property which simultaneously has different values for different people. It also has constantly increasing (changing) values for each person.

B.5 Reification (or, recasting of statements)

Reification is simply recasting of statements from the direct to the referential. "Rodda is the author of this dissertation" is a direct statement. "Clarke commented that Rodda is the author of this dissertation" is a reified version of the first statement and asserts a property of Clarke and not of the dissertation or of Rodda. See Figure 19 following in this section.

Reification properties

```
<rdf:Description about="...">
        <roddans:originator>Rodda</roddans:originator>
</rdf:Description>
```

is reified thus:

```
<rdf:Statement>
        <rdf:subject resource="..."/>
        <rdf:predicate resource="...#author/>
        <rdf:object>Rodda</rdf:object>
</rdf:Statement>
```

B.6 RDF Schemas

- Describe rules for using RDF properties and are expressed as RDF
- Are NOT XML Schemas, which are meant as ultimate DTD replacements

B.7 RDF Classes

RDF classes are

- groups of Web resources
- identified by URLs

Additionally, there is a special class consisting of all possible RDF strings – stated as "rdfs:Literal".

Property-centric classes

The majority of Object Oriented classes specify completely what properties they have and what types are included. Extended into RDF, each property specifies what classes of subjects and objects it relates. This permits addition of properties without altering the class.

Specifying a class is done by creating an RDF resource of type rdfs:Class: <rdfs:Class id="RoddaClass">

```
    <rdfs:label>Rodda Personal Class</rdfs:label>
    <rdfs:comment>Rodda personal data class</rdfs:comment>
</rdfs:Class>
```

Further, specifying a property is done by creating an RDF resource of type "rdfs:Property"

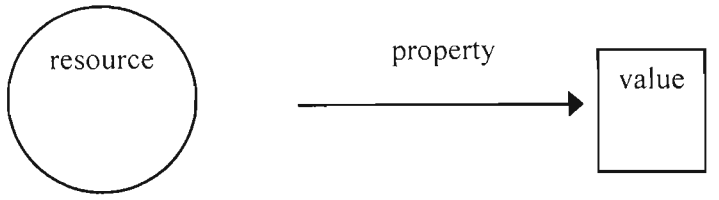
```
<rdfs:Property id="RoddaPropertyOne">
    <rdfs:comment>Rodda security assertion</rdfs:comment>
    <rdfs:domain resource="#RoddaClass"/>
    <rdfs:range resource="..#Literal"/>
<rdfs:Property>
```

B.8 Schema URIs

Ordinary XML namespace URIs only guarantee uniqueness. No assertion or assumption of useful reference may be made at all. Where used in RDF, namespaces ought to refer to an RDF schema document. Those RDF schemas which are referred to at the hidden or concealed level within a SCRIPSIT entity may additionally refer to schemata and namespaces carried within SCRIPTSIT entities.

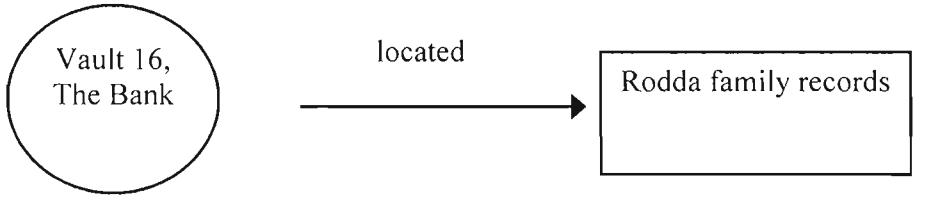
Following are diagrammatic representations of aspects of RDF as described in the first part of Appendix B.

RDF model as a statement



- simple triple-based model (resource – property – value)
- resources represented by nodes with URI(s)
- property is an attribute of the resource
- values literal or pointers to other statements
- statements (1 to n) about a resource forms the description

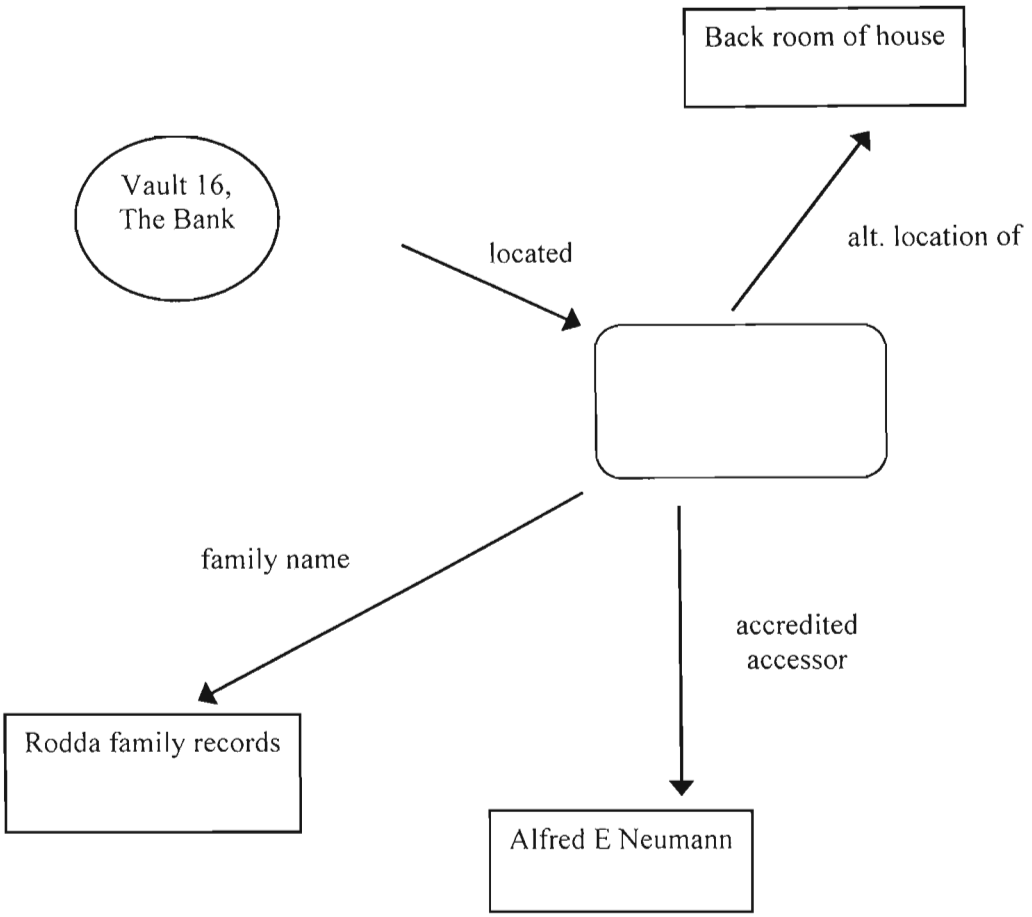
RDF expanded in an example



- The *Rodda family records* are *located* at *Vault number 16 at The Bank*.
- There is nothing preventing multiple values pointing at one resource, nor to prevent multiple resources from being pointed at by one value.

Figure 17 - RDF model as statement

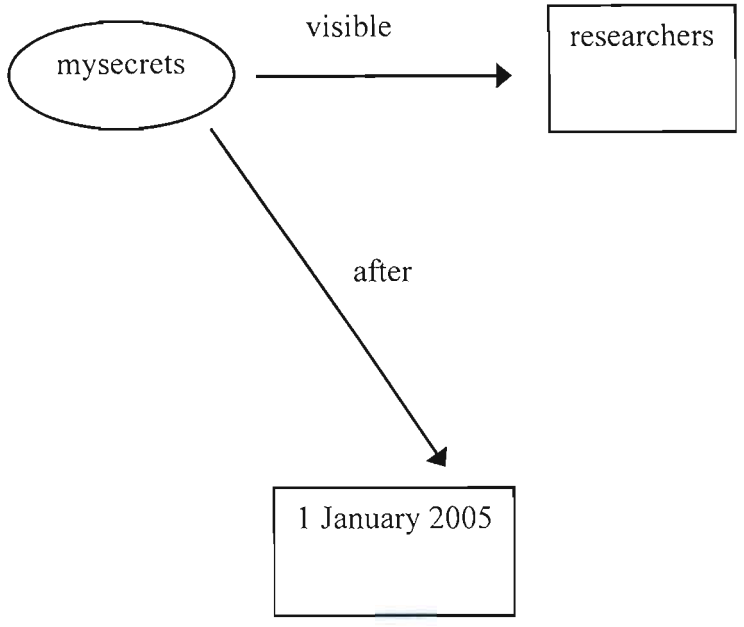
RDF - structured example



- structured metadata realised by replacing text value with value of another node.
- The *Rodda family records* are *located* in *Vault 16, The Bank.*
- Further, *Alfred E Neumann* is an *accredited accessor* of the contents of *Vault 16, The Bank.*
- The *back room of (the) house* is an *alternative location for a copy of* the contents of *Vault 16, The Bank.*

Figure 18 - RDF structured example

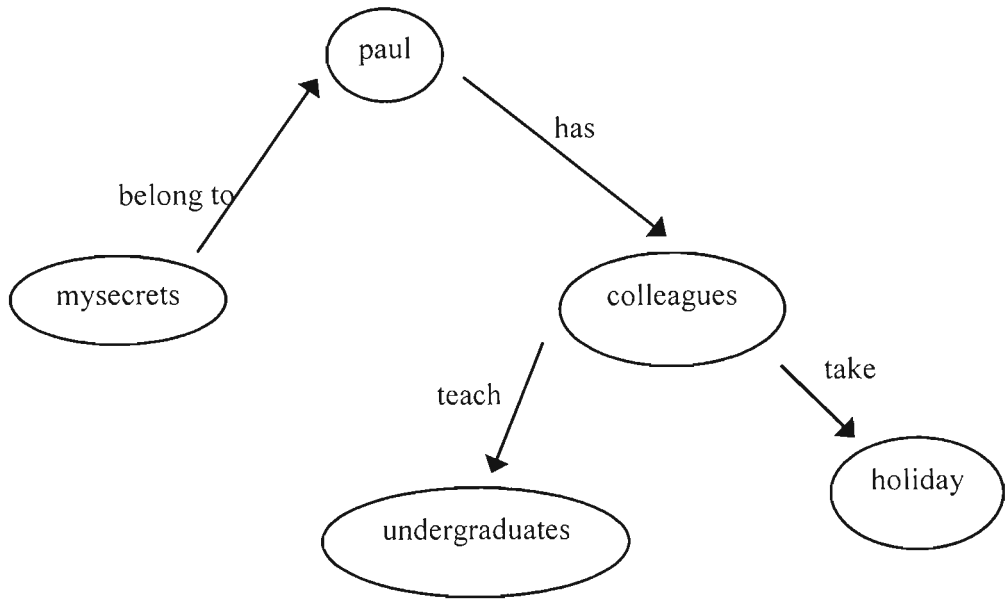
RDF - reification



- Reification is making statements about statements
- In this example, *mysecrets* are *visible* to the group of users known as *researchers*, but only *after* the *1st of January 2005.*
- Changing either of the values (researchers or 1 January 2005) leads to the overall meaning and implication of the RDF statement changing.

Figure 19 - RDF reification example

RDF – edge directed graphs



- A number of *edge-directed* assertions are made above:
  - (belong to, mysecrets, paul) – mysecrets belong to paul
  - (has, paul, colleagues) – paul has colleagues
  - (teach, colleagues, undergraduates) – colleagues teach undergraduates
  - (take, colleagues, holiday) – colleagues take holiday
- This is a trivial example of an edge-directed graph of RDF nodes. In SCRIPSIT's case, nodes are SCRIPSIT entities, with RDF elements making up the edges.

Figure 20 - RDF edge directed graphs

Appendix C – Entity skeleton

```
P3P exposed
       XML exposed wrapper
              XML wrapper
                     P3P concealed
                     XML partial key
                     [ Engine sits here inline with encrypted
                     [ content. Also required java classes present.
                     [ After this, execute SECURE engine with the
                     [ following binary base64 encoded content. All
                     [ data is UTF-7/Base64.
                     <encxml:binary xmlns:encxml="http://www.scripsit.org/encxml">
                            7E57A909B9C9D9449DA5F51A907A5F51A
                            909B9C9D9449DA5F51A909B9C9A907A5F
                            51A909B9C9D9449DA5F51A909B9C7E57A9
                            09B9C9D9449DA5F51957A909B9C9D9449D
                            A5F519A907A5F51A909B9C9D9449DA5F51
                            </encxml:binary>
                     LINKS
                     end P3P concealed
              end wrapper
              LINKS
              PLAIN
       end exposed wrapper
end P3P
```

```
<?xml version="1.0" encoding="UTF-8"?>

<!-- ------------------------- -->
<!-- Uninodal form of SCRIPSIT -->
<!--                           -->
<!-- 1 July 2004               -->
<!--                           -->
<!-- ------------------------- -->

<!DOCTYPE rdf:RDF [
  <!ENTITY rdf      "http://www.w3.org/1999/02/22-rdf-syntax-ns#">
  <!ENTITY rdfs     "http://www.w3.org/2000/01/rdf-schema#">
  <!ENTITY p3p      "http://www.w3.org/2002/01/p3prdfv1#">
  <!ENTITY pubnote  "http://www.scripsit.org/2004/pubsecurity#">
  <!ENTITY privnote "http://www.mydiary.org/2004/mysecurity#">
]>

<rdf:RDF
  xmlns:rdf ="&rdf;"
  xmlns:rdfs="&rdfs;"
  xmlns:p3p ="&p3p;"
  xmlns:scripsit =&scripsit;">

  <p3p:Policy rdf:ID="web browser">
    <p3p:disclosure
       rdf:resource="http://www.displacedpersons.org/secure.htm"/>
    <p3p:entity rdf:parseType="Resource">

      <p3p:authority.name rdf:value="University"/>
      <p3p:contact.email rdf:value="library@ukzn.ac.za"/>

    </p3p:entity>

    <p3p:access rdf:resource="&p3p;AccessClass-nonident"/>

  </p3p:Policy>
</rdf:RDF>
```

Archival and reuse of Qualitative Research Data : Seeking a realisation framework in an African Context

Paul Rodda
Centre for Information Technology in Higher Education · ITEd · University of Natal, Durban
roddap@nu.ac.za 031 260 1124

Appendix D – Web 2003 Poster

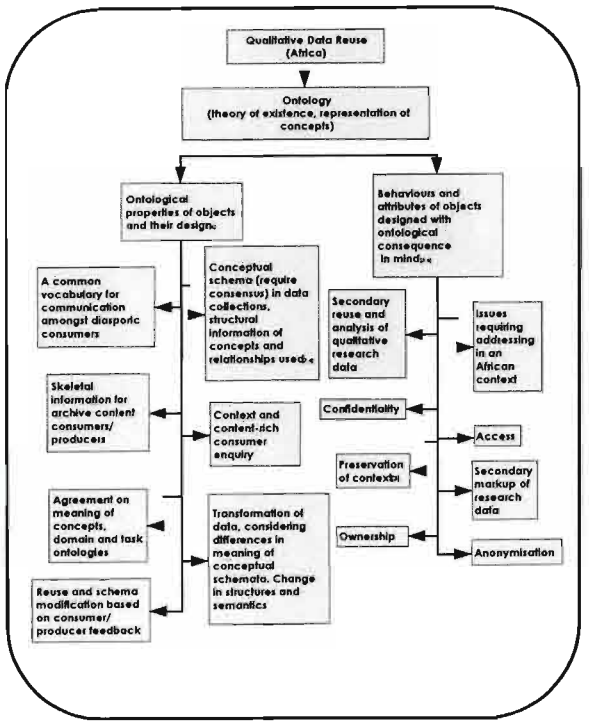
❶ Abstract

Qualitative research lags behind the quantitative world with respect to archival and reuse of research data. Issues around access to and reuse of qualitative research data have come increasingly to the fore in recent years. Central to a relevant and universally applicable archival and reuse strategy is accommodation of qualitative and quantitative data. Collaborative research between quantitatively-driven science and technology and the qualitatively-driven social sciences is particularly relevant locally (and globally) – with particular reference by this research to Indigenous Knowledge Systems (IKS), Information and Communications Technology (ICT) and Societies in transformation (NRF research focus areas – Distinct Opportunities).

Capture, encoding, classification and multi-level, multi-party reuse of qualitative research data have distinct requirements in terms of access, confidentiality, ethics, security and ownership. It is not sufficient to capture and encode qualitative research data: such data must be accessible. Technology underpinning these requirements exists in relatively well-defined XML, RDF [2,4,7,8] and related technologies. As crucial are the means of securing, accessing and preserving the archived data – questions of [3] confidentiality, ownership, anonymisation, security and secondary enrichment of original data require consideration.

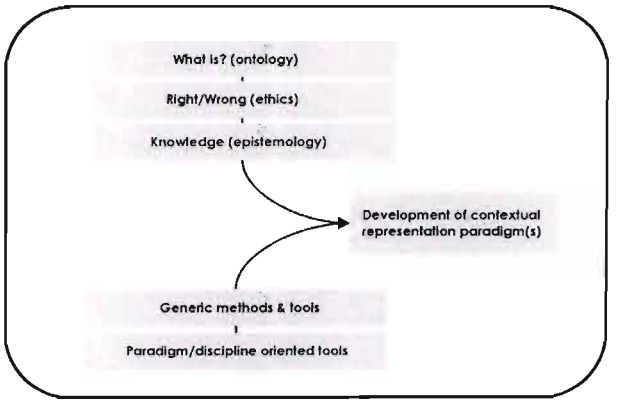
❷ Research Tree

The inverted tree below has an ontological base. Before it is possible to archive context and meaning, concepts and complex relationships need to be defined. A representation of concept and a means of representing the existence of context is required.

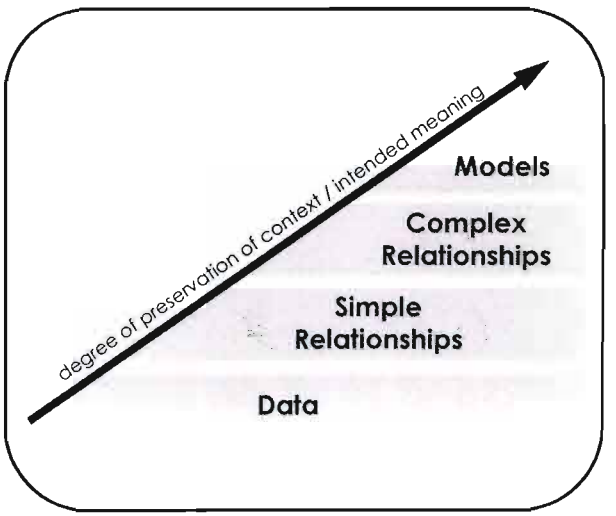


❸ Distinguishing expert management from authentic encoding

A distinction is drawn between knowledge management and contextually authentic archival and accessing of qualitative (social science) research data [4]. Knowledge management tools are well known and accepted in commercial and narrowly-defined domains (usually created and accessed by domain experts only).



Widespread adoption of interoperable metadata standards is pursued by the Dublin Core Metadata Initiative (DCMI) [8]. This is not the only metadata initiative in the international arena, though it is one of the better examples. Central is the development of specialised metadata vocabularies for describing resources, enabling better targeted data recovery tools and methods. Feeding into this are the epistemological and ontological domains from different areas of content. Specific 'Africanised' metadata definitions are essential to the furtherance of qualitative data archiving and reuse in a local context. Further research into the creation of such metadata sets is expected.



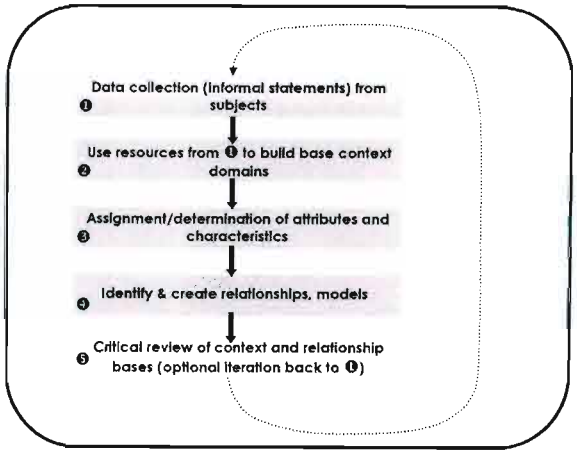
❹ Expected outcomes and interim comments

Preservation of context and meaning in archived data is greatly dependent upon explicit encoding of relationships (simple and complex schemata) and the eventual creation of super-schemata representing multiple contexts and perspectives. Context existing in the minds of the original researchers and subjects is largely inaccessible by secondary users (including academics, public and private organisations, communities and casual data browsers).

Third-party access to archived social science research is an extreme situation with respect to data reuse. Collaborative access to and reuse of archived data consists of a mix of producers and users of knowledge. Appropriate application [2,5] of expert domain knowledge and preservation of contextual prompts is the norm. The creation of networks of multi-level accessible and cost-effective qualitative research data appears to be a challenge which can only improve understanding and application of oft-repeated qualitative research. Methodologies and models to promote greater collaboration between the natural and social science domains is an expected outcome of this research.

A significant shift in priorities and problems is found with research data reusers who are not domain experts, seeking expert and contextually relevant results from investigations. Formulation of queries, identification of appropriate resources are a few of the challenges faced.

The premise on which this research was started is simply this: It is possible to successfully preserve context and relevance in qualitative research data in such a manner as to facilitate 3rd party access and reuse. It is the objective of this research to add to the overall body of knowledge with particular reference to the needs of qualitative research reusability in an African context.



❺ References

1. Beys P., Jansen M., (1999, September). "Automatic Reuse of Knowledge: A Theory", Faculty of Psychology University of Amsterdam Social Sciences Informatics, 1018WB Amsterdam, The Netherlands <email: (pascal.jansen@swi.psy.uva.nl>. online at <http://sern.ucalgary.ca/KSI/KAW/KAW99/papers/Beys1/autoreuse.pdf>. (originally submitted for KAW'99, Twelfth Workshop on Knowledge Acquisition, Modeling and Management Banff, Alberta, Canada, 16-21 October 1999). (last accessed June 6th 2003).
2. Carmichael P. (2002, May). "Extensible Markup Language and Qualitative Data Analysis", Forum Qualitative Sozialforschung (Online Journal), 3(2). online at <http://www.qualitative-research.net/fqs/fqs-eng.htm> (last accessed June 10th 2003)
3. Corti L., Day A., Backhouse G., (2000, December). "Confidentiality and informed Consent: Issues for consideration in the preservation of and provision of access to qualitative data archives", Forum: Qualitative Social Research (On-line Journal), 1(3). online at <http://qualitative-research.net/fqs/fqs-eng.htm> (last accessed May 23rd 2003)
4. Fischer W.J., (2001, June). "Knowledge Reuse: The Roles of Human and Technical Intermediaries", Thesis submission to Faculty of the Graduate School of Arts and Science, Georgetown University. online at <http://ccf.georgetown.edu/theses/billfischer.pdf>. (last accessed August 28th 2003)
5. Markus, M.L. (2001) "Toward a Theory of Knowledge Reuse: Types of Knowledge Reuse Situations and Factors in Reuse Success", Journal of Management Information Systems (M.E.Sharpe), 18, 1 (Summer): 57-93. online at <http://www.ingentaselect.com/vl=1/cl=5/nw=1/rpsv/catchword/sharpe/07421222/v18n1/m_cp=1.htm>
6. Mizoguchi R., Ikeda M., "Towards Ontology Engineering", The Institute of Scientific and Industrial Research Osaka University, 567 Japan, Technical Report AI-TR-96-1, I.S.I.R., Osaka University. online at <http://www.ei.sanken.osaka-u.ac.jp/pub/miz/miz-onteng.pdf>. (last accessed July 7th 2003).
7. Roberts K.A., Wilson R.W., (2002, May). "ICT and the Research Process: Issues Around the Compatibility of Technology with Qualitative Data Analysis", Forum: Qualitative Social Research (On-line Journal), 3(2). online at <http://www.qualitative-research.net/fqs/fqs-eng.htm> (last accessed May 23rd 2003).
8. Dublin Core Metadata Initiative, website at <http://dublincore.org/about/>. (last accessed 30th August 2003)

Open Access : A draft model for controlling access and content visibility in archived Qualitative Research Data

Paul Rodda
Centre for Information Technology in Higher Education · ITEd · University of KwaZulu- Natal, Durban
roddap@nu.ac.za 031 260 1124

Abstract and rationale

Secure multilevel, multiuser access to Qualitative Research Data (QRD) requires rigorous implementation of policies and mechanisms which not only secure access to QRD portals, but also control the depth and breadth of the QRD returned. Qualitative research may be defined as an interdisciplinary and transdisciplinary domain which transects and traverses the humanities and the social and natural sciences. Ownership, ethics, secondary enrichment of original data, confidentiality and anonymisation of QRD have the potential to be addressed via the structures and mechanisms afforded by the Platform for Privacy Preferences Project (P3P). P3P is aimed primarily at online repository privacy practices in a form suitable for automated retrieval and interpretation by user agents. P3P user agents allow users to be informed of repository practices and to automate decision-making based on these practices when appropriate. eCommerce and public access to records have been the primary areas of interest with P3P to date. It is proposed that P3P may constitute a workable and trustable basis for mediating access to archived QRD.

Collaborative research between quantitatively-driven science and technology and the qualitatively-driven social sciences has become especially relevant locally and globally. African foci are found in the areas of Indigenous Knowledge Systems (IKS), Information and Communications Technology (ICT) and Societies in transformation (NRF, South Africa).

An initial mechanism for QRD access mediation is proposed in this poster. This is based upon P3P and RDF from the perspective of open access to QRD. This is in contrast to the eCommerce and public record driven implementations and proposals present in current literature. The same access model appears to be applicable where QRD content is to be archived and made available for reuse. This poster describes the broad access control and management structures and relationships envisaged as part of the external support for archived qualitative research data. This has its origins in the pursuit of a set of methods to facilitate the reuse of qualitative research data. Models included in this poster are applicable beyond the original motivation for the research into reuse of qualitative research data. Semantic and ontological questions arising from this are beyond the scope of this presentation.

Confidentiality, ownership, custodianship, anonymisation and ethics

Questions arising from requirements for access to archived (qualitative) research data:

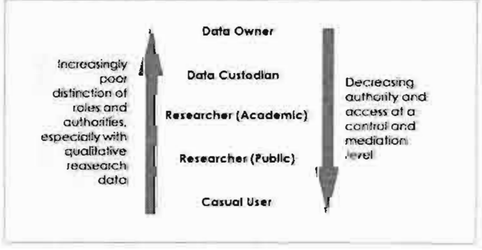
- As subjects of qualitative research, do we have any control over who or what has access to aspects of the archived research?
- As qualitative researchers, how do we selectively and specifically allow and/or disallow access to parts (any or all) of the archived research data?
- As interested third parties, what access do we have to any given research data repository? May greater or altered access be requested?
- As users of research data repositories, how much anonymity is available? What attention to confidentiality is assured?

Role definitions and authorities are greatly complicated by the requirements of:

- Confidentiality
- Anonymisation
- Ownership and custodianship
- Selective access on individual, group, institutional or any other domain levels

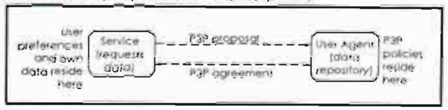
The figure following alludes to two of the areas of procedural and ethical dilemma faced by architects of qualitative research repositories with target audiences covering a wide range of roles and types of data interest:

Figure 0 : Authorities and roles



From P3P to access mediation of archived qualitative research data

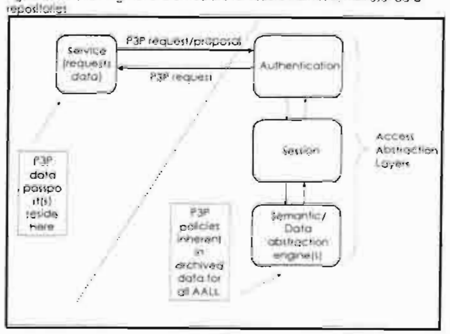
Figure 1 : Service and User Agent as P3P black boxes (Conventional application of P3P for website policy advertisment and acceptance)



P3P is aimed primarily at open advertisment of website data privacy policies.

Access mediaion and control requirements of qualitative research data are significantly complicated by the requirements for strict observance of confidentiality, of anonymisation, of ethical use and the need to allow distinction and control of custodianship and ownership of data to be effected in a decentralised manner Figure 2 diagrams an access mediation model based on P3P/RDF with bilateral exchange of P3P-based policies and data authority.

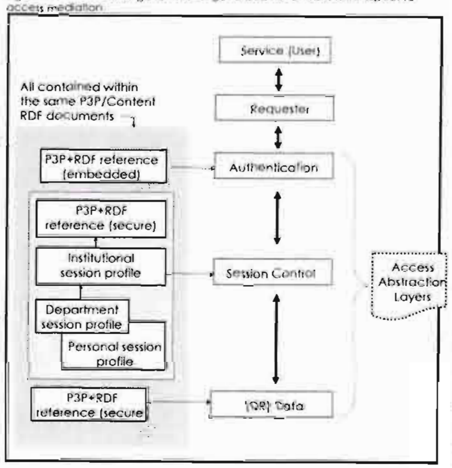
Figure 2 : P3P/RDF High-level architecture for secure multilevel, multiuser data repositories



Profiles, personae and other authentication data included in a P3P-based 'data passport' refer both passively and actively to the embedded policies and authorisations embedded as P3P in the document matrix making up the (qualitative) data repositories.

Where there is no P3P data passport available or offered, access defaults to a maximum of whatever unrestricted public access is allowed by any part of a repository.

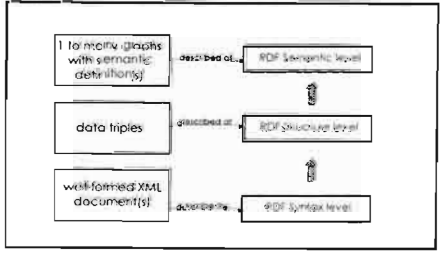
Figure 3 : P3P embedding / functional generalisation of data with respect to access mediation



RDF provides the basis for embedding of simple to complex structures, semantic networks and elements.

It also forms part of the basis of P3P policies and passports as proposed in this poster.

Figure 4 : RDF as semantic support for embedding of 'extended meaning' in archived data



From data vehicles to the meaning of it all ...

Metadata (or "data about data") can be used to label and categorise data for searching and processing by computer. A formal alternative is "structured descriptions of resources". Metadata form the matrix enabling large collections of data to function as organised libraries which seldom exist as single instances.

P3P uses RDF schema defined according to requirements of the environment. A proposed RDF schema is available on the w3c.org website

Resource Description Framework (RDF) data consists of nodes and property/value pairs describing nodes. A node is any object which can be pointed to by a URI. Properties are attributes of nodes, values are either atomic values for the attribute or other nodes.Information about a research topic (a node), may include the property "Owner". The value for the Owner property may be a string of text, a URI pointing to another document or a persona definition. RDF defines metadata processing frameworks and data models based on triples (subject/resource, predicate/property, object/property value). Data graphs with unique identifiers may be formed with these data triples. RDF forms the basis of tools able to link, classify and extend data and add subjective value. An example is the aggregation of a collection of XML documents into an RDF model. Document collections may be complete and fully formed, they may be data fragments and they may also be networks of multiply-linked XML documents. This forms the essential basis of RDF/XML used as dynamic and extensible repositories. Semantically-dependent queries against knowledge encoded in an ontology are available via RDF/XML document networks.

eXtensible Markup Language (XML) is an appropriate medium for metadata because it is widely understood and processable. XML provides a facility to define tags and the structural relationships between data or metadata forming a syntactic (and later, a semantic) tree. There is no predefined tag set and no prescribed semantics. The semantics of an XML document are defined by the XML itself. Interpretive semantic operations are imposed upon the data and are not inherent in it. XML is a natural format for transmitting metadata independent of platform and application. Goldfarb (2000) succinctly defines XML as *"XML data is smart data HTML tells how the data should look, but XML tells you what it means but XML data isn't just smart data, it's also a smart document and you don't have to decide whether your information is data or documents; in XML, it is always both at once You can do data processing or document processing or both at the same time".*

URI (Uniform Resource Identifier), the W3C's codification of the name and address syntax of present and future objects on the Internet. URI is the umbrella term for URNs, URLs, and all other Uniform Resource Identifiers.

Sources and references

1. Carmichael, P. (2002, May). "Extensible Markup Language and Qualitative Data Analysis". Forum Qualitative Sozialforschung (Online Journal), 3(2), online at <http://www.qualitative-research.net/fqs/fqs-eng.htm> (last accessed June 10th 2003)
2. Corti L. Day A. Backhouse G. (2000, December). "Confidentiality and Informed Consent: Issues for consideration in the preservation of and provision of access to qualitative data archives". Forum Qualitative Social Research (On-line Journal), 1(3), online at <http://qualitative-research.net/fqs/fqs-eng.htm> (last accessed May 23rd 2003)
3. Dublin Core Metadata Initiative, website at <http://dublincore.org/about/> , (last accessed 5th January 2004)
4. Fischer W.J., (2001, June), "Knowledge Reuse: The roles of human and technical intermediaries", Thesis submission to Faculty of the Graduate School of Arts and Science, Georgetown University, online at <http://cct.georgetown.edu/thesis/BillFischer.pdf>, (last accessed August 28th 2003)
5. Goldfarb, C.F. (2000), "XML in an Instant: A non-geeky introduction", website at <http://www.xmlhandbook.com/press/nongeeky.htm> , (last accessed 11th September 2003)
6. Mizoguchi R., Ikeda M., "Towards Ontology Engineering", The Institute of Scientific and Industrial Research, Osaka University, 567 Japan, Technical Report AI-TR-96-1, I.S.I.R., Osaka University, online at <http://www.ei.sanken.osaka-u.ac.jp/pub/miz/miz-onteng.pdf>, last accessed July 7th 2003).
7. Platform for Privacy Preferences (P3P) Project, website at <http://www.w3.org/P3P/> , (last accessed 6th January 2004)
8. Roberts K.A., Wilson R.W., (2002, May). "ICT and the Research Process: Issues Around the Compatibility of Technology with Qualitative Data Analysis". Forum: Qualitative Social Research (On-line Journal), 3(2), online at <http://www.qualitative-research.net/fqs/fqs-eng.htm> (last accessed May 23rd 2003)