Defining Interoperability Standards

A case study of Public Health Observatory Websites

Richard Dean

MSc Thesis

Abstract Page

The Association of Public Health Observatories (APHO) is a group of region-based health-information providers. Each PHO publishes health-related data for their specific region. Each observatory has taken a national lead in one or more key health area – such as 'cancer' or 'obesity'. In 2003, a project was initiated to develop 'interoperability' between public health observatory websites, so the national resources published by one lead observatory could be found on the websites for each other PHO.

The APHO interoperability project defined a set of requirements for each PHO – websites should comply with the current government data standards and provide webservices to allow data to be searched in real-time between different PHOs.

This thesis describes the production of an interoperable website for the North East Public Health Observatory (NEPHO) and the problems faced during implementation to comply with the APHO interoperability requirements. The areas of interoperability, e-Government and metadata were investigated specifically in suitability for NEPHO and an action list of tasks necessary to achieve the project aims was drawn up.

This project has resulted in the successful introduction of a new NEPHO website that complies with the APHO and e-Government requirements, however interoperability with other organisations has been difficult to achieve. This thesis describes how other organisations approached the same APHO interoperability criteria and questions whether the national project governance could be improved.

Acknowledgements

For their help in the production of this thesis and the work contained within I would like to thank the following people:

From the Department of Computer Science, Durham University – Dr Elizabeth Burd and Professor Keith Bennett

From the Centre for Applied Social Sciences, Durham University – Professor Gyles Glover and Di Barnes

From the North East Public Health Observatory – Professor John Wilkinson and Kathryn Bailey

For all their love and support through everything I have ever done – my parents – Gillian and Peter Dean, and my great auntie – Win Harvey.

Richard Dean

November 2005

Table of Contents

List of Figures

List of Tables

Declaration

This thesis has not been submitted for a degree at Durham University, or any other institution. The research conducted in the thesis is the work of the author except where indicated otherwise.

List of Acronyms

CORBA......... Common Object Request Broker Architecture

DCOM........... Distributed Component Object Model

DSOM........... Distributed Systems Object Model

e-Gov............. Electronic Government – Generic term covering government IT projects

e-GMS........... Electronic Government Metadata Standard

e-GIF .... Electronic Government Interoperability Framework

FTP................ File Transfer Protocol

GCL............... Government Category List – a controlled vocabulary for keyword tagging resources.

HealthPromis. Thesaurus used by HDA (superseded by NPHL)

HTML .... HyperText Markup Language

HTTP............. HyperText Transfer Protocol

NPHL .... National Public Health Language

OWL .... Web Ontology Language

PHITS .... Public Health Information Tagging Standard (superseeded by NPHL)

PHORMS ...... Public Health Observatory Resource Management System

RDF............... Resource Description Framework

RMI............... Remote Method Invocation

SOAP .... Simple Object Access Protocol

SMTP .... Simple Mail Transfer Protocol

UDDI............. Universal Description, Discovery and Integration

URI / URL..... Uniform Resource Identfier / Locator

WSDL .... Web Service Description Language

WWW .... World Wide Web

WYSIWYG... What You See Is What You Get

XML.............. Extensible Markup Language

List of Organisations

APHO............ Association of Public Health Observatories

CMO .... Chief Medical Officer

DoH............... Department of Health

EMPHO......... East Midlands Public Health Observatory

ERPHO .... Eastern Region Public Health Observatory

HDA.............. Health Development Agency – Merged with NICE in 2004

IBM............... International Business Machines

INISPHO....... Ireland and Northern Ireland's Population Health Observatory

LHO .... London Health Observatory

NCSA............ National Centre for Super computer Applications at the University of Illinois

NEPHO .... North East Public Health Observatory

NICE .... National Institute for Health Care and Clinical Excellence

NWPHO........ North West Public Health Observatory

OMG .... Object Metadata Group

PCT .... Primary Care Trust – There are 302 PCTs providing health care across the United Kingdom

PHO............... Public Health Observatory

SEPHO.......... South East Public Health Observatory

SWCIS .... South West Cancer Intelligence Service

SWPHO......... South West Public Health Observatory

WCH .... Wales Centre for Health

WMPHO .... West Midland Public Health Observatory

YHPHO......... Yorkshire and the Humber Public Health Observatory

Chapter 1 Introduction

1.1 Project History

The Government's 2000 health whitepaper '*Saving Lives*' [1] called for "a clearer national picture of health and health inequality so that we can track changes over time". As a result, a Public Health Observatory (PHO) was created in each English region. In February 2000, eight PHOs were established, with a ninth created in 2002 as a result of a NHS regional realignment. *Saving Lives* recognised that "many agencies are involved in collecting and using information about health and disease in the population – yet in some areas this data may not be available or may be unreliable". PHOs were tasked with drawing together information from different sources in new ways to improve health. The main activity of the PHOs is to obtain, use and then disseminate information, primarily reports and data but also about people, organisations and events [2]. Each observatory takes a national lead in one or more specific key health areas – for example, the North East Public Health Observatory (NEPHO) holds the national lead in mental health, consequently it publishes data about mental health for the whole of England. A problem faced by each observatory is in disseminating information beyond their region – while a mental health report for England could be made available on NEPHO's website, a user visiting a different observatory's website would be unaware of the resource.

1.2 Objectives

In January 2003, PHOs agreed to adopt electronic government standards for interoperability and metadata. If users could find, via their local website, information stored by any PHO, or even that a wider range of organisations hold on their websites, PHOs could increase dramatically the health intelligence available to their users. PHOs were commissioned by the Department of Health to develop methods to share information about resources between each observatory's website. Standards developed should be suitable for implementation within other healthcare information providers, such as the DoH, HDA and PCTs.

1.3 Scope of the thesis

This thesis will describe work undertaken by the Association of Public Health Observatories Interoperability Group up to June 2005. The North East Public Health Observatory is used as the main case study organisation, although solutions implemented by other PHOs are also described. The thesis describes the need for wider interoperability between PHOs and other healthcare information providers, however the implementation of such solutions are beyond the scope of this thesis.

1.4 Aims of this thesis

This thesis is being written as the summation of two years work reviewing, summarising and investigating the technologies involved in implementing interoperability within the healthcare information domain. It has the following aims:

1. To investigate **interoperability**, **metadata** and standards developed for the UK's **electronic government** programme in suitability for use by NEPHO

2. To analyse the existing NEPHO website and create an action list of tasks necessary to develop a site that would comply with requirements defined by the APHO interoperability group

3. To design, implement and review the new interoperable NEPHO website.

4. To describe the approach taken by other PHOs to comply with the same requirements, to develop interoperability with these PHOs and to critically examine the successes and failures of the project at a national level.

1.5 Thesis Structure

This thesis is split into chapters, each of which are referenced and numbered in the index. The chapters are organised as follows:

Chapter Two describes the formation of the internet and world wide web as a technology that led to the development of high-speed computer networks. The chapter describes common standards that facilitate web-services which can be used as a basis for developing interoperability between organisations. The chapter concludes by describing the concept of 'the semantic web' – a project that intends to create a universal medium for information exchange by giving meaning (semantics), in a manner understandable by machines.

Chapter Three describes a background survey of Electronic Government (e-Gov) policy in the UK – including the mandatory Electronic Government Interoperability Framework (e-GIF). The chapter describes the electronic Government Metadata Standard (e-GMS) in detail, along with describing different types of controlled vocabulary.

Chapter Four gives a more detailed look at the Public Health Observatory interoperability project. The North East Public Health Observatory will be used as a case study. An analysis of their existing website is provided, resulting in an action list of tasks to make the NEPHO website interoperable.

Chapter Five describes both the design and implementation of a new website that complies with electronic government initiatives and will be capable of interoperating with other organisations.

Chapter Six provides an overview of the project status by June 2005, and describes the work done by other PHOs in connection with the national interoperability project

Chapter Seven examines progress in redeveloping the NEPHO website and developing interoperability by June 2005.

Chapter Eight draws conclusions about the project to-date, and outlines future plans for developing interoperability with PHOs and beyond.

Chapter 2 Technical Interoperability

2.1 Introduction

Interoperability is the ability to share information seamlessly between a variety of organisations. It is a government priority in the UK, Europe and internationally [3].

This chapter describes a number of important milestones in computer communication technologies. In the early years of computing, it was acknowledged that it was far more efficient for applications to perform different tasks on specialist computers – a database server, for example would be much more efficient for its particular task than a generic computer that did not specialise in a specific function. This gave birth to the concept of distributed computing (Chapter 2.2). This was a significant step forward as it allowed computers to start communicating on a large scale, however there were problems. The rush to develop distributed computing led to a number of competing standards for communications between each component in a distributed computer system. A business could not simply integrate applications with their business partners because of the variety of incompatible standards.

Research into different computer networking techniques eventually led to the wide-spread adoption of packet switching networks and the adoption of a common inter-networking protocol named TCP/IP. The ability of TCP/IP to work over pre-existing networks allowed for ease of growth, and paved the way for the development of the internet. The internet and the World Wide Web (Chapter 2.3) can be ranked as one of the most important and profound inventions of the twentieth century [4]. Applications have the ability to talk to hundreds of millions of computers. The Web has the ability to expedite and simplify our work – it has changed business focuses and people's personal lives. Web services operating over the internet and web-based protocols like HTTP extend the web's capability to include communication between computer programs. The reminder of Chapter Two consists of a series of sections which outline such relevant technologies and the relevant background to the subject of the thesis.

Chapter 2.4 describes the use of XML to store data. Chapter 2.5 describes the principle of caching information available via the Internet and World Wide Web. Chapter 2.6 takes an in-depth look at web services – a method to send XML-based messages between computer systems via the Internet, thus achieving interoperability without the necessity to specify low-level communications protocol. Finally, Chapter 2.7 provides a brief overview about the 'Semantic' web – a research initiative to create a universal medium for information exchange by providing meaning to the content of documents available on the Web.

2.2 Distributed computing

It's often more efficient for large applications to perform different tasks on different computers. With the emergence of powerful computers and networks, distributed computing became a widespread phenomenon. An organisation's computing capability is split over networks of specialised computers, rather than a single central computer. N-tier applications split their computing power over a number of computers. For example a three-tier web based application may rely on a database computer server to collect data which is passed to a web server computer to create a web page which is passed to a user's computer to display the website in the user's web browser.

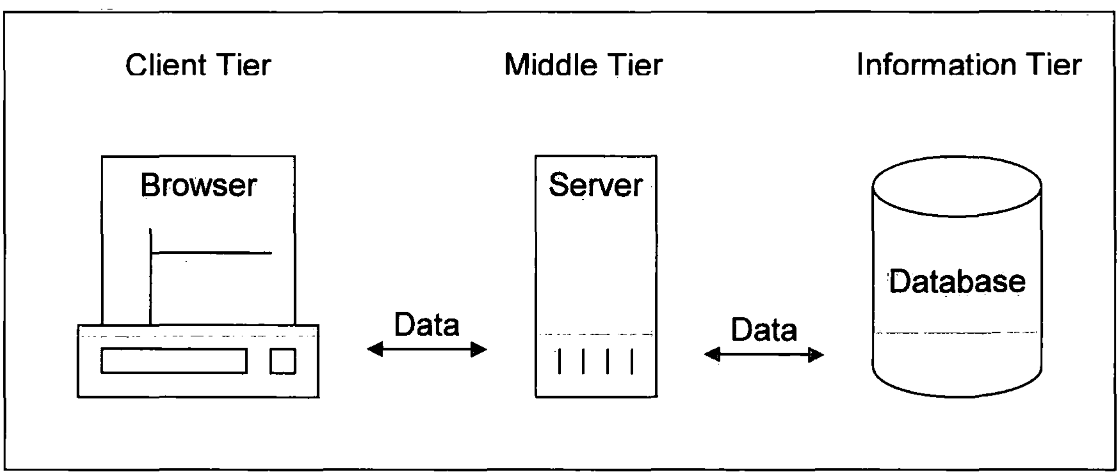


Figure 1 Three-tier Web based application (From [4])

For a distributed system to function, components on different computers in a network must be able to communicate. In the early 1990s a number of organisations saw the need for this functionality and developed their own solutions to enable communication between distributed computers. The Common Object Request Broker Architecture (CORBA) is OMG's specification for developing interoperability between distributed computing nodes. CORBA "wraps" code into a bundle containing additional information on the capabilities of the code inside, and how to call it. The resulting wrapped object can be called from other programs (or CORBA objects) over a network [5,10]. Similar specifications were created by rival organisations: Sun created the Remote Method Invocation (RMI) [6], Microsoft developed the Distributed Component Object Model (DCOM) [7,8] and IBM launched the Distributed Systems Object Model (DSOM) [9].

The emergence of these new technologies was significant – each technology allowed programs running in different locations to communicate as if they were running on one computer. Businesses could integrate applications with their business partner's computing systems. While distributed computing was possible using a single technology, interoperability between the competing standards remained difficult. For example, DCOM and CORBA can't communicate easily – a COM/CORBA bridge is needed. If either underlying protocol changes, the bridge needs updating. Such competing standards impeded the ability to integrate components of distributed computing. A common internetworking protol was required to facilitate business process integration and automation. Such a protocol was offered by TCP/IP and similar technologies developed to support the Internet, as described below.

2.3 The Internet and World Wide Web

The Internet and the World Wide Web are not synonymous: the Internet is a worldwide, publicly available collection of computer networks that transmit data by packet switching using the standard Internet Protocol (IP). The World Wide Web is a collection of interrelated documents and resources addressable by URLs. The Web is accessible via the internet, as are many other services such as email, file sharing, data collaboration and so on [85,86].

Research into various networking technologies began at the US Department of Defence in the 1950s. Packet switching was chosen for being highly robust and reliable. In the early 1970s, researchers decided that a common internetwork protocol was required to hide the differences between individual network protocols. This reduced the role of the network to the bare minimum and it became possible to join any networks together, no matter what their characteristics were. The package of standards developed included the Transmission Control Protocl (TCP) and the Internet Protocol (IP). Combined, these form the Internet Control Suite (TCP/IP) [86]. TCP/IP-based networks started to appear from January 1983. The ability for TCP/IP to work over pre-existing communication networks allowed for great ease of growth.

In the early 1990s, Tim Berners-Lee began creating HTML, HTTP and the first web pages. The 'World Wide Web' project was publicised by CERN on August 6th 1991 [13,85]. The first popular web browser, Mosaic was developed in 1993 at the National Center for Super Computer Applications (NCSA) in the University of Illinois [11, 12, 13]. Within a year more than two million users were browsing the web. By June 1997, 960,000 people in the UK alone had access to the internet [14]. This figure quickly rose to 4.3 million by March 1998 and has continued to climb rapidly (Figure 2).

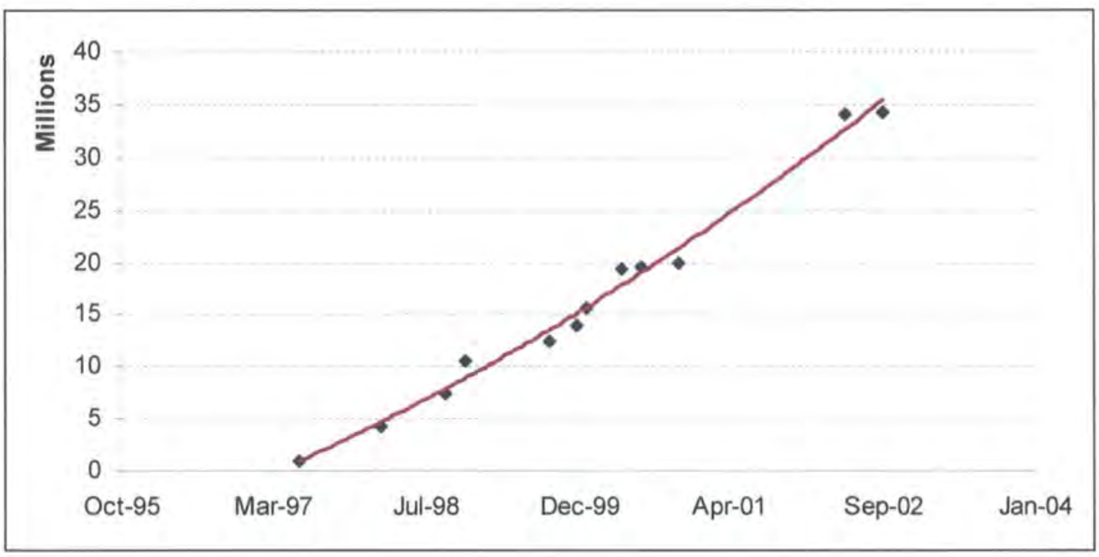


Figure 2 Number of UK citizens online Source: [14]

Over the course of the 1990s, the Internet successfully accommodated the majority of existing public computer networks [85]. This growth has been attributed to the lack of central administration coupled with the non-proprietary open nature of the internet protocols, which encourage vendor interoperability and prevents one organisation from excreting too much control over the network.

In the following sections, we will look at some of the technologies developed for use over the Internet.

2.4 XML

Extensible Markup Language (XML) is a standard for document markup. It defines a generic syntax used to mark up data with simple, human-readable tags and a standard format for computer documents that is flexible enough to be implemented on a wide variety of domains [16].

Data are included in XML documents as strings of text. The data are surrounded by text based markup that described the data. The basic unit of data in XML is called an 'element'. The XML specification defines the exact syntax this element should follow: how elements are delimited by tags, what a tag looks like, what names are acceptable for elements, where attributes are placed, and so forth [16].

```
<PERSON person_id= "100" sex= "M">
  <person_name>
    <given_name>Richard</given_name>
    <surname>Dean</surname>
  </person_name>
  <organisation>University of Durham</organisation>
  <country>England</country>
  <contact_details>
    <email>richard.dean@durham.ac.uk</email>
    <phone>+44 191 334 1479</phone>
    <phone>+44 191 334 0405</phone>
  </contact_details>
</PERSON>
```

Figure 3 An example of an XML document with metadata tags (surrounded by <...>) identifying the meaning of following data. (Adapted from [17])

2.5 Data Caching

In real life, people often store important information such as telephone numbers for their friends and colleagues. Directory enquiries services can provide people with contact details for those not already listed. Such services cost money, and therefore most people prefer to store the number in an address book for future reference. This is an example of caching data – storing a local copy of information to speed up retrieval time and reduce the cost.

In computer science, a cache is a collection of data values stored elsewhere or calculated earlier, where the original data are expensive (usually in terms of access time) to fetch or compute relative to reading the cache. Once the data are stored in the cache, future use can be made by accessing the cached copy rather than refetching or recalculating the original data, so that the retrieval time is lower [15].

Caching is widely used in computing for many processes ranging from CPU processes to speeding up access to data on hard disks, and is frequently used in accessing the internet, where files such as web pages and associated images are stored on local file systems. In the Public Health Observatory interoperability project, data caching could be used to periodically store a copy of other organisations databases. This approach is discussed in more detail in later chapters, specifically Chapter 7.7.1

2.6 Web services

Web services represent the next phase of distributed computing, building on the work established by the previous technologies such as DCOM and CORBA [18]. Architects of each new technology borrowed the concepts and ideas from the previous generations, making the task of developing distributed applications easier. However, each technology still lived, for the most part in its own world, and the interoperability between different enterprise infrastructures was difficult [18]. Web services have two major goals – to make distributed computing easier for the business programmer and to enhance interoperability [18]. These goals are aided by the loose coupling between the requesting program and the service provider, and the use of XML, which is platform and language neutral.

The definition of a web service varies a lot. Beznosov defines Web services as an *XML-based messaging interface to computing resources that is accessible via Internet protocols, in particular the Simple Object Access Protocol (SOAP)* [18]. IBM provide the following description [19]:

> *Web services are self-contained, self-describing, modular applications that can be published, located, and invoked across the Web. Web services perform callable functions that can be anything from a simple request to complicated business processes. Once a Web service is deployed and registered, other applications can discover and invoke the deployed service.*

Web services are not limited to the Hypertext Transfer Protocol associated with the web – they may use a variety of other transport protocols such as FTP and SMTP. The use of existing protocols means the deployment of web services can be delivered with little change to existing infrastructures – use of standard internet protocols means that most organisations already have the communication software and infrastructure in place to support web services. Existing development environments and languages can be used.

Web services improve distributed computing capabilities and use open standards. This means web services can communicate regardless of the platforms or technologies that created each component. The text-based communications protocols employed by web services make debugging easier than the binary communications protocols of DCOM/CORBA.

2.6.1 Advantages of Web Services

Web services have many advantages over earlier attempts at cross-domain interoperability. Web services have a number of characteristics that set them apart from solutions that came before them and make web services more likely to succeed. Besnosov [18] lists the following advantages:

- Active support by major software vendors such as Microsoft, IBM and Sun.
- Loosely coupled processing. Earlier attempts at interoperability assumed a common application environment at both ends of a transaction. Web services allow the subscriber and provider to adopt the technology that is most suited to their needs.
- XML provides a flexible model for data interchange that is independent of the computing environment.

Use of Internet standard protocols means that most organisations already have much of the communications infrastructure needed to support web services. Few new protocols need to be supported, and existing development environments can be used.

2.6.2 Web service building blocks

The web services framework is divided into three areas – communication protocols, service descriptions, and service discovery [20]:

- **Communication protocol** – The Simple Object Access Protocol (SOAP) enables communications between web services;
- **Service Description** – The Web Service Description Language (WSDL) provides a formal, computer readable description of web services;
- **Service Discovery** – The Universal Description, Discovery and Integration (UDDI) directory that is a registry of web service descriptions.

These are described in detail below.

2.6.2.1 Communication Protocol – SOAP

SOAP is a XML-based protocol specification that defines a uniform way of passing XML-encoded data. It provides a common data format for exchanging information between independent data sources and platforms. It also defines a way to perform remote procedure calls (RPCs) using HTTP as the underlying communication protocol [3,21].

2.6.2.2 Service Description – WSDL

The Web Services Description Language (WSDL) provides a way for service providers to describe the basic format of web service requests over different protocols or encodings. WSDL is used to describe *what* a web service can do, *where* it resides, and *how* to invoke it [21].

A WSDL document uses the following elements in the definition of network services [21]:

- Types – a container for data type definitions using some type system.
- Message – an abstract, typed definition of the data being communicated.
- Operation – an abstract description of an action supported by the service.
- Port type – an abstract set of operations supported by one or more endpoints.
- Binding – a concrete protocol and data format specification for a particular port type.

- Port – a single endpoint defined as a combination of a binding and a network address.
- Service – a collection of related endpoints.

2.6.2.3 Service Discovery – UDDI

UDDI provides a mechanism for clients to find where specific web services are provided, and who provides them [3,21]. Business can dynamically connect to services provided by external business partners by using a UDDI interface. A UDDI registry has two kinds of client – businesses that want to publish a client, and clients who want to obtain a service of a certain type and bind.

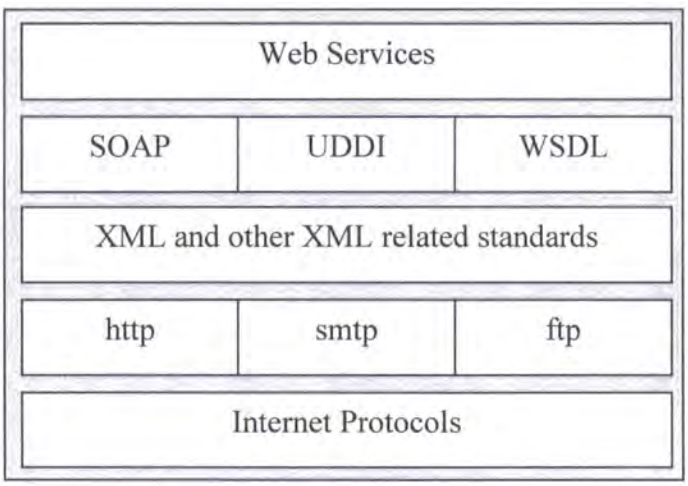


Figure 4 Web service building blocks [3]

SOAP, UDDI and WSDL are the underlying technologies upon which Web services are based. Using these formats and protocols (shown in Figure 4), systems from different domains, independent environment, or with different architectures can engage in a cooperative manner to implement business functions.

2.6.3 Reservations about web services

There are two reservations about web services. First, that web services are not a proven technology; some view web services with a suspicion that they are the 'latest fad', and like many other promised solutions, they will not deliver. While this cannot be disproved, web services have significant advantages over past solutions.

The second reservation about web services centres around its reliance on XML. Firstly, use of XML expands the size of data several times over, increasing storage requirements. Spanning processing domains requires a flexible representation. Improvements in the amount of data storage available and increased speed of computer networks have offset this problem to some extent. Once a message is within a single environment, implementers must decide the extent to which XML is required [18]. Secondly, the use of XML does not solve the semantic mapping problem (see Chapter 3.4.1). There are lots of potential ways to represent the same entity in different XML schemas. One important research initiative aimed at addressing semantic differences is termed the Semantic Web. It is described in more detail in the following section.

2.7 The Semantic Web

The 'Semantic Web' is a research initiative that intents to create a universal medium for information exchange by giving meaning (semantics), in a manner understandable by machines, to the content of documents on the Web [22]. To date, the Web has developed as a medium for documents for humans rather than of information that can be manipulated automatically. Computers can parse Web pages for layout and routine processing – headers and links to other pages are adeptly processed – but in general, computers have no reliable way to process the semantics [23]. This section provides an overview of the need for the Semantic Web, and explains that the work done in this thesis will help NEPHO build a foundation layer for future semantic web developments.

Keyword-based search engines, such as Altavista, Yahoo, and Google are the main tools for using today's Web. Serious problems associated with their use include [24]:

- High recall, low precision. Search engines often recall thousands of pages which match search terms. Even if the main relevant pages are retrieved, they are of little use if they are buried amongst thousands of mildly relevant or irrelevant documents also returned.
- Results are highly sensitive to vocabulary: Often our initial keywords do not get the desired results; in these cases the relevant documents use different terminology from the original query. This is unsatisfactory because semantically similar queries should return similar results.

- Results are single web pages. If we need information that is spread over various documents, we must initiate several queries to collect the relevant documents, manually extract the partial information and put it together.

The main obstacle to providing better support to web users is that, at present, the meaning of Web content is not machine-accessible. Tools can be developed to retrieve texts, split them into parts, check the spelling, and count their words. But when it comes to *interpreting* sentences and extracting useful information for users, the capabilities of current software is still very limited [24].

Despite improvements in search engine technology, the difficulties remain the same. It seems that the amount of Web content outpaces technological progress. One solution to improve the current situation is to use the content as represented today and to develop increasingly sophisticated techniques based on artificial intelligence and computational linguistics. This task has been followed for some time, but the task still appears too ambitious [24].

An alternative approach is to represent Web content in a form that is more easily machine-processable and to use intelligent techniques to take advantages of these representations.

The Semantic Web's aim is to bring structure to the meaningful content of Web pages, thereby creating an environment where software agents roaming from page to page can readily carry out sophisticated tasks for users. Berners-Lee provides the example of a physical therapy service – a webpage may be marked up so that a human can view details about the service, but also a semantic web agent can understand the semantics of the data. In Berners-Lee's example, the agent will know that the page not only has keywords such as "physical therapy, medicine, treatment", but that a doctor "*works* at this *clinic* on *Mondays*, *Wednesdays* and *Fridays* and that the script takes a *date range* in *yyyy-mm-dd format* and returns *appointment times*" [23].

2.7.1 Layered Approach

The development of the Semantic Web is predicted to proceed in steps, each step building a layer on top of another.

Development of the full semantic web may take anther ten years to be realised to its full extent (as envisioned today, of course). Two principles should be followed when building a layer of the semantic web:

- Downward compatibility – agents fully aware of a layer should also be able to interpret and use information written at a lower level.
- Upward partial understanding – agents fully aware of one layer should take at least partial advantage of information at higher levels.

Figure 5 shows the "layer cake" of the semantic web, which describes the main layers of the Semantic Web and vision.

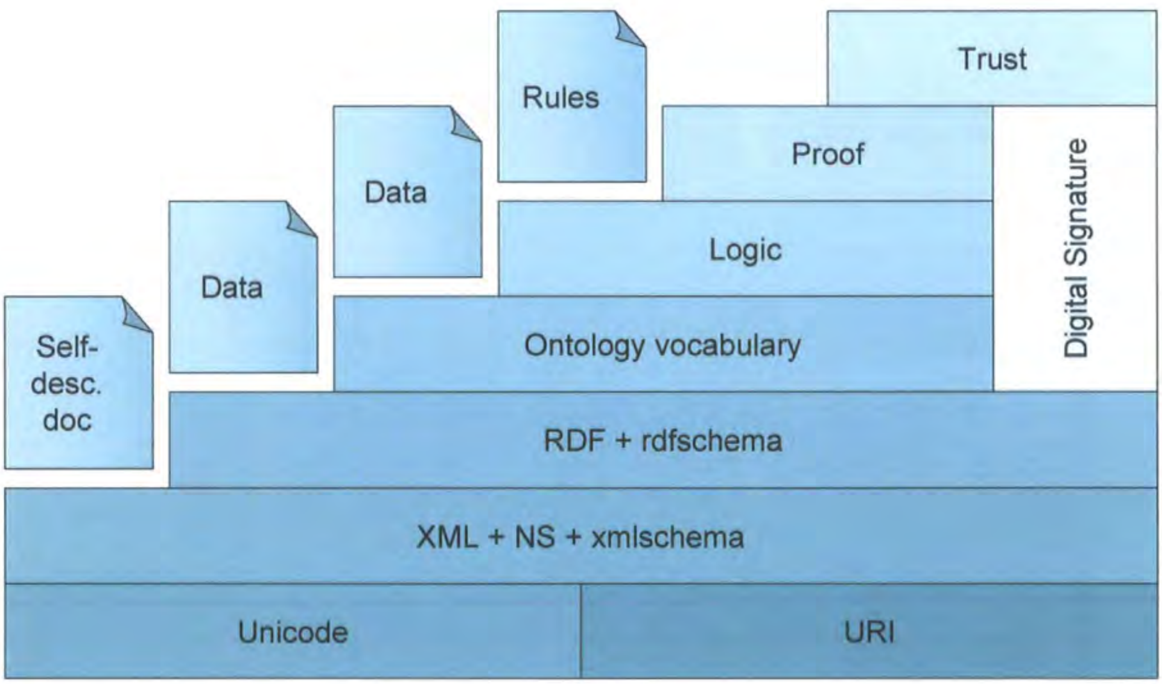


Figure 5 A layered approach to the semantic web [24]

The foundation layers of the semantic web are based upon XML (described in section 2.4), a language that lets one write structured web documents with a user-defined vocabulary. XML is particularly suited for sending documents across the web.

RDF is a basic data model for writing web objects. The RDF model does not rely on XML, but RDF has an XML-based syntax. The Web Ontology Language (OWL) allows the expression of ontologies, which define the meaning of terms used in RDF statements [25]. Ontologies are described in more detail in section 3.5.4.

RDF Schema provide modelling primitives for organising Web objects into hierarchies. Key primitives are classes and properties, subclasses and sub property relationships, and domain and range restrictions. RDF Schema can be viewed as a primitive language for writing ontologies, but there is a need for a more powerful ontology layer to represent the complex relationships between Web objects.

The logic layer is used to enhance the ontology layer and to allow the writing of application-specific declarative knowledge.

The proof layer involves the actual deductive processes as well as the representation of proofs in Web languages and proof validation.

Finally, the Trust layer will emerge through the use of digital signatures and based on recommendations by trusted agents. Trust is a high-level and crucial concept for the development of the semantic web. The web will only achieve its full potential when users have trust in its operations and in the quality of information provided [24].

2.8 Summary

This chapter has investigated some important concepts of interoperable computer systems. There are many benefits to developing distributed systems where dedicated machines can be built to perform specialised functions. A number of differing standards such as CORBA and DCOM (section) were developed to allow interoperability between distributed computing nodes however interoperability between competing standards was difficult (Chapter 2.2).

Section 2.3 described the development of TCP/IP – the Internet Standards Protocol and the phenomenal rise of the Internet and the World Wide Web. This aided interoperability because it used a common set of protocols to join computers to a single distributed world-wide network.

The chapter continued by discussing XML as a data format for sending information between different computer systems in a platform-independent, easy to understand format. Data caching was briefly discussed as a technology to reduce costs associated with fetching data from remote systems. Section 2.6 described Web services, an XML-based messaging interface to computer resources that is accessible via Internet Protocols, in particular SOAP [18].

Interoperability is the ability to seamlessly share information between computer systems. Previous attempts at developing interoperability between computers have been difficult due to competing communication models.

Web services facilitate information sharing over computer networks such as the internet by using open standards such as XML. Web services are based on three concepts – SOAP, WSDL and UDDI.

Finally, section 2.7 described the Semantic Web, which is an initiative that aims at improving the current state of the internet through the use of machine-processable information. Building the semantic web will proceed in layers. It may take another decade to reach the 'trust' layer. Key technologies include explicit metadata and ontologies.

Chapter 3 Electronic Government

3.1 Abstract

Electronic government (e-Government) encompasses a wide range of services: dissemination of information, commerce with the private sector, services to individual citizens and businesses, and participatory democracy [27]. The Parliamentary Office of Science and Technology uses the term e-Government to cover projects that exploit new technologies to modernise government and secure benefits for citizens and businesses [26]. Chapter 3.2 describes early work behind this new branch of government.

By 1998 there were 300 local government websites in the UK, however there were few standards to categorise the information available. The continued growth in the number of websites being produced to represent governmental organisations saw the clear need to develop standards. These included an interoperability framework (e-GIF), described in Chapter 3.3 and the adoption of a metadata standard (e-GMS), described in Chapter 3.4 to ensure that resources are consistently classified across different domains.

A mandatory part of the e-GMS is the use of keywords to classify resources. Chapter 3.5 describes different types of controlled vocabularies. Every resource must be classified by one or more keyword from the Government Category List (GCL), which is a taxonomy consisting of 420 preferred terms covering the entire range of government activities from agriculture to science. While a good starting point, the GCL is too broad to accurately classify all the resources published by a public health observatory, therefore a more specific health-related vocabulary is required.

3.2 Background and Introduction to e-Government

In the mid 1990s, government agencies began publishing strategies that could take advantage of the capabilities offered through the internet. Early research highlighted clear benefits to developing e-Government, for example, in the United States, Chenery [28] estimated that for every government transaction that could be moved online, a potential cost saving of $400 could be achieved. In the UK, *The National Strategy for local egov* [29] estimated that electronic procurement could reduce procurement costs by about two thirds, whilst simultaneously speeding up the process. The strategy paper

Two years on [30] estimates that "Councils expect to deliver £1.2 billion in efficiency savings by 2007/08 directly as a result of e-government investment.

Local and central government agencies were driven towards development and implementation of e-Government by the ability to reduce administrative and operational costs while simultaneously enhancing the services they offer to businesses, citizens and the general community at large.

Early information and policy strategies that were created were not fully understood by those who were supposed to implement them [27]. Between 1995 and 1998 the number of UK local government websites rose from 40 to 300 [31].

Early websites were static and focussed on the dissemination of information. The 1998 Competitiveness whitepaper [33] recognised the "extraordinary progress" in ICT, setting a target to deliver 25% of government services over the internet by 2002. This target was extended to include 100% of government services by 2008 in the 1999 whitepaper *Modernizing Government* [34]. This timescale was brought forward to a target of 100% of government services by 2005 [35, 36] as a result of European initiatives such as the Interchange of Data between Administrations (IDA) programme [36, 37, 38, 39, and 40] and the Prime Minister's aim of making the UK "the world's leading Internet economy" [35].

Since 1998, the number of government websites has spiralled to a number that is now unknown [27] – sparking calls from organisations such as the British Chamber of Commerce to reduce the number of government websites to a more manageable number [32].

In his forward to a 1999 Cabinet Office study, titled *e-commerce@itsbest.uk: A Performance and Innovation Unit Report,* the Prime Minister noted that, although the UK had world-class telecommunications technology and fairly good levels of Internet penetration into homes and workplaces, complacency was a danger: "There are signs we are not capitalizing on our strengths and keeping up with the pace of change," he wrote. "A recent survey of Directors in the UK showed that (only) 2 percent of UK Board Directors believe that the Internet poses a serious competitive threat. That cannot be right... ." The Prime Minister was aware of the opportunities and competitive risks represented by the internet. [85,88]

Subsequently, in September 1999, the UK government established the office of the e-Envoy with four principal work areas: e-Policy, Service Transformation, e-Delivery and e-Communications [89]. The e-Envoy developed a set of policies, technical standards and guidelines covering ways to achieve interoperability between public sector data, information and resources, information communications technology, and electronic business processes. The resulting electronic-Government Interoperability Framework is described below.

3.3 The Electronic Government Interoperability Framework

The Electronic Government Interoperability Framework (e-GIF) may be considered as a "highway code" but for computer systems, rather than road users. The e-GIF encompasses the following aims to:

- enable the seamless flow of information across government and public service organisations
- set practical standards using stable well supported products
- provide support, guidance and toolkits to enable the standards to be met
- provide a long term strategy that is able to accommodate and adapt.

The e-GIF comprises a framework of technical policies and specifications governing information flows across government and the public sector [3,41]. It defines policies which cover interconnectivity, data integration, e-services access and content management for the exchange of information between government systems and the interactions within UK Government and between UK Government and citizens, intermediaries, businesses (worldwide) and other governments. 'UK Government' includes central government departments and their agencies, local government and the wider public sector, such as the National Health Service [41].

The selection of policies and specifications has been driven by the need for interoperability, the promotion of openness, market support, scalability and international standards [41]. Only specifications that are relevant to systems' interconnectivity, data integration, e-services access and content management metadata are specified.

A number of key policies are outlined in the e-GIF. These policies ensure that all systems developed for a UK government agency share common standards in terms of interfaces, data architecture, communication protocols and metadata standards. Key policy decisions are outlined below [41]:

- alignment with the Internet: the universal adoption of common specifications used on the Internet and World Wide Web for all public sector information systems
- adoption of XML as the primary standard for data integration and presentation tools for all public sector systems
- adoption of the browser as the key interface; all public sector information systems are to be accessible through browser based technology; other interfaces are permitted but only in addition to browser based ones
- the addition of metadata to government information resources
- the development and adoption of the e-GMS (e-Government Metadata Standard, described in section 3.4.3 below) and the development and maintenance of the GCL (Government Category List)
- adherence to the e-GIF has been mandated throughout the public sector.

The e-GIF architecture (Figure 6) consists of the e-GIF framework and the e-GIF registry. The e-GIF framework covers high-level policy statements, technical policies and management, implementation and compliance regimes.

The e-GIF registry encompasses a range of standards, including the e-Government Metadata Standard (e-GMS) and Government Category List (GCL), the Government Data Standards Catalogue (GDSC), XML schemas and the Technical Standards Catalogue (TSC).

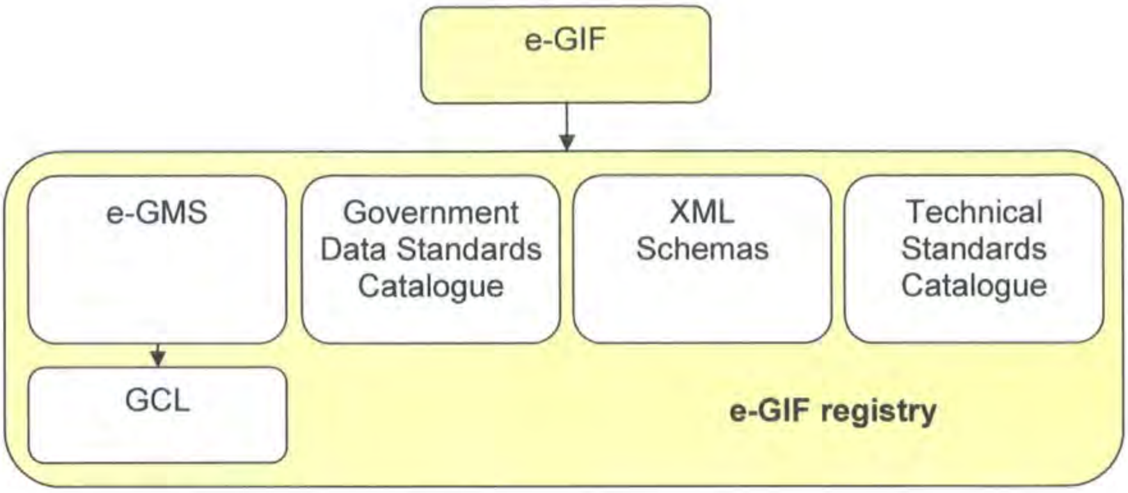


Figure 6 e-GIF architecture [41]

In summary, compliance with the e-GIF is mandatory across the public sector. e-GIF compliance requires a browser interface for access, XML as the primary means for data integration, and the use of the e-GMS for the basis of content management. While use of the e-GMS and the GCL is mandated through e-GIF, it is not expected that either standard will be implemented in full in every system – it is expected that local standards based upon e-GMS / GCL will be developed within differing domains. The reminder of this chapter takes a more detailed look at the development of metadata standards such as the e-GMS and the use of controlled vocabularies such as the GCL.

3.4 Use of metadata to classify resources

Chapter 2 investigated technical interoperability, especially the use of XML schema. There is a difference, however, between the ability to send and receive messages between two electronic systems and for those systems to be able to *understand* the messages received. This section considers how systems interpret this data – there are many ways to represent the same information. This is called a semantic mapping problem – how can two parties know what each other is communicating about and be able to share exact information.

The most commonly used definition of metadata is 'data about data' [47]. The Office of the e-Envoy describes metadata as 'structured information about a resource'. For example, a catalogue selling household items gives the metadata about those items – the brand, price, colour and capacity. Metadata enables a resource to be found by indicating what the resource is about and how it can be accessed with a series of structured descriptions [41, 47].

3.4.1 Constraints of semantic mapping

Graham et al [42] identifies the problems of semantic mapping in a non-technical way. He states that "If you send a letter in English to someone who only speaks Japanese, you won't be able to communicate without the help of a translator. If you send temperature readings in Celsius units to someone expecting them in Fahrenheit units, there is potential for confusion and error unless someone knows how to make the right conversion". This problem is rampant in the computer industry [42]. Because there is no clear separation between syntax and semantics, much of the effort in integrating applications is spent negotiating between different data formats. Graham et al realised that many companies deliver data-mapping tools and technologies, but they typically solve a point integration problem between two applications and do a poor job of dealing with the next application that becomes part of the integration. In the following section, two metadata standards – Dublin Core and e-GMS are described. Both of these standards aim to ensure that government applications use a consistent set of semantics for metadata, thus reducing the semantic mapping problem.

3.4.2 Dublin Core

The Dublin Core Metadata Initiative (DCMI) [43] traces its roots to Chicago at the second International World Wide Web conference in October 1994 [44]. A discussion on semantics and the web revolved around the difficulty of finding resources, even then, with only about 500,000 addressable objects on the web. A workshop to discuss metadata semantics was held in Dublin, Ohio in March 1995 [44]. People discussed how a core set of semantics for Web-based resources would be extremely useful for categorising the Web for easier search and retrieval. The result was dubbed "Dublin Core metadata" based on the location of the workshop. The Dublin Core Metadata Set consists of 15 optional metadata elements. This has since become a recognised international standard (ISO 15836) [43, 46].

The development of Dublin Core was important because it defined an internationally recognised metadata standard designed to specifically consider the rapidly expanding internet. It is used by governments worldwide as the basis of national standards – so even where different governments have developed national standards, many metadata elements can be tracked back to their Dublin Core routes.

The elements defined in Dublin Core have been adopted by many organisations, including the UK government who used Dublin Core as a basis for its own metadata element set.

3.4.3 e-Government Metadata Standard

The UK Government's Electronic Government Metadata Standard [41, 47, 48] defines the elements, refinements and encoding schemes to be used by government departments when creating metadata for information about resources or defining search interfaces. The e-GMS is needed to 'ensure the maximum consistency of metadata resources across public sector organisations' [41]. The UK first version of the government's metadata standard was originally laid out in the e-Government Metadata Framework (e-GMF) [49]. This used simple Dublin Core elements [43].

3.4.3.1 Why is metadata important?

Reasons for developing a metadata standard are laid out in e-GMS version three [47]:

- Improved use of official information, joined-up systems and policies, and services designed around the needs of citizens is called for by government whitepapers, such as 'Modernising Government' [34]
- Standardisation of government information systems will allow them to be accessed easily from central portals
- New systems for the handling of electronic records are being devised. Official records will not always be stored in paper format.
- Metadata makes it easier to manage or find information, be it in the form of web pages, electronic documents, paper files or databases.
- For metadata to be effective, it needs to be structured and consistent across organisations.
- The e-GIF is mandated across all government information systems. By association, so is the e-GMS.

3.4.3.2 e-GMS in practice

Dublin Core continues to be the cornerstone of the e-GMS. The e-GMS is viewed as an overall standard. It is unlikely when developing a system that it will require all of the elements listed. e-GMS expects that local standards are created based on a 'cut-down' version of e-GMS. One such cut-down version of e-GMS, called *e-GMS for websites* [48], has been released to help in the development of local metadata standards for websites. The elements in *e-GMS for websites* are listed in Table 1, below.

Mandatory	**Recommended**	**Optional**
Creator	Coverage	Audience
Date	Language	Contributor
Subject.Category		Description
Title		Disposal
Accessibility		Format
Identifier		Relation
Publisher		Rights
		Source
		Status
		Type

Table 1 List of e-GMS elements

e-GMS defines attributes for each element. These attributes are important for defining data types, cardinality, syntax examples and possible refinements for each element. Each attribute is described in Table 2, below.

Attribute	**Description**
Definition	The formal definition of the element, taken from Dublin Core wherever possible.
Obligation	These are: • **Mandatory**: this element must have a value • **Mandatory if applicable:** this element must be given if the information is applicable • **Recommended:** This element should be given if the data is available for the given resource • **Optional:** This element may be given if the data is available for the given resource. In e-GMS for websites, [48] non-applicable resources were removed and other elements which are marked as mandatory if applicable were made mandatory (for example, all resources on a website have a unique URI, but no physical location). The obligation applies to the element as a whole. Values can be added to the unqualified element, or to one or more refinements, except in the case of the Subject element, where the Category refinement is mandatory.
Purpose	Giving the purpose of the element, background information and other factors
Notes	Additional information which is considered useful to the element or element refinements.
Not to be confused with	Provides clarity over the appropriate use of the element. For example, guidance is provided to clarify the difference between similar elements such as *publisher* and *creator*
Refinements	Used to make the meaning of an element narrower and more specific. For example, the date element has a number of refinements such as date created, date issued, date copyrighted and so on.

Examples and HTML syntax	Examples of how the element could be completed are provided for a variety of resources, including the HTML syntax used in the header of a HTML file.
Encoding schemes include	Encoding schemes are used to regulate the value of an element. They include contextual information or parsing rules that help interpret a term value. These include controlled vocabularies which are described in section 3.5, below.
Mapping	Lists the elements in other metadata schemes that the element maps to, including Dublin Core and similar metadata schemes used by the American and Australian governments.

Table 2 Attributes of each e-GMS element [41]

3.4.4 Metadata summary

The e-GMS is the UK government standard for defining 'data about data'. All resources stored on electronic systems by any UK government body must use e-GMS as a basis for classification. By using a single standard throughout government, the semantic mapping problem is greatly reduced – for example, the standard defines a single way to represent concepts such as a date, which if left open to different systems developers could be represented in many different ways.

The e-GMS mandates that keywords are selected to classify resources. Keywords are used in the e-GMS 'subject' element. There are two refinements where keywords should be used:

- *Subject.category* is a mandatory requirement; at least one term from the Government Category List (GCL) must be used to describe each resource.
- *Subject.keyword* allows words or terms to be taken from other controlled vocabularies to accurately describe the resource.

Keyword classification is unique in the e-GMS because it uses a predetermined set of terms as opposed to plain text fields such as 'title', 'description' or 'author'. The use of keyword libraries is examined in more detail in the following section.

3.5 Controlled Vocabularies and Keyword Classification

A 'controlled vocabulary' refers to a list of terms or headings, where each has an assigned meaning [50]. It is designed for classifying, indexing, and searching information resources. The benefit of such an approach is that if everyone uses the same name for the same concept, things become much easier to find. Types of controlled vocabulary include taxonomies, thesauri and ontologies. Each concept is outlined below.

3.5.1 Taxonomy

A taxonomy is a structured list or 'tree', formed into a hierarchy with broader terms at the top. Ideally, each item (or taxon) in a taxonomy should be mutually exclusive and unambiguous, so if 'mouse' appears in one place referring to the small furry mammal, it shouldn't turn up elsewhere referring to the pointing device attached to most computers.

When web designers need to describe the structure of a website, they often create a 'site map' taxonomy view of the site, however the pages on a site are related in many ways and can usually be reached from many other places, so the structure is more complex than a traditional taxonomy allowed.

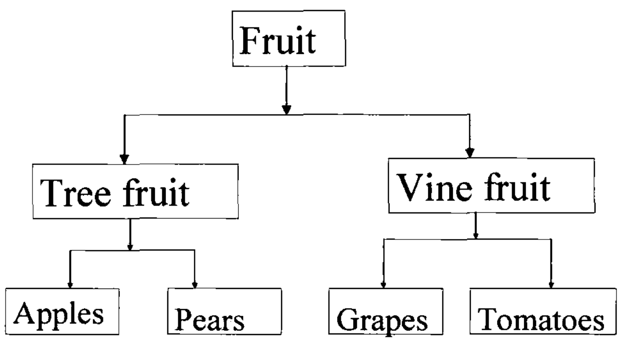


Figure 7 A traditional taxonomy - each item exists only once and in one place (From [50])

3.5.2 Polyhierarchical taxonomy

A polyhierarchical taxonomy allows items to appear in more than one place in the relationship diagram, so a user can reach a given taxon by a variety of routes. For example, botanists classify tomatoes as a fruit, however, they are commonly classified as a vegetable. After an import tax was imposed on 'vegetables' but not 'fruit', the U.S. supreme court ruled that "in the common language of the people", tomatoes should be classified as a vegetable [51].

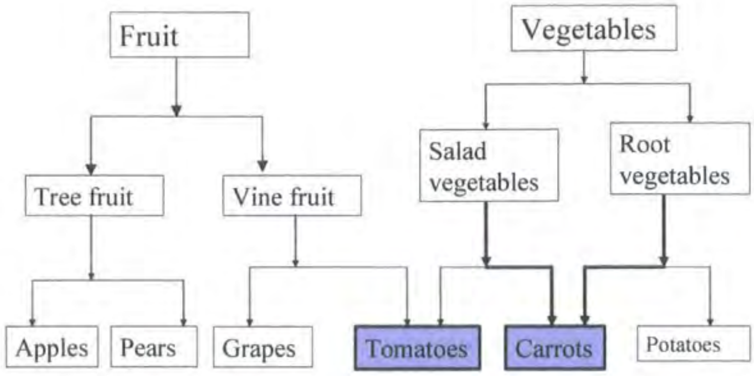


Figure 8 A Polyhierarchical taxonomy: Botanists classify tomatoes as a fruit; however the most people may search for 'tomatoes' in the vegetable section of their local greengrocer (From [50])

3.5.3 Thesaurus

A thesaurus is a taxonomy with extras [50]. It shows lateral connections (such as 'related' and 'see also' terms). It has an underlying index showing words that people may use but shouldn't be used for tagging ('non-preferred terms') and tells you what to use instead ('preferred terms'). While a taxonomy is designed to classify things, a thesaurus is designed to help you find the right words or phrases to describe what you're looking for.

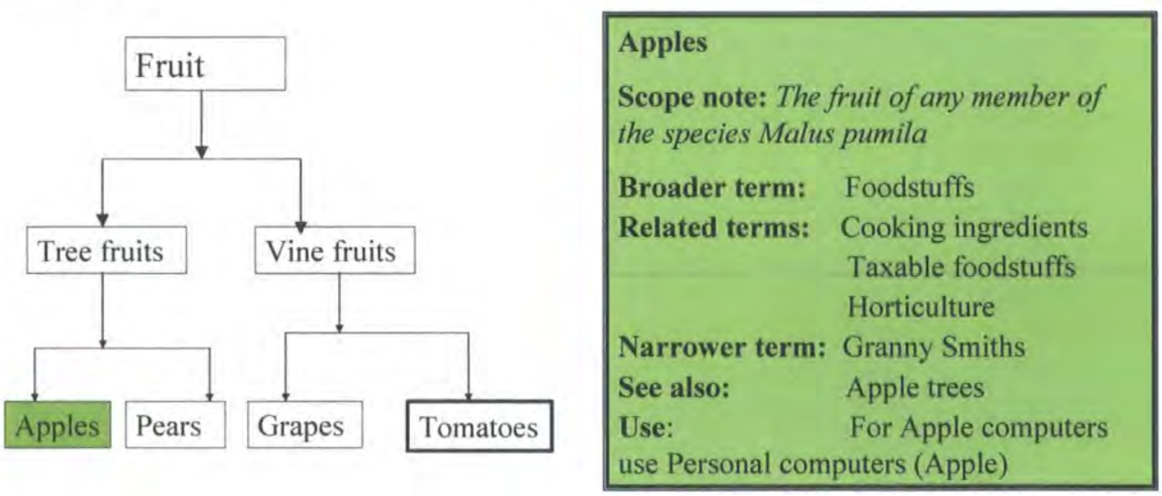


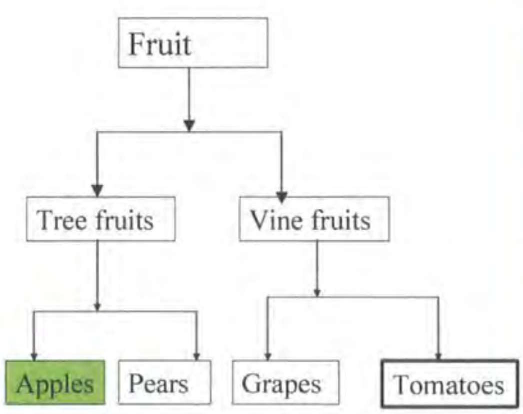
Figure 9 Thesaurus. Behind the scenes, notes show some of the additional features of a thesaurus, including a variety of related terms (From [50])

3.5.4 Ontology

'Ontology' comes from the study of philosophy and came to its current use by artificial intelligence workers who used ontologies to describe the set of rules governing an artificial world.

Ontologies are important for developing the semantic web. They allow information to be marked up in a way that machines can understand, thus an electronic web 'agent' can tell the difference between a granny smith apple and an Apple iPod.

The UK Government's e-Envoy describes an ontology thus [50]:
"An ontology is therefore a thesaurus gone mad. It is more specific in defining a concept or item and its relationships. A dog will be a noun, and an animal, and a mammal, and possibly a pet. Instead of having 'puppies' as a narrower term of 'dogs', it might have puppies as offspring of dogs. Ontologies often cover all elements of metadata, not just other subject terms. For example, a document could be connected to a person by the relationship 'created by', this person may then connect to an organisation by the relationship 'works for'; to an address by 'lives at' and other documents by 'author of'. This can lead directly from one document to others written by the same person, or by others working for the same organisation that the author works for. Ontologies can be machine generated from good metadata. The semantic web will be built on ontologies."



Apple *Noun* **Apples** *Noun plural*
Scope note: *The fruit of any member of the family Malus pimula*
Price *£0.90/kilo*
Supplier *Bob's fruit store* **Delivery area** *SW12*

Agent: Are apples expensive?

Apple is a noun. Expensive is an adjective. Adjectives are used to distinguish nouns. An Apple is either a computer or a fruit in any context. Apple in the context of computers is type of computer. If an Apple is perishable or has pips it is fruit. Apple in the context of fruit is a type of fruit. When Apples are computers, expensive means price >£1500/ea. When Apples are a fruit, expensive means price >£1/kilo.

Therefore; these apples are not expensive.

Figure 10 This shows some of the data an ontology might hold about 'apple' so that the agent can answer the simple question about the cost. The agent needs a lot of details to follow its logic pattern.

Extra details would be needed to find out about delivery, payment methods, or, if the apples were British-grown organic granny smiths (From [50])

In conclusion, there are a variety of schemes which vary in complexity from a simple list to an ontology to describe the complex relationships between data, creators, organisations, and so on.

3.5.5 Some Controlled Vocabularies

One problem with using a controlled vocabulary is that there is no single vocabulary that covers everything in the world in enough detail. The Government Category List (GCL) is a taxonomy consisting of 420 preferred terms covering the entire range of government activities [52]. The GCL was designed for use in meta-tagging electronic resources and building directory structures [53].

The GCL consists of 12 top-level terms ranging from 'agriculture and environment' to 'science and technology'. The GCL encompasses the entire range of government activities; therefore the majority are irrelevant to a health-specific domain. A number of controlled vocabularies designed specifically for the healthcare domain have been created.

Public Health Observatories had developed a taxonomy called PHITS ('Public Health Information Tagging Standard), the Health Development Agency used a thesaurus called 'HealthPromis', a European multilingual thesaurus on health promotion [54, 55] was developed, and in the United States, one hundred different controlled vocabularies were combined to create a Unified Medical Language System Metathesaurus [56], containing over one million terms. Unfortunately, the variety of different controlled vocabularies heads back to a semantic mapping problem – how could one system be sure that it is able to retrieve an accurate list of resources from another system if they do not share a common vocabulary? One solution is to create a mapping from one vocabulary to another. The solution that was adopted by the public health observatories and their partners in organisations such as the HDA was to merge their two vocabularies to create a new health-wide taxonomy called the National Public Health Language (NPHL).

3.6 Chapter Summary

In March 2000, the Prime Minister outlined the UK's aim to become "the world's leading Internet economy" [35]. In order to achieve this goal, the Electronic Government Interoperability Framework was mandated across the public sector. This chapter has described the aims and key policies involved in implementing the e-GIF.

By 2001, the number of government websites in the UK was unknown [27]. With a target of delivering 100% of government services online by the end of 2005, the number of government services available over the internet has continued to increase. Calls for simplification will continue until interoperable methods can be developed to ensure services can be accurately located.

The role of accurate metadata to describe resources is crucial for information retrieval. This chapter has described the Electronic Government Metadata Standard, mandated as part of the e-GIF. The chapter went on to describe types of controlled vocabulary, including taxonomies and ontologies. A number of controlled vocabularies were described, including the Government Category List, which is a required element of the e-GMS, along with controlled vocabularies dedicated to the health-care domain.

Chapter 4 Background and Analysis of NEPHO website

4.1 Chapter Abstract

The aim of this chapter is to introduce the case study organisation. It will give an introduction to Public Health Observatories, including a project aimed at developing interoperability between each observatory's website. The case study focuses on the North East Public Health Observatory. A review of their website at the start of the interoperability project is included and a list of actions is drawn up.

4.2 Public Health Observatories

The Government's 1999 health whitepaper 'Saving Lives: Our Healthier Nation' [1] describes the need for Public Health Observatories, saying:

> *'We need a clearer national picture of health and health inequality so that we can track changes over time. Many agencies are involved in collecting and using information about health and disease in the population. Yet in some cases information may not be available, or may be unreliable ... At local level data may be even patchier.'*

The solution to this problem was to set up eight Public Health Observatories in England, one in each NHS region of the country, to strengthen the availability and use of information about health at local level [58].

Figure 11 PHO Organisational Map, June 2000 [58].

The PHOs were launched in February 2000 by the Parliamentary Under Secretary of State for Public Health, and the Chief Medical Officer. On the same day, a national website (www.pho.org.uk) went live.

PHOs main tasks are to support local bodies by:

- Monitoring health and disease trends and highlighting areas for action;
- Identifying gaps in health information;
- Advising on methods for health and health inequality impact assessments;
- Drawing together information from different sources in new ways to improve health;
- Carrying out projects to highlight particular health issues;
- Evaluating progress by local agencies in improving health and cutting inequality;
- Looking ahead to give early warning of future public health problems.

PHOs are also required to strengthen public health input into the broad range of cross-government initiatives aimed at improving health and reducing inequalities [58].
Saving Lives: Our Healthier Nation [1] states that there should be a high level of collaboration between Public Health Observatories such that they will be linked

together to form a national network of knowledge, information and surveillance in public health. In addition, each regional Public Health Observatory is expected to provide a national link on key policy areas.

The Association of Public Health Observatories (APHO) was established in June 2000 to facilitate this collaborative working at national level. It comprises representatives from all regional Public Health Observatories and a number of partners such as the Department of Health, the Office for National Statistics, the Health Development Agency, and the National Centre for Health Outcomes Development [58].

The APHO provides an important link between regional Public Health Observatories and the national governing arrangements and is a valuable forum for both disseminating good practice and for co-ordinating action across Public Health Observatories.

4.2.1 PHO realignment

Public Health Observatories were originally established on the basis of NHS regions. In 2002, NHS regions were replaced by a regional public health function for each government office region [57]. In April 2003 a ninth observatory was created in order to align observatories with the nine government office regions. The old Northern and Yorkshire PHO was rebranded as the North East PHO. Trent PHO was rebranded as East Midlands PHO, and a new PHO was created to cover the Yorkshire and Humber region. Observatory-like functions are subsequently being developed in Wales, Scotland and Ireland.

4.3 Outputs of a PHO

The first APHO annual report *Progress and Prospects 2000/01* [58] notes that 'All Public Health Observatories are keen to make use of web-based technologies to facilitate sharing and dissemination of information'. During 2000, websites for each of the eight PHOs were launched. The annual report highlights 23 examples of work completed in 2000/01 and hosted on one of eight (now nine) regional websites.

Typical outputs of each PHO include:

- National reports on a lead area
- Local reports on wide range of health-related subjects
- Newsletters
- Events – including conference and training
- Data

PHOs use a range of distribution channels for information, including print media, email groups and conferences. Virtually all the outputs of a PHO are archived on their website, often including additional resources such as raw data in the form of spreadsheets or interactive analysis tools.

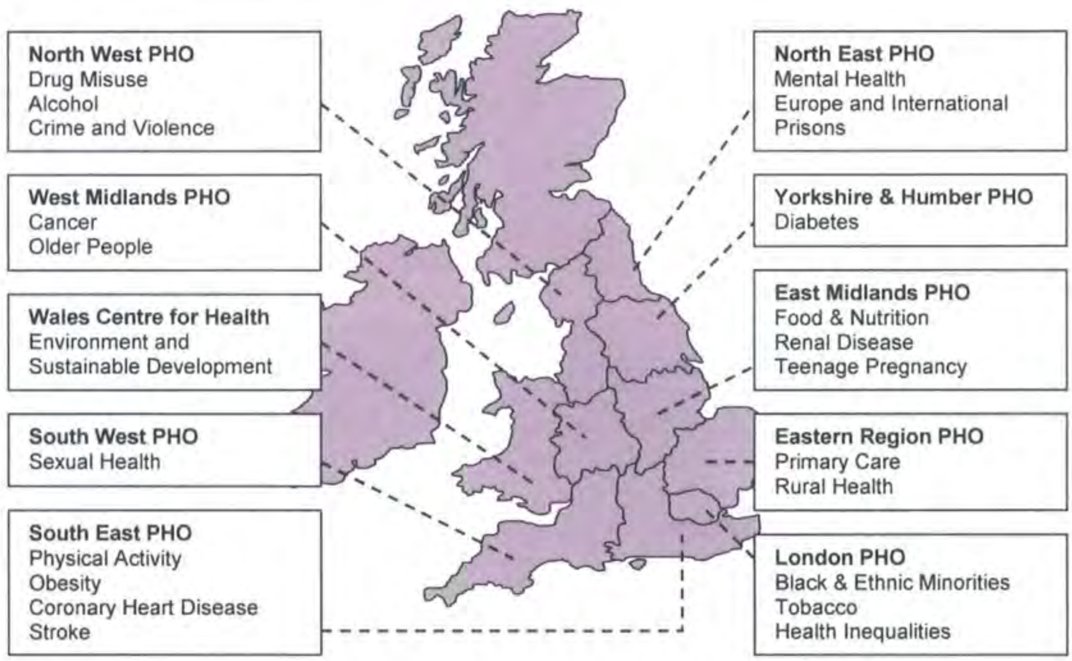


Figure 12 PHO regions and national lead areas post April 2003

By 2002, the number of resources hosted on each PHO's website had continued to increase. As each PHO had a number of specific national lead areas (Figure 12), many published reports covered the whole of England, but were only published on the particular regional website. For example, the London Health Observatory had the lead area on smoking and tobacco, therefore they published reports about tobacco consumption for the entire UK and hosted the reports on the LHO website. The exposure granted on the other observatory sites was poor, a user searching the NEPHO website for tobacco consumption would not find the report published by LHO, despite it covering tobacco consumption in the North East of England.

4.4 APHO interoperability project

In 2002, discussions between Public Health Observatories led to the formation of an interoperability group who would look at ways to share metadata about resources across each organisation's website. The group consisted of a representative from each observatory – including web developers, data specialists, a librarian, and a clinician.

As outlined in section 1.2, the Association of Public Health Observatory directors agreed to develop interoperability of websites over a two-to-three year time frame at their meeting on 7 January 2003. A vision was proposed where "each organisation has its own one-stop-shop but, that shop enables the searching of other organisations' data and resources, using whatever search mechanism is used" [2]. Other organisations including PCTs, the Department of Health, the Office for National Statistics, the Health Development Agency, the Health Protection Agency and the nine regional bodies should eventually all interoperate.

4.4.1 APHO interoperability levels

The APHO interoperability group defined three levels of interoperability between organisations. These were termed 'basic', 'intermediate' and 'advanced' level. Each level was defined from a technical and a user's viewpoint. Each PHO would aim to become interoperable with other PHOs at the 'advanced' level; however it was expected that interoperability would have to be implemented in phases, so an organisation may initially implement a 'basic' solution while developing more advanced interoperability. The choice of requirements for each level was heavily influenced by the decision to support the e-GIF. Public Health Observatories believed they were outside the scope of the government's mandate to require e-GIF compliance, however even if e-GIF was not mandated for PHOs, many PHO partner organisations would be required to fully implement e-GIF and PHOs believed that basing the requirements on a common standard would be beneficial. Table 3 highlights interoperability levels from a user's viewpoint, as presented in a report to PHO directors in November 2003 [59].

Basic	Simple text search Indeterminate results -> on a par with a typical search engine. For example, a search for "heart disease" would return a report with the description "This report is about circulatory diseases except heart disease", but it would fail to return a report about "ischemic circulatory disease"
Intermediate	An ability to refine the search – for example, only view reports published after a certain date, or related to a certain term in a controlled vocabulary
Advanced	Users could "search other websites as effectively as their own", including access to restricted information For example, a search for "smoking" would return documents about "tobacco consumption"

Table 3 User View

Table 4, below deails the technical requirements defined by the APHO interoperability group. These requirements were based upon requirements set out in the e-GIF and associated e-GMS – including the selection of either the PHITS or NPHL controlled vocabularies, described in section 3.5.5, for resource classification to be used in the 'subject' e-GMS element. In addition, a controlled list of resource types, called PHRTES, was agreed to describe the format of resources.

	Basic	Intermediate	Advanced
Is fully e-GIF compliant	✗	✗	✓
Ability to use PHITS or NPHL for tagging and retrieval	✗	✓	✓
Resources keyword tagged with PHITS or NPHL categories	✗	✗	✓
Use common access control	✗	✗	✓
Have a dynamic website (i.e. using a database)	✓	✓	✓
Use APHO standard for type encoding (PHRTES)	✗	✗	✓
Use APHO standard for minimum metadata set	✗	✓	✓
Use APHO standard for common XML Schema	✗	✗	✓

Table 4 Technological and functional requirements for the 3 levels of interoperability

While this requirements analysis was satisfactory for the APHO board of directors, it should be noted that a more detailed technical requirements analysis may have been beneficial for the project at this stage. The list of requirements did not specify methods for physically passing messages between differing systems. As Chapter 2 has outlined, there are many potential ways to share data between systems however the interoperability group did not specify how one 'advanced' level system would actually communicate with another 'advanced' level system – this was subject to further investigation at a later stage once observatories had developed e-GMS compliant databases. Subsequently, NEPHO's action list described below includes 'interoperable methods to access metadata', however there was no guarantee that other PHOs would develop compatible methods and as Chapter 6 describes, significant time and resources were wasted due to vague requirements written in plain English. This problem is further assessed in Chapter 6.3.

4.4.2 Summary of APHO requirements

These views mean that:

At **basic** level, there is some ability to search and be searched

At **intermediate** level, interoperability is to a high level, but is not fully eGIF and/or not all resources are tagged with a controlled vocabulary, although the vocabulary is useable

At **advanced** level, websites are fully eGIF complaint and adheres to all APHO standards.

All PHOs will aim to develop advanced-level interoperability with each other.

4.5 Project Requirements

This section lists the requirements for work to be undertaken in order to redevelop the NEPHO website as part of this thesis.

NEPHO's website needs to comply with the 'advanced level' interoperability criteria defined by the APHO interoperability group. This means that the NEPHO website will have to be:

- e-GIF compliant
- All resources will have to be tagged with the PHITS tagging system (or an equivalent)

- APHO standards will have to be used to classify resource types

At the highest level, e-GIF compliance means that:

- Access will have to be provided through a browser interface (not FTP package)
- XML should be used as the primary means of data integration
- Internet and World Wide Web standards should be used.
- All resources should comply with the e-GMS.

4.6 Assessment of existing NEPHO website

The following section reviews the status of the NEPHO website at the start of the thesis. As part of this thesis, an action list has been created in order to highlight the tasks which need to be completed during the implementation phase in order to develop a website that complies with the 'advanced' criteria defined by the APHO interoperability group [59].

The North East Public Health Observatory website (Figure 13) consists of four resource types. At the start of the project, the site hosted the following resources:

- 27 events
- 75 reports classified in one of 11 report types
- 28 news items
- 46 web pages

The site was launched in April 2003, when PHOs were realigned. The site includes archived reports published by NYPHO and consists of a mixture of static and dynamic content.

Figure 13 NEPHO Homepage

The site consists of a page template, and 46 PHP/HTML based pages which contain content for each page. Pages are organised into a three-tier menu hierarchy. The pages are edited in a web design program called Macromedia Dreamweaver, and uploaded via FTP. Page structure is stored in a database table called 'menu_data', which contains the following fields: Page ID, Filename, Page Title, Menu Sequence and a second page ID, used to classify pages as being in a sub-group of the associated page. An online management function is provided (Figure 14). That allows administrators to add pages, edit metadata about the page title, filename, and menu sequence.

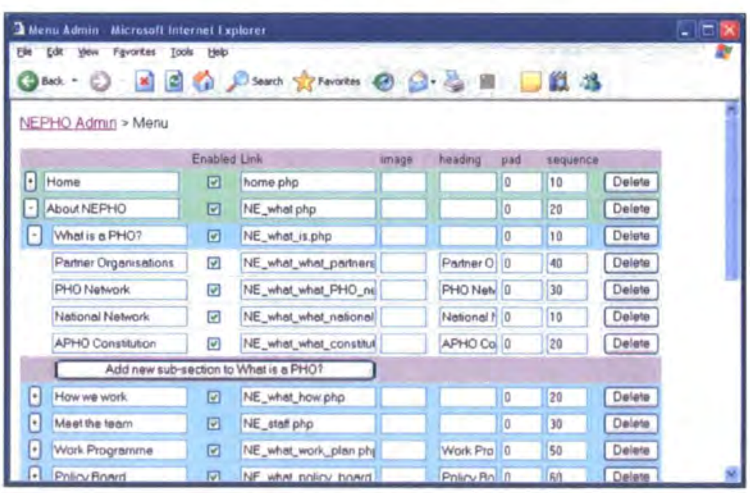


Figure 14 Menu administration

Some page files include PHP code which is executed when the file is loaded by the template file. These include special sections for news items, events, reports, and a contact form.

The majority of resources stored on the NEPHO website are reports. Reports hosted on the existing NEPHO website were organised into eleven categories, with a total of 126 associated files. Categories include 'occasional papers', 'project reports', 'work in progress' and 'archived reports'. Each file is uploaded by FTP and stored in a subfolder of the main website. A web-based report administration function is provided (Figure 15).

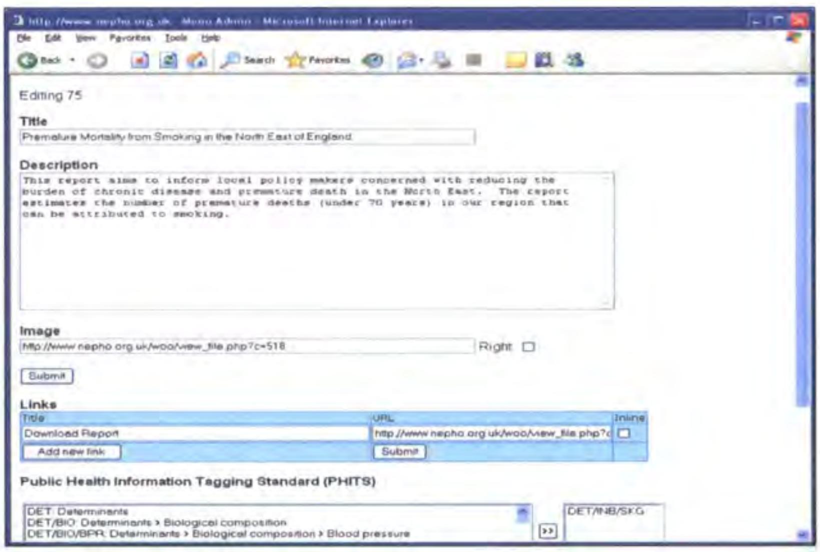


Figure 15 Editing a report

This administration function is linked to a database table called 'reports' which includes the following fields: Report ID, Category, Title, Description and a link to an optional thumbnail image file. Two further tables are provided – 'report_links' links reports to uploaded files and includes the fields ID, Report ID, filename, and title. NEPHO has also implemented a trial of the PHITS keyword tagging ststem. Tags are selected from a list box. The selected tags are stored in the second table 'report_PHITS', linking reports IDs with PHITS IDs.

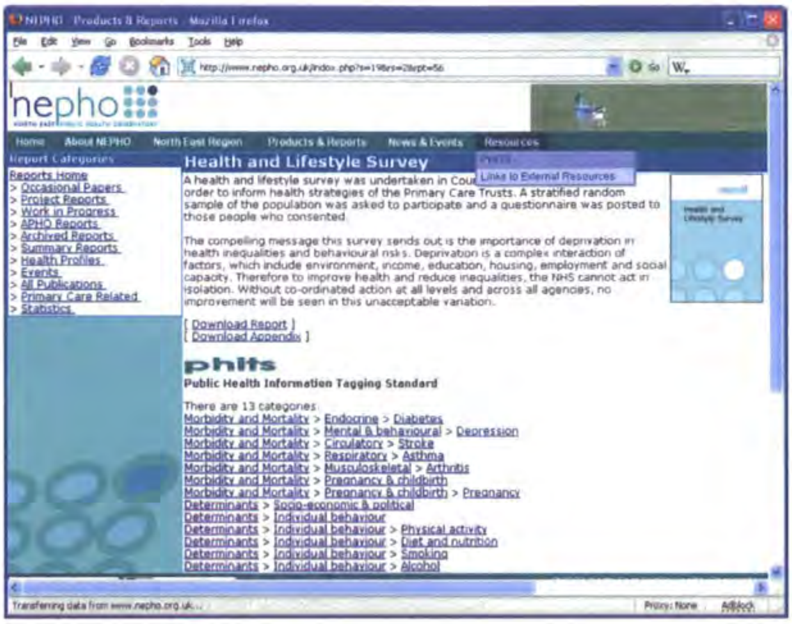


Figure 16 A typical report on the NEPHO website. Metadata stored about the report includes title, description, thumbnail URL, file descriptions / URLS, and related keywords.

The reports section appears on the page menu as a single page. The PHP file loaded by the template includes further instructions for displaying reports (Figure 16). By default, this lists report categories. Selecting a specific category displays all reports in the category, selecting a specific report displays information about the report – the report title, description, thumbnail image, links to files, and related PHITS categories.

Further administration functions have been programmed for news (data stored in a 'news' table) and events (data stored in an 'events' table). Scripts handling these resources are handled in the same way by the menu administration function and page template. The further functionality is programmed into the linked PHP script.

4.7 Action List

As a result of the investigation into the requirements of the organisation, an action list was drawn up. The action list consists of the tasks that need to be completed to achieve the objectives. The tasks have been listed in order of development. For example, database structure and administration functions for entering metadata must be implemented before resources can be tagged. Interoperability with other organisations can only proceed once the database has been populated.

4.7.1 Action One: Develop online management for metadata elements

There are a lot of metadata elements defined in Dublin Core, e-GMS, and APHOMS. Most of these elements share common attributes. To manually program information entry, storage, and retrieval mechanism for each element would be repetitive, time consuming, and lead to difficulties in maintenance. Each element has a set of attributes such as Dublin core syntax, a requirement level (mandatory, recommended or optional), and a data type (integer, date, string, or linked to a controlled vocabulary). For each element, an administration function will need to be developed, and the element would need to be described in web services. Generic solutions should be developed based upon element attributes, rather than the individual elements themselves. This approach will improve reliability, introduce code reuse and subsequently reduce development effort.

4.7.2 Action Two: Develop online resource management

Accurate metadata is vital for describing resources. Every resource stored on the NEPHO website needs to comply with the e-Government Metadata Standard and APHOMS. Resources should be managed through an online administration function based on the APHOMS elements defined in the metadata management function (4.7.1). As e-GIF requires the browser to be used as the primary interface, the new administration function should include the ability to upload files such as reports or thumbnail images without the need to use FTP.

4.7.3 Action Three: Create interoperable methods to access metadata

e-GIF requires XML-based methods to be used as the primary means of data integration. An XML-based web service should be created to facilitate metadata search and retrieval. Additional methods of facilitating interoperability such as XML-based RSS feeds should also be investigated.

4.7.4 Action Four: Develop a front-end for viewing metadata-based resources

Resource metadata should be used to create web-pages for each resource. Typically, this should include the resource title and description, plus a list of relevant categories from a controlled vocabulary. Users should be able to browse other resources which are explicitly or implicitly linked in the database. Explicit relationships are where an administrator has created a link between two resources – for example, a PDF report file may be linked to a thumbnail image of the report. Implicit relationships are automatically inferred from the resources metadata – for example, a link to all other resources with the same keyword tag, information about the publishing organisation, or created in the same month.

4.7.5 Action Five: Work with other observatories to develop interoperability

Other organisations should be able to use the web services described above to search and retrieve metadata from the NEPHO website. Tools also need to be created to search the solutions created by other observatories and wider organisations who work with NEPHO, including national health-related organisations and regional organisations such as the North East Regional Intelligence Network.

4.8 Chapter Summary

This chapter described the organisation being used in the case study – NEPHO. It also described the Association of Public Health Observatories, the aim of the interoperability project and what the organisation is trying to achieve from it. Finally it described their current systems, leading to a requirements specification and an action list of activities to be included in the solution.

Chapter 5 Design and Implementation of the action list

5.1 Chapter abstract

This chapter describes work undertaken as part of this thesis in order to fulfil the requirements set out in section 4.5. The development will proceed in five stages as defined in the action list created in section 4.7:

Action One	Development of a tool to manage metadata element definitions
Action Two	Development of a resource management system using the elements defined in the previous step
Action Three	Development of methods to extract metadata using XML-based methods
Action Four	Development of a front-end using these methods
Action Five	Working with other organisations to develop interoperability between different solutions.

5.2 Action One: Develop online management for metadata elements

Information about element attributes will be stored in a database table called 'aphoms_elements' (Table 5). An online management function will allow administrators to edit data for each element which is stored in this table. The data stored in the aphoms_elements table will be used to define data types for metadata stored about each resource – it will be used to define the administration function, display in the front-end, and build complex types for the web service.

Field	Details
element_id	A unique ID number for each element / refinement (e.g. '9')
label	Text displayed in administration function (e.g. 'Publisher')
mapping	The metadata scheme element is taken from - typically Dublin Core (e.g. 'DC')
element	The element within the mapped metadata scheme (e.g. 'publisher')
refinement	Any further element refinement (e.g. 'dateCreated')
syntax	Syntax used for identifying elements, used in web service and <meta> tags in the header of web pages. (e.g. 'DC.publisher')
min_occ	The minimum number of times an element must appear for each resource. This attribute can ensure that mandatory elements are completed. (e.g. '1' [for a mandatory element])
max_occ	The maximum number of occurrences an element can appear
type	The data type used to store each resource, such as 'string', 'int', 'date'
linked_table	A number of elements will be linked to a controlled list of potential values. These attributes describe the table storing this information. (e.g. 'Countries', 'Country_id', 'Country_name')
linked_field	
displayed_field	
stage_id	Allows elements to be displayed in stages

Table 5 Information stored in the aphoms_elements table

The aphoms_elements table was populated with 'label', 'mapping', 'element', 'refinement' and 'syntax' information from a spreadsheet detailing APHOMS elements as proposed by ERPHO and discussed at a meeting of the APHO interoperability group. Data type formats for each element had to be added. An online administration function was developed to administer each element (Figure 17).

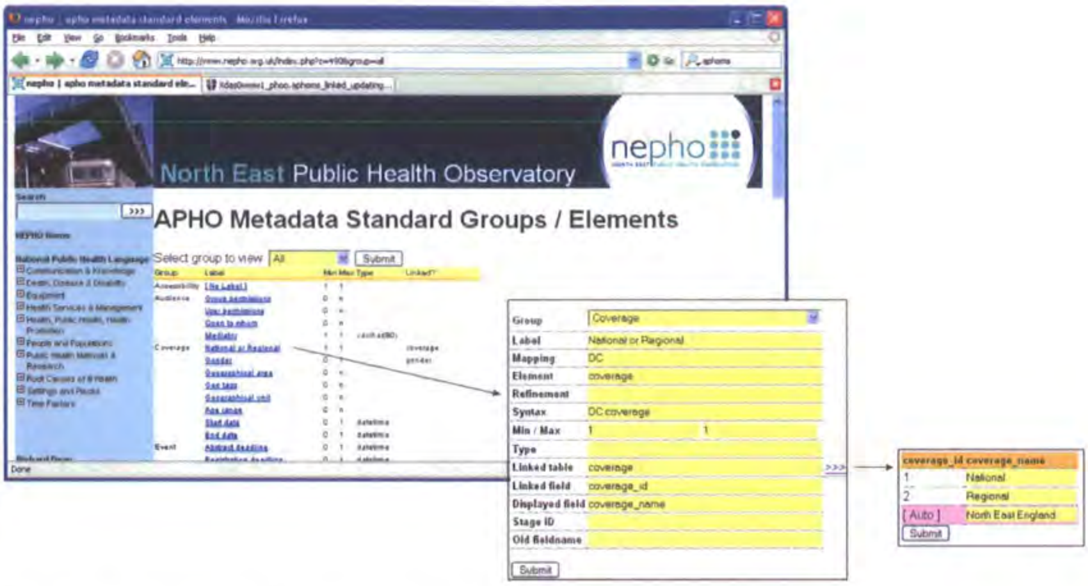


Figure 17 - Screenshot of APHOMS administration function

Once definitions for all applicable elements had been completed, element definitions were used to update the database structure. There are a variety of different element types. Information was stored in a number of tables, as described below;

- If the element is linked to controlled list of values stored in another table, then store this metadata in a table called 'aphoms_content_linked'
- Else, if the element has an unlimited maximum cardinality, store in a table called 'aphoms_content_multi'.
- Else, store the metadata in a table called 'aphoms_content' – which will have a field named after the element syntax, and a type as defined in the aphoms_elements table.

The online management form successfully created the database structure.

The APHOMS management page was programmed to update the database when a new element was added. On a few occasions, an existing element's data type was changed to reflect the client's wishes. For example, the 'address' refinement was originally defined as a varchar (i.e. one line of text), but was subsequently changed to a text type (i.e. multi-line text). Changes to an existing element were stored in the aphoms_elements table, however the data type stored in the aphoms_content table had to be manually changed as it was felt that changing the data type for existing elements could potentially loose data already stored in the database – for example, by reducing the length of a text field. As this type of change is very rate, the additional cost of developing such functionality outweighed the benefit of implementing such a solution.

5.3 Action Two: Develop online resource management

Resources stored on the existing NEPHO website included news, events, reports and html pages (see section 4.6). These needed to be imported into the new data structure defined by the APHOMS administration function.

The requirements for developing an online resource management function are:

- Develop an online page-management function based on element types defined in section 5.2.
- Investigate methods to edit HTML code in a browser.
- Store metadata in APHOMS structure defined in section 5.2.
- Develop a method to upload files without using FTP.
- Import existing resources into the new data structure.

In the United States, Crystal [60] investigated interface design for metadata creation, which has become 'even more vital' with the exponential growth of the World Wide Web.

> 'While the value of metadata is well understood, appropriate methods for creating and using web resource metadata are still in development. Initial research suggests that resource authors can create metadata suitable for use in surrogates than can facilitate information seeking and retrieval. Usability problems limit authors' facility with metadata creating and suggest a need for better interfaces' [60].

In Crystal's experiment, resource authors were asked to create Dublin Core metadata using a web-based application. Screen capture software was used to record the user interface activities. The experiment concludes that for decentralized metadata creation to become a reality, better designs are needed to reduce users' cognitive load and lower barriers to efficient metadata entry [60].

5.3.1 Resource metadata administration

Section 5.2 described the database structure for storing metadata about resources. The metadata elements described by this function were used to develop a resource management tool. Each element was defined according to its data type and its cardinality. Resource metadata is stored in the tables aphoms_content, aphoms_content_linked and aphoms_multi.

Figure 18 Linked metadata

Four different edit functions were developed – one for each of the following data types:

- Linked tables – with / without a maximum number of occurrences (Figure 18)
- Dates (Figure 19)
- Short text fields (one line, e.g. 'title')
- Long text fields (multiple line, e.g. 'description', described in more detail below)

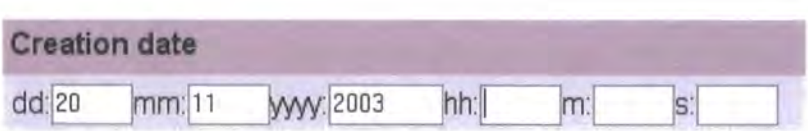


Figure 19 Metadata element with 'date' format

Administrators could edit resources by clicking an 'edit' or 'add' button. This led to an initial metadata entry page detailing the resource title and description metadata. Further metadata elements were provided via a number of tab buttons (Figure 20). The tabs linked to a page called 'All APHOMS', which listed alphabetically all the APHOMS elements and refinements which had been defined in the APHOMS administration function. As this page was quite long, links were also provided to subsets of the elements, grouped alphabetically (Figure 20). Further tabs were provided for special administration functions, such as uploading files or keyword tagging with the NPHL.

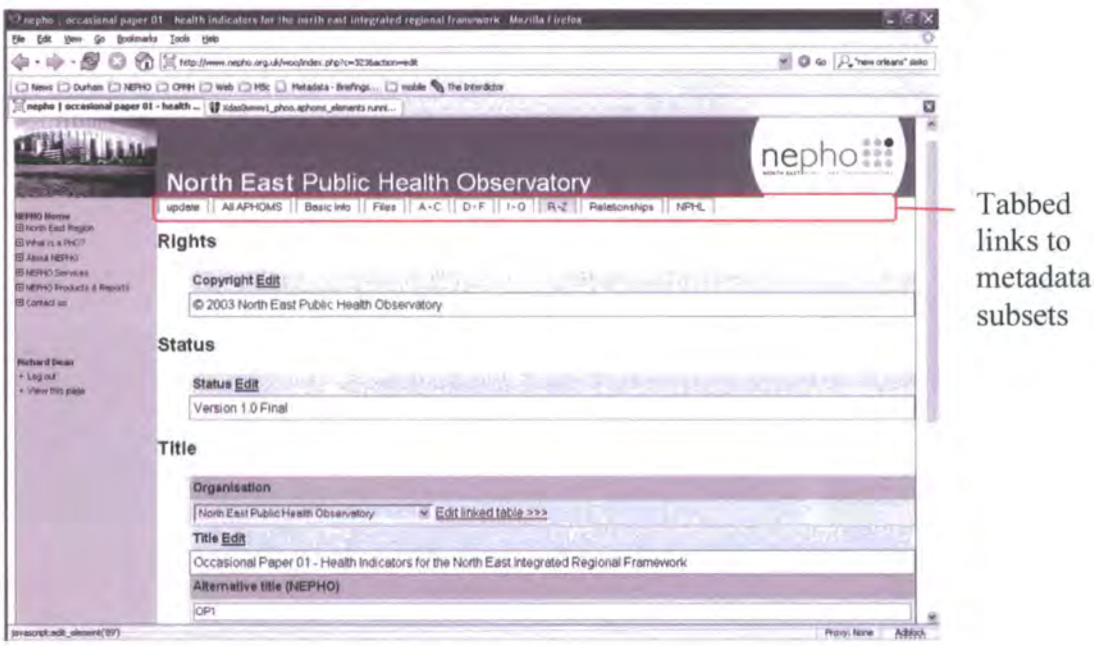


Figure 20 Entering metadata about a resource on the NEPHO test site.

5.3.2 WYSIWYG Editor for long text fields

NEPHO wished to use multi-line long text fields to store marked up information, allowing basic formatting for pages, such as bold and italic text, and bulleted lists. In the database, long text fields were stored as the 'text' data type, allowing up to 65,535 characters, including HTML markup. Three options were investigated:

- Using a HTML 'textarea' input type to type in the HTML code
- Using a simplified markup such as BBCode [61]

- Using a third-party JavaScript application for displaying and editing in a WYSIWYG (What You See Is What You Get) environment.

The first option – to allow HTML source to be edited directly in a text-box was discounted because the client did not know how to program in HTML, and it would be very simple to miss some closing tags, or break the page layout. These concerns are addressed in the second option investigated – to use a guided plain text edit function, such as BBCode.

BBCode (Figure 21) is a special implementation of HTML designed to improve the user-friendliness of entering mark up code in a textfield [61]. BBCode was developed for the internet bulletin-board community in order to provide a safer, easier and more limited way of allowing internet message board users to format their messages. Previous methods of posting messages relied on users entering HTML code which could easily break the bulletin board page layout, or run unwanted JavaScript.

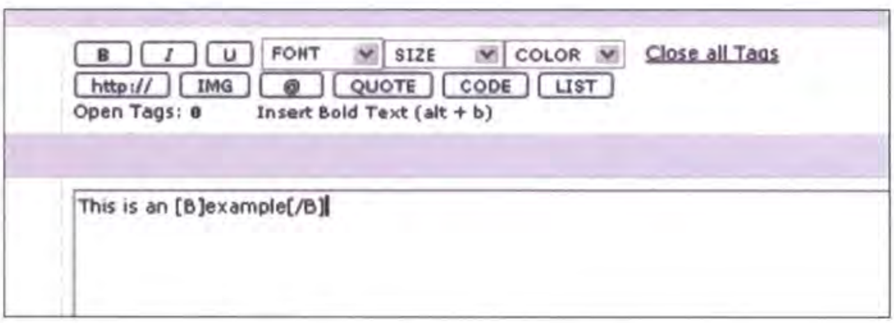


Figure 21 Example BBCode data entry

Recent versions of Internet Explorer and Mozilla have included sophisticated HTML editing functionality directly inside web pages. Mozilla's Midas specification [62] enables rich-text editing in browsers based on Mozilla, including Netscape, which is the default browser used by NEPHO. The Microsoft HTML editor [63] is a built in "function rich text editor" for Internet Explorer. Both solutions are browser-specific – A form using the Mozilla editor function would not work in internet explorer and vice-versa. Roth [64] developed a rich-text editor that would detect the browser and invoke browser-specific editing functions through a common interface. Roth's code was used to test a WYSIWYG editor (Figure 22) which proved acceptable as it worked in all browsers used by the client, provided an easy-to-use interface that the client could understand, and the code was made freely available with no restrictive license terms.

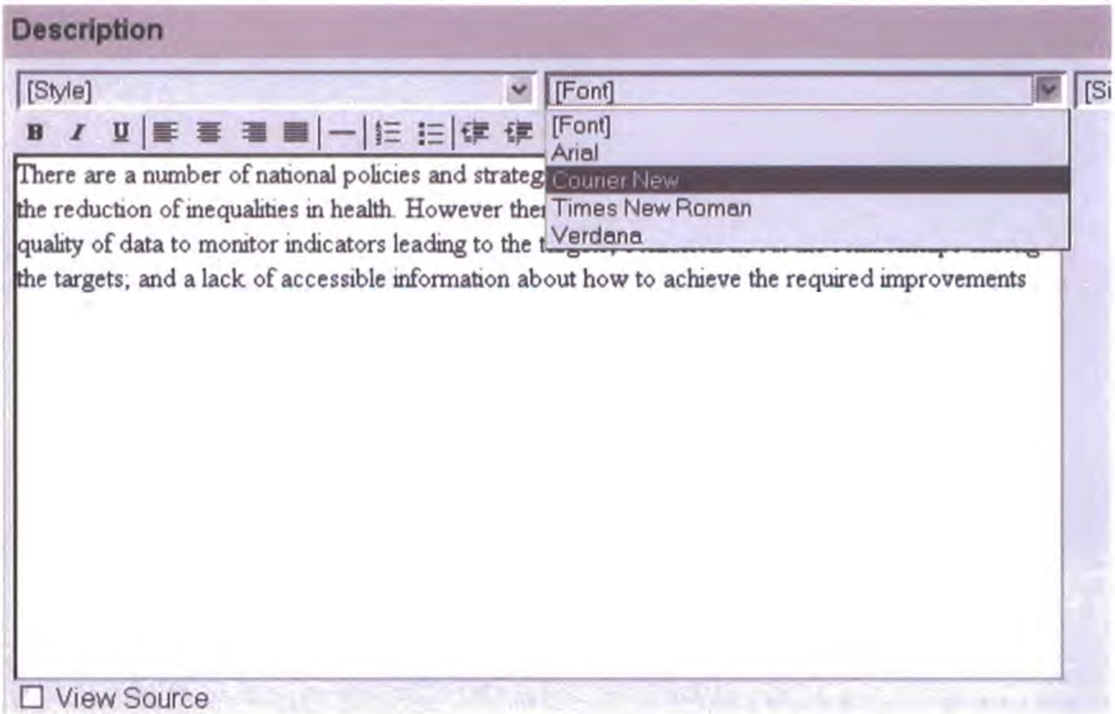


Figure 22 WYSIWYG editor on NEPHO's administration page

5.3.3 File Upload Function

The requirements analysis indicated that the existing method of uploading files via FTP was too difficult for most administrators. e-GIF specifies that the browser should be used as the key interface. The action list therefore specifies that a file upload function should be developed. Uploading files through a web-browser is quite straight-forward using a multipart MIME-encoded form. In PHP, the uploaded file is stored in a temporary location and is destroyed when the script finishes execution. Two solutions for storing the uploaded file were considered:

- Writing the file to a website sub-folder
- Storing the file as a binary data type in the database.

A prototype for each storage model was implemented, based on a web-form which allowed a user to upload a test file. This was stored in a temporary location by the upload script, which then moved the file to a sub-folder of the website, or stored the file in the database. Each prototype is evaluated in the subsections below.

5.3.3.1 Disk storage

Resource files on the existing NEPHO site were stored in a sub-directory called 'resources/files'. This file system structure was replicated in the prototype.

During trials of this approach, a file permission problem emerged with uploaded file once moved to a subfolder. Scripts executed on the PHP server run from a 'generic web user' account. Files uploaded through a web-browser and subsequently placed into a user's sub-folder are therefore 'owned' by the 'generic web user' account, rather than the user account specific to the NEPHO domain. As a consequence, the file system refused to allow deletion, movement, or replacement of the file via FTP or via a shared network drive connection. Furthermore, as files uploaded did not belong to the specific NEPHO account, the file space allocation was taken from elsewhere on the service provider's disk allocation. Further security concerns were raised because the uploaded files belonged to the same user account as scripts created by other users, raising a possibility that files could be accessed, modified or deleted by another script.

5.3.3.2 Database storage

Databases provide a concept of a Binary Large OBject (BLOB) which can store binary data. The concept of a BLOB has been a reality since the 1980s. IBM's DB2 introduced VARGRAPHIC fields in 1984. Oracle introduced LONG and LONG RAW data types (c 1987) and Dbase introduced "memo" fields circa 1988. BLOBS in modern databases such as MySQL provide an alternative to file system storage.

The SQL standard defines "binary data" as "a sequence of octets that does not have either a character set or collation associated with it." [65]

One major disadvantage often cited by critics of storing BLOBs in a database is the additional overhead involved in retrieving information from the database. Atkinson [66] concluded that storing files in a relational database reduces performance by approximately a third, however Waterson [67] argues that storing binary data in a large database may infact be faster as there are no filesystem overheads to contend with. His benchmarking results indicated that there is little difference between requesting a file 100,000 times directly and requesting the same file from a database binary object. Such results fail to consider database and file system caching which could significantly alter the results – a more effective demonstration would be to request 100,000 different files.

The database-file-upload prototype used a single table called 'files' to store metadata about the file including:

- Filename (varchar)
- MIME type (varchar)
- File size (integer)
- File data (blob)

To reassemble the file, a script called 'view_file.php' was developed. This output appropriate header information, including file type and name, before outputting the file data. The prototype worked well for small file sizes, however large files were less successful due to the size of the data being stored in a single data field. The PHP server software could run out of memory if it read the entire file contents at once. The database was modified to store file data in a separate table called 'file_data'. Large files were read in chunks and split data over a number of rows. This approach ensured the PHP script did not run out of memory while working with a large file as only a section of the file had to be read and stored in the database at a time. This solution solved the problem, allowing large files to be uploaded through the browser and stored securely in the database.

5.3.3.3 File storage security

Some PHOs have implemented a user registration function and restricted access to some files. At the commencement of the project, NEPHO do not have any restricted files, and did not plan to introduce restricted areas of the site. File security was considered, however, because the two approaches differ greatly on this issue.

The database BLOB storage approach requires execution of a specific PHP script in order to build the file. The PHP script can include authentication functions to restrict access. An authentication system has already been developed to allow authenticated users to access administration functions on the NEPHO site. The file-system storage approach was less secure because the file is stored at a specific URL and no authentication script is executed to deliver the file to the user. There are a number of possible options to restrict access to the file, including the use of a password-protected sub-folder, or storing the file in a non-web-accessible folder, and then using a script to forward the file contents if the remote user could be suitably authenticated.

5.3.3.4 Evaluation of prototypes

The file-based prototype caused problems with user accounts, and raised security concerns. The performance implications of the database approach were not as significant as feared, while they were offset by the increased security and the elegance of storing the data alongside the metadata in a single database. Both approaches were discussed with the service provider who agreed that they would prefer the data to be stored in a MySQL database.

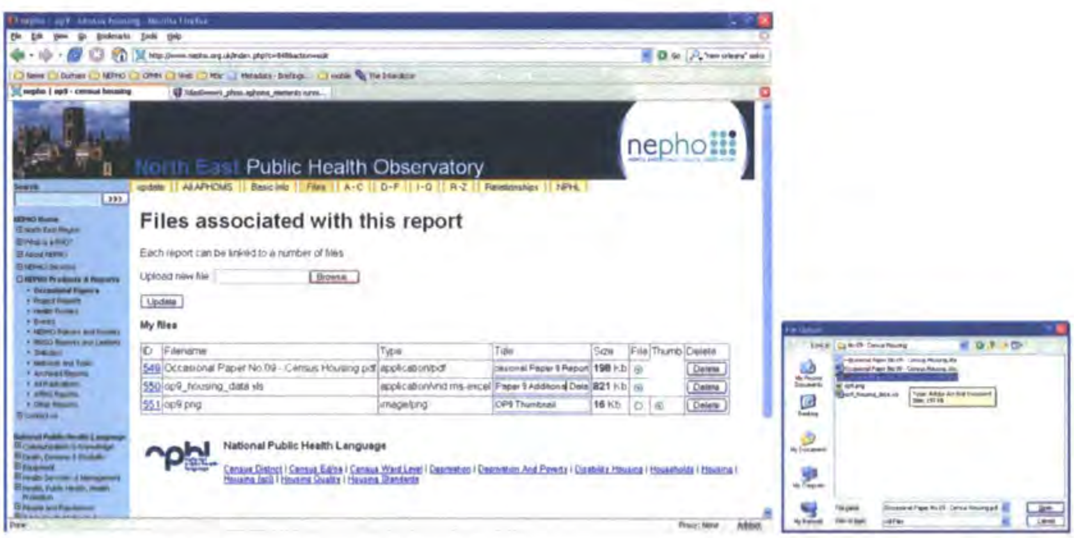


Figure 23 File Management

5.3.4 Keyword tagging with the NPHL

The National Public Health Language (NPHL) was chosen to classify resources. The NPHL provides more depth and breadth than the previous PHITS tagging system. A number of prototype tagging interfaces were developed. The first prototype required an administrator to navigate the NPHL one term at a time, following relationship links between terms. With a total of 1,400 terms, this method was very slow to accurately tag resources. Administrators felt unfamiliar with resources, and often found there was a lot of back-tracking involved in moving about the NPHL structure. Further implementations included a tree-based view, and showing each narrower term as a sub-category. This approach was welcomed, however displaying all the terms at once created a very long page, requiring a lot of scrolling. The page filesize was larger than usual, at 31kb – however as the site administrators would use a high-speed connection, this was not considered to be a major issue, which might limit the ability to display all 1,400 tags at once. A compromise was reached by using CSS and JavaScript to hide narrower terms by default, but allow them to be shown by clicking a JavaScript link that switched the visibility property of the narrower terms (Figure 24). This allowed the entire thesaurus to be loaded onto one page and narrower terms to be shown / hidden without the need to constantly refresh the page.

Figure 24 NPHL Tagging

A keyword search method was developed to allow administrators to enter potential keywords from the document and then view matching NPHL terms. This function proved to be very popular. The default one-line text box was expanded to a multi-line box in to allow multiple terms to be entered at once, speeding up the tagging process.

Occasionally, incorrect terms may be selected. HTML forms using checkboxes only submit selected values. Two types of checkboxes were therefore implemented – a checkbox to 'add' a term to a resource, and a checkbox to 'delete' a term from a resource. 'Add' checkboxes were highlighted in green, while 'delete' checkboxes were highlighted in red. Terms that had already been selected for a specific resource were provided with a red 'delete' checkbox while non-selected terms were provided with the default green 'add' checkbox.

5.4 Action Three: Create interoperable methods to access metadata

e-GIF requires XML to be adopted as the 'core standard' for data integration and management' [41]. XML-based web services were needed for search-and-retrieve functions. In addition, XML-based RSS-feeds were developed to allow syndication of resources, such as a 'latest news' feed. This mirrored similar incentives being developed by other observatories and the National Electronic Library for Health.

5.4.1 Web services

PHP does not include native support for web services. A number of PHP-based SOAP implementations were investigated. NuSOAP [68] was chosen because it could implement web services from a standalone class file while most other packages required code to be included in a server build operation, which would be difficult to implement because this action would require work to be undertaken by the service provider.

5.4.2 Design and implementation of web services using NuSOAP

In NuSOAP, four stages were used to develop web services to represent PHP classes.

Stage One: Development and test the class as normal

Stage Two: Describe the classes function's and message format in terms of WSDL

Stage Three: Add the function to the web service

Stage Four: Create a proxy class based upon the web service

Once the proxy class has been created, functions can be called in exactly the same way as a PHP class would be used if it were not based on webservices. Two functions were developed (Table 6):

	Function One	**Function Two**
Function name	searchMetadata	getContent
Action	Search the database and return basic information such as the resource title	Return all the metadata related to a specific resource
Input	text	content id (integer)
Output	Array of matching resources. For each resource: • id • title • description • link	Array of all metadata related to resource with the matching content id.

Table 6 Web service functions

5.4.3 RSS feeds

RSS is a format for syndicating news and the content of news-like sites [69]. Up-to-the minute news feeds are well suited to a syndication format such as RSS, however pretty much anything that can be broken down into discrete items can be syndicated via RSS. Once information about each item is in RSS format, an RSS-aware program can check the feed for changes and react to the changes in an appropriate way.

RSS feeds in the UK healthcare domain were championed by the National Electronic Library for Health (NELH) who provide a number of RSS feeds including 'What's new?' [70], 'Document of the Week' [71] and 'Hitting the Headlines' – a regularly updated feed analysing medical stories in the press [72,73]. In December 2004, RSS feeds were discussed at a workshop on interoperability in public health. Seidl [74] describes RSS as 'technologically straightforward, built on widely accepted standards, and offering simplistic functionality'. However, he also notes that similar 'disruptive' innovations have toppled established solution providers – not because the providers were unaware of the innovation, but because they believed it incapable of satisfying the demands of their established market.

An RSS feed for news items was developed by querying the APHOMS metadata and selecting only a subset of metadata, including title, description, published date, and a link to the resource. A number of further options, such as specifying a limit on the number of resources returned were made available through the querystring. The news RSS feed was reused in its entirety to develop a feed for event items, however a modification was made following feedback from NEPHO which modified the date ordering. Instead of ordering by published date, events were reordered to list upcoming events in ascending event date order, followed by historic events, listed in reverse chronological order. Both news and events feeds were used on the NEPHO homepage to list the top three items in each category.

The news RSS feed was reused again to produce a report feed. The code was slightly modified to include links to thumbnail images of each report, thus allowing the feed to be used for displaying the most recently published reports on the NEPHO homepage.

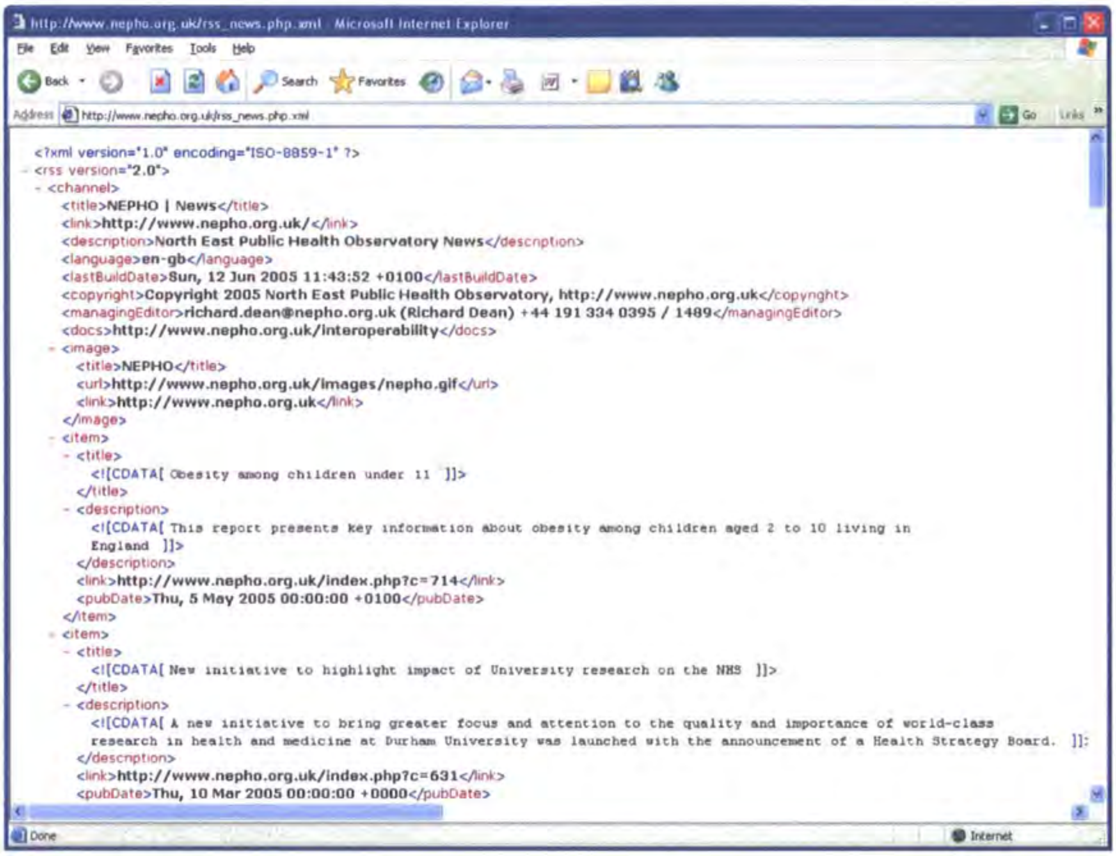


Figure 25 NEPHO RSS News Feed

An RSS feed administration function has been developed for the NEPHO website (Figure 26). This function took the URI of an RSS feed, and then downloaded feed information, such as the feed title and description. Feeds were regularly refreshed. Each item in the feed was stored in a database table. This was very easy to implement because the RSS standard specifies the message format and only a simple set of metadata is provided about each item. The feeds were programmed to automatically refresh every 24 hours by default. The storage of information about feed items in a database table is a simple example of metadata caching.

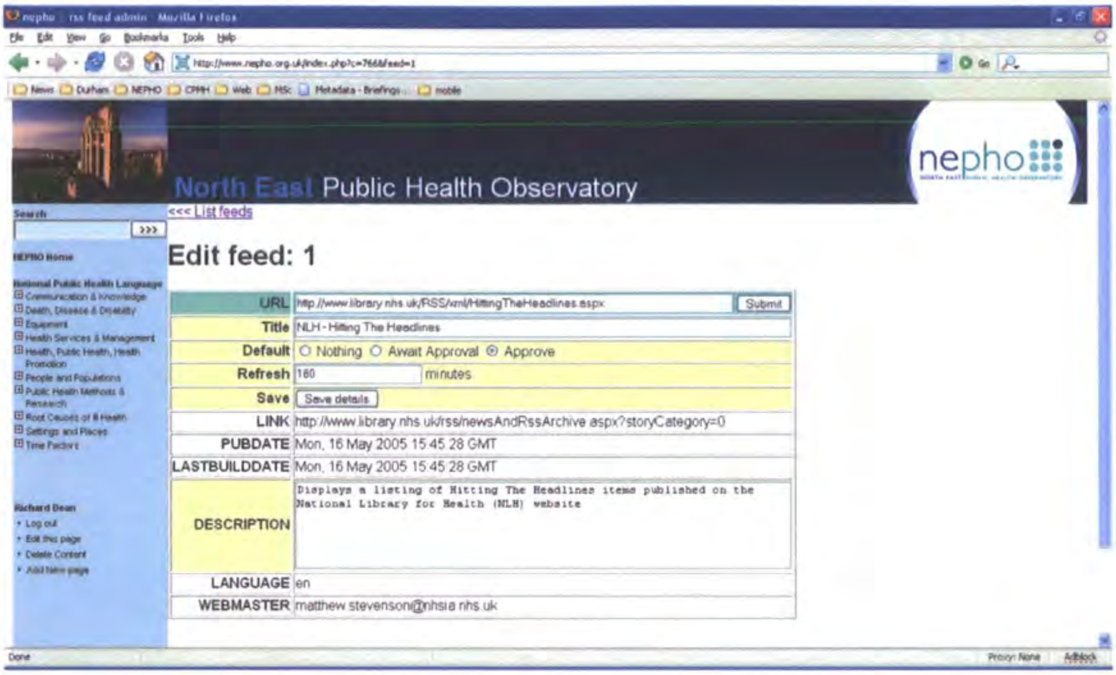


Figure 26 RSS feed administration screen

The RSS feeds have not yet been publicly promoted because NEPHO believed that external content should be moderated to select only suitable or interesting resources. The suitability of feeds also varies. Feeds such as the NPHL's 'Hitting the Headlines' feed [71,73] is aimed at the same audience as UK PHO websites and is unlikely to need moderation. Feeds from news organisations, commercial providers and specialist organisations are likely to need moderation to ensure the content is suitable to be published on NEPHO's website. A moderation function for items has been developed (Figure 27), however the problem of who should be responsible for moderating resources will not be addressed until the new member of staff is appointed.

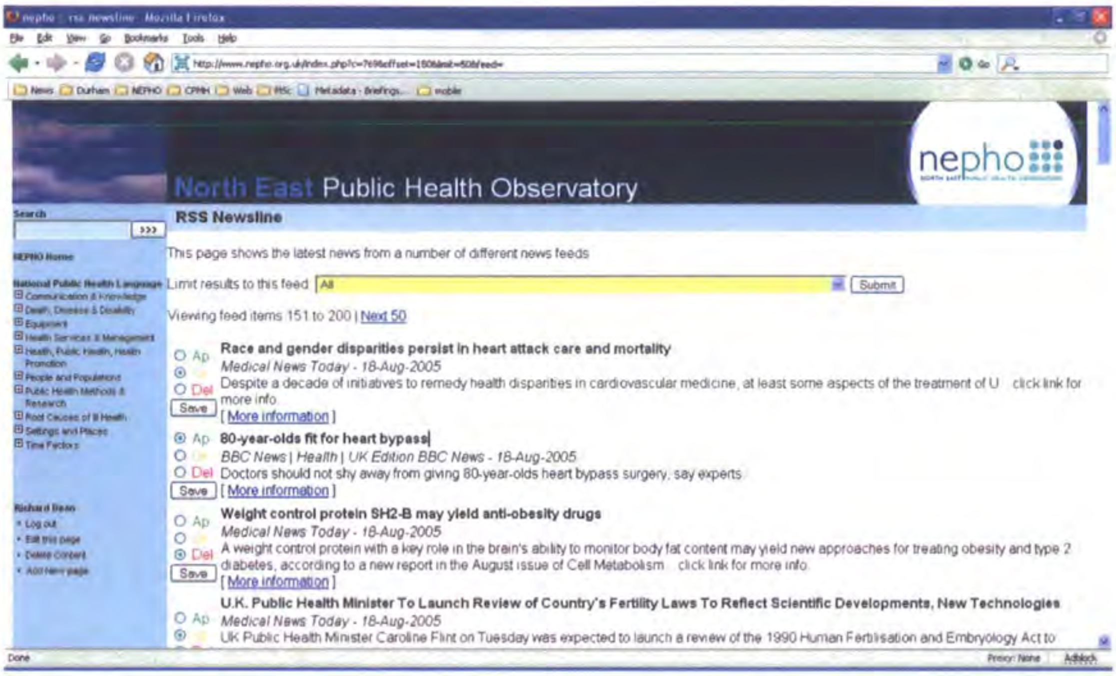


Figure 27 Syndicated RSS Feeds on the NEPHO website including administration functions

5.5 Action Four: Develop a front-end for viewing metadata-based resources

5.5.1 Navigation

The new NEPHO website used a different navigational structure to the existing three-tier hierarchy. A new 'relationships' table was used to store the relationship between each resource. Resources were also classified as parent / children of other resources. The NEPHO homepage could be considered as being at the top of the family tree with a number of child relationships to pages such as 'North East Region', 'What is a PHO?', 'About NEPHO' and so on. A child page such as 'About NEPHO' had a number of children of it's own such as 'Meet the Team' and 'How we work' along with 'NEPHO Homepage' as a parent.

A new menu function was programmed. Some users reported that they preferred the original drop-down JavaScript menu, however this option was not accessible to some users and did not fit the more complex relationship model. Links to each resource were created by linking to the index.php script, with the unique resource ID number as a querystring. For example, the 'About NEPHO' page has the unique resource ID number 122. A link to the resource was created by appending the querystring c=122. Where resources provided an alternative title, the menu script used this title if shorter than the main title. This was used for resources with long titles such as a page titled 'Regional Development Agencies (RDAs)', which was shortened to 'Reg. Development Agencies' to ensure the link remained on one line.

5.5.2 Use of web services

The new front-end used the basic APHOMS webservice to retrieve metadata for each page. No direct queries were made to the database in order to retrieve resource metadata. A template was used to display specific metadata in specific areas of the page. The template placed metadata such as title, keywords, publisher and language in the head of the html page. In the body of the page, the template placed the resource 'Title' at the top of the page, followed by any related thumbnail images on the right-hand-side, then the resource description, information about related files, and other metadata, including publisher, author, and dates of publication, where appropriate.

The basic searchMetadata web service was implemented on the site. More advanced search options should be added once resources have been sufficiently tagged.

The news, events and reports RSS feeds were used to create dedicated news, events and reports pages. All feeds were used on the NEPHO homepage. Figure 28 illustrates the metadata elements, XML feeds, and web services used to create the NEPHO homepage.

Figure 28 Summary of metadata used on NEPHO homepage

5.6 Action Five: Work with other observatories to develop interoperability

Three organisations – NEPHO, SWPHO and WMPHO used an incremental development approach. Basic search mechanisms were developed to allow searches based on plain text. The levels of interoperability described by the APHO interoperability group were commonly cited as lacking sufficient technical information. Most of the other public health observatories formed a consortium to redevelop the 'PHORMS' system developed by ERPHO and SEPHO. This system had been highlighted at the start of the project as an example of a site that was already interoperable. NEPHO developed a prototype search tool in November 2003, however the subsequent redevelopment of the site meant that this web service was withdrawn and access to new web services will not be available until the consortium are ready to begin testing their new site.

Work done to implement search-and-retrieve web services between NEPHO and WMPHO and between NEPHO and SWPHO are described below. Further details about the approaches taken by other observatories are provided in Chapter six.

5.6.1 Trial web services with SWPHO

SWPHO provided a web service with a function called 'BasicSearch'. The function used a simple string to query SWPHO's database and return metadata about matching resources.

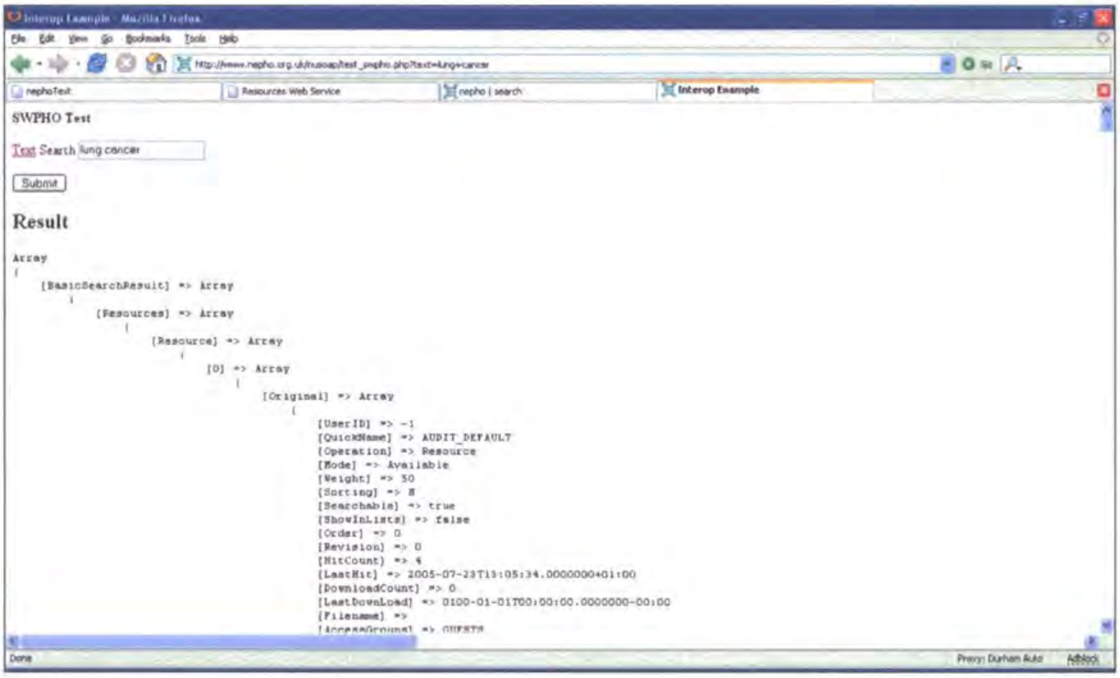


Figure 29 NEPHO website searching SWPHO metadata.

The SWPHO webservice returned a complete set of metadata about matching resources. This proved to be slightly troublesome because some searches returned a large number of results, causing the NuSOAP parser to run out of memory. Most of the metadata fields were incomplete, and no guidance was given about which fields would eventually be completed. SWPHO's web developers acknowledged that getting resources tagged was proving more difficult than expected and agreed that a final webservice would use fewer metadata fields than in the trial version.

SWPHO conducted tests on the web service provided by NEPHO. SWPHO's web developers were able to connect to the web service, call the methods provided and get results back though Visual Studio.

5.6.2 Trial web services with WMPHO

WMPHO provided a similar webservice to SWPHO. Their webservice ran a basic search based upon a plain-text search string. The code used to test SWPHO's webservice was reused. The only change required was to the webservice URL, the function name and the parameter name. This change was very easy to implement. As the format of the returned message was different, the metadata message formats need to be semantically mapped. This proved to be much more time consuming.

For each resource returned, WMPHO's web service returned the following metadata fields: id, title, type, description, source, creator, date, time period, and URL.

Figure 30 WMPHO website searching NEPHO metadata.

WMPHO developed a webpage to test two NEPHO web services – a free text search and an NPHL search. This trial was successful, however the trial has not been progressed to being implemented on public area of WMPHO's website.

NEPHO developed a basic caching mechanism based on WMPHO's webservice. Whenever the WMPHO search function was called, the results were stored into a table called 'my_cache', along with the original parameters. When the same original search was next used, data could be retrieved from the cache. This trial was implemented on NEPHO's NPHL page, as a test to investigate how well the NPHL tagging mechanism used by NEPHO could work with organisations such as WMPHO who did not use the same tagging mechanism. Most NPHL tags returned results from the WMPHO database, thus indicating that the use of differing tagging systems will allow interoperable methods to view resources; however the quality of the results returned may vary.

5.7 Chapter Summary

This chapter explained key developments implemented as work of this thesis in order to make the NEPHO website comply with the e-Government standards, and the interoperability criteria defined by APHO. An important part of the work was to develop a common method to implement each metadata element. The reasoning behind this approach is that there are a large number of metadata elements which would be time consuming to program individually

Chapter 6 Reality

6.1 Chapter Abstract

This chapter describes progress made by other PHOs on the interoperability project by June 2005. The range of solutions reflects technical and political reasoning within each organisation. Five different solutions have been developed. The North East, North West, West Midlands and South West PHOs developed their own solution. The remaining observatories formed a consortium to develop a solution based on ERPHO's existing website.

The rest of this chapter contains an overview of work undertaken by the West Midlands PHO, South West PHO, North West PHO, and the consortium of other PHOs. The progress of the interoperability project as a whole is dependant upon different organisations working closely together. The chapter describes work undertaken to develop interoperability between organisations. By June 2005, most organisations had begun development work on their new sites. A number had developed trial web services; however no organisation had completed the project criteria. Communication between organisations was infrequent. For a project about developing 'interoperability', this should prove cause for concern.

6.2 Progress of other PHOs

In June 2005, the interoperability group were required to produce a progress report that would be submitted to a meeting of the APHO board of directors meeting in early July. Each PHO completed a table of progress and submitted a short written statement on their progress. Each organisation's progress was discussed at a meeting of the interoperability group in late June, and a final report was submitted to the APHO board of directors on 6th July [76]. The report has been used as a starting point to describe the progress made by each observatory by June 2005. Further information has been collected from organisations via telephone interviews and via email.

6.2.1 WMPHO

West Midlands PHO's website was originally a mix of static html pages and a dynamic library of public health 'grey' material – a collection of documentary material not commercially published or generally available [75]. This was administered in a Microsoft Access database and dynamically linked to the website using the Active Server Pages scripting language (ASP). The interoperability project required the entire site to be catalogued in a database, and subsequently an amount of redevelopment work was required. The new site was developed in SQL server.

By June 2005, the existing resources had been transferred to the new database which was being populated with complete metadata tags. WMPHO's grey material library had already been tagged using the PHITS language. WMPHO decided to continue using PHITS rather than the NPHL because of the substantial costs involved in developing a new tagging system and then retagging existing resources. WMPHO felt that the NPHL should be given time to 'bed in' and that the PHITS system was more established.

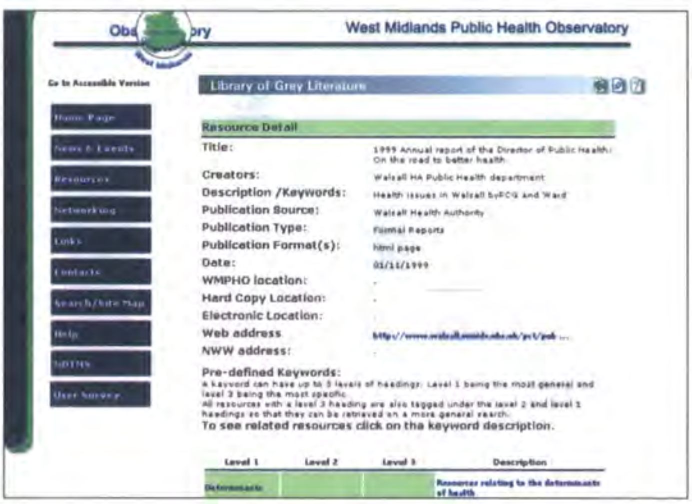


Figure 31 WMPHO Grey Literature Library, including PHITS keyword tagging

NEPHO and WMPHO regularly communicated with each other. Each has tested solutions developed by the other, as described in section 5.6.2.

6.2.2 SWPHO

In late 2003, plans to merge SWPHO with the South West Cancer Intelligence Service (SWCIS) were announced. There were a number of phases in the merger. The planning phase commenced in January 2004 with the aim of setting up a shared resource by April 2004 with a view to a full merger in 2005.

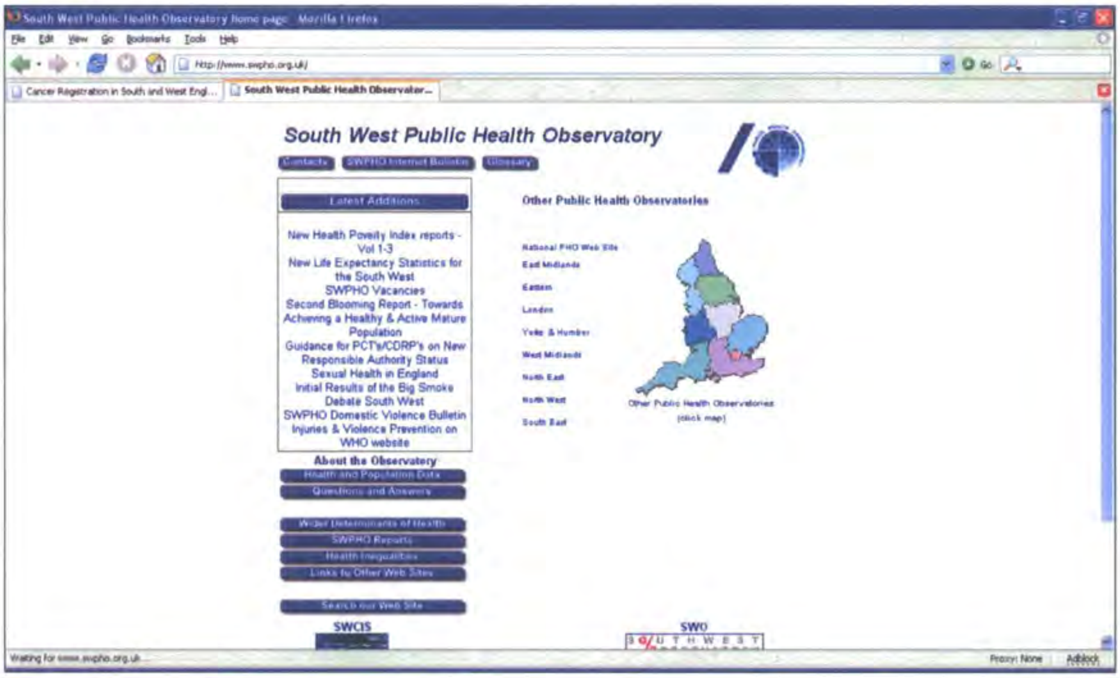


Figure 32 Existing SWPHO Website

The existing SWPHO and SWCIS websites were both collections of static HTML files. SWPHO investigated a variety of solutions, including the consortium led by ERPHO. They decided that a custom built solution would be most appropriate. A new developer was appointed to build a new site which would incorporate the existing SWPHO and SWCIS resources, and meet the interoperability criteria.

Development of the new site initially focussed on defining the database setup and administration functions for entering resources and defining metadata tags. This work was completed by May 2005. The programmers then concentrated on developing the user interface, design and operation of the website while the SWPHO staff were expected to tag resources in order to populate the resource database. This is a similar development strategy to NEPHOs.

While the technical implementation proceeded smoothly, the web developers were "continually frustrated at the lack of progress on the resource tagging" [76]. This was seen as a major stumbling block for the entire project which could lead to a considerable delay in the launch of the new site originally due for launch in September 2005.

NEPHO and SWPHO have successfully implemented a basic interoperability trial, as described in section 5.6.1. These trials were successful. Further work remains to be done as SWPHO's new site is still in development and both sites remain largely untagged.

6.2.3 NWPHO

NWPHO led a consortium of public health organisations based in North West England developing a new region-wide public health portal. This uses a common database, based on Microsoft's Sharepoint software [77]. Each organisation will be able to create their own branded area and upload resources. The single database will allow easier resource sharing between the organisations.

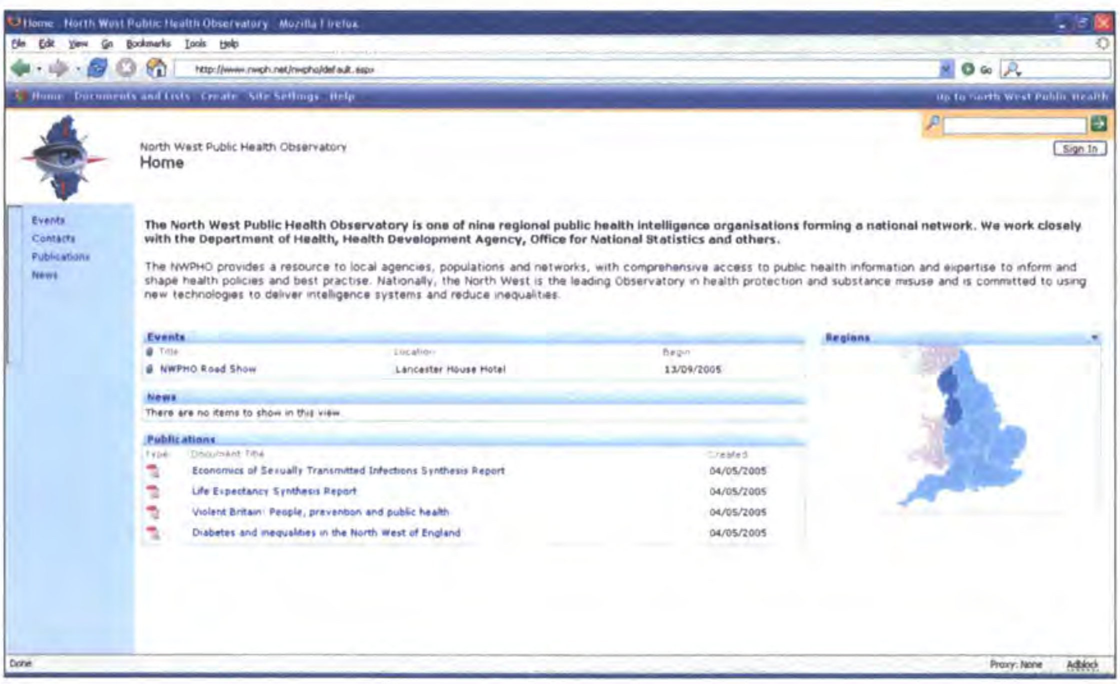


Figure 33 North West Public Health Portal - NWPHO homepage (in development)

As NWPHO's new site was being developed with other health organisations in the North West, agreeing to APHO standards proved difficult. For example, the APHO interoperability requirements stated that each PHO would use a common list of keyword tags. This functionality was not required by other health-based organisations in the North West – although it could be argued that PHOs should be encouraging regional health organisations to use the classification scheme. NWPHO's existing website contained more archived material than other observatories which would require keyword tagging. NWPHO decided not to comply with the keyword tagging requirement, and did not implement a keyword tagging system on their website.

NWPHO developed a test search function which could accept keywords and search for appropriate resources by matching the keywords to NWPHO's non-keyword based metadatabase and returned appropriate results. NWPHO claimed this search mechanism would provide a similar level of accuracy as when searching keyword tagged resources. NWPHO have not made any web services available for NEPHO to test. Therefore the accuracy of this search method could not be verified. The consequences of NWPHOs decision are evaluated in Chapter 7.

6.2.4 Other PHOs

The Eastern Region Public Health Observatory (ERPHO) lead early work in commissioning a metadata-based website. The solution was called 'Public Health Observatory Resource Management System' (PHORMS). PHORMS version one was released in Autumn/Winter 2002 to mixed feedback. Many found it comprehensive but a common criticism was that people found it 'too complex'. In 2003, a number of modifications were developed, leading to a 'version two' development. This release was developed in conjunction with the South East PHO whose existing website needed redevelopment. ERPHO could share development costs, while SEPHO didn't have to start from scratch. After the interoperability project was approved, the Yorkshire and Humberside, East Midlands and London Health Observatories joined the PHORMS consortium in order to develop a new e-GIF compliant and interoperable version of PHORMS.

Public sector procurement projects with a budget over £99,000 are subject to EC treaty principles of non-discrimination, equal treatment and transparency [78]. Tenders above this threshold must be published in the Official Journal of the European Union (OJEU, formerly the OJEC). The consortium developing PHORMS were therefore legally bound to follow the OJEC tendering process. Requirements specifications were developed between consortium members and the external company who had developed earlier versions of PHORMS. After the contract was advertised in the OJEU, a number of companies responded. Each company was required to submit specifications of their solution and to go through a number of selection processes. The consortium blamed the protracted tendering process for early delays in implementing the project. After several months of describing requirements to potential suppliers and reviewing tenders, solutions proposed by other suppliers were deemed unsuitable, and, perhaps unsurprisingly, the company who had developed the requirements specification and the two previous versions of PHORMS were selected.

During the implementation phase of the PHORMS project, a number of measures were made in order to reduce development costs. The existing ERPHO-SEPHO implementation used web services to communicate in real-time. In the new development, the live search was supplemented with a metadata cache. Each night, the PHORMS sites would each download each other's entire metadata in order to refresh the local cache. The consequences of using cached metadata are evaluated in Chapter 7.

6.2.5 Other organisations

The APHO interoperability project was funded to develop interoperability standards between the nine English PHOs. A Public Health Observatory function in Wales is also being developed after the Health (Wales) Act 2003 [79] stated that the Wales Centre for Health must, 'develop and maintain arrangements for making information about matters related to the protection and improvement of health in Wales available to the pubic in Wales'. Similar observatory functions are being established in Scotland by Health Scotland and in Ireland as part of 'Ireland and Northern Ireland's Population Health Observatory'. Both Wales and Ireland are developing new websites to agreed APHO / e-GIF standards. Representatives have been added to the APHO interoperability group, but funding has to be sought elsewhere.

6.3 Interoperability between PHOs

Interoperability with other organisations has been slow because partners were often at different development phases or were too busy working on their own solution to develop joint solutions.

An interoperable search in conjunction with other organisations can only be developed once more basic steps had been completed, including:

- Requirements Specifications
- Design for database and administration functions
- Implementation of database and administration functions.
- Resources are entered into the system

The APHO requirements were written in plain English. The PHO directors had appointed a clinician to be chair of the interoperability group. This was because they believed a "non-technical" person could report back to the directors in a language that they would understand. The flip-side of this argument was that the requirements were too vague for computer programmers. Sommerville [80] identified three types of problem associated with requirements written in plain English:

- Lack of clarity – requirements should be precise and unambiguous.
- Requirements confusion – the functional and non-functional system requirements are often mixed up with the system goals and design information.
- Requirements amalgamation – requirements should be expressed individually.

Solutions developed by the West Midlands and South West PHO have been tested by NEPHO and these tests are described in Chapter 5. The solutions being developed by NWPHO and the PHORMS consortium are still in development, and have not been made available for NEPHO to test. However, a number of PHOs claimed to be interoperable with one another in the report submitted to directors in June 2005, despite being unable to test any web services in conjunction with NEPHO. These claims are investigated in the following section.

6.4 Progress to June 2005

The interoperability group's report to directors in June 2005 [76] listed a number of examples where organisations have developed trial web services with other organisations.

NEPHO had developed a number of interoperable solutions including text and keyword based search webservices which would fulfil the APHO interoperability criteria at a basic/intermediate level. The NEPHO webservices were made available to the interoperability group between the start of March and the end of June and each organisation was contacted several times, however other organisations were unable to test the webservices or provide details of their own solutions because they were too busy developing their own solutions. This proved to be frustrating because other organisations claimed in the report to directors that they were developing interoperability with certain other organisations while not mentioning to NEPHO's persistent requests. The report to directors therefore stated that NWPHO was interoperable with WMPHO and ERPHO, while ERPHO was interoperable with SEPHO and NWPHO.

In the following section, the thesis briefly looks at the interoperability developments claimed in the June report.

6.4.1 NWPHO - WMPHO

NWPHO claimed to be interoperable at an 'intermediate' level with WMPHO, however there is no mention of this on WMPHO's site, and no ability on NWPHO's site to actually perform the search – the only search function provided on the NWPHO website is through a Google search box that still listed Trent PHO and NYPHO, indicating the page hasn't been updated since the PHO reorganisation in April 2003 (section 4.2.1).

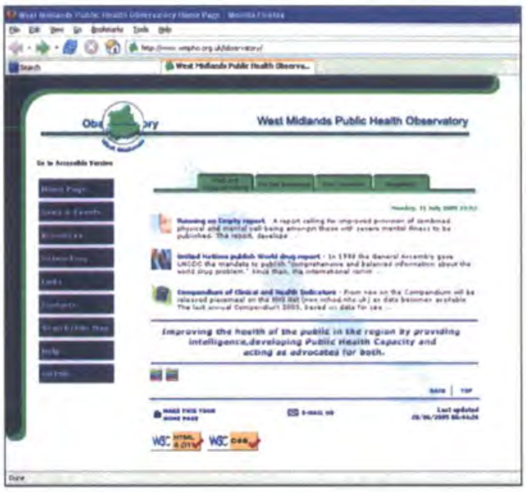


Figure 34 WMPHO homepage. The top story is titled "Running on Empty"

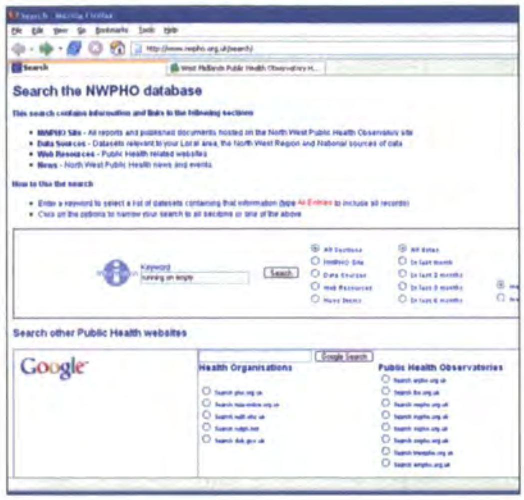


Figure 35 NWPHO's search page is quite complicated. A Google search is provided.

Upon further investigation, WMPHO claimed that NWPHO had successfully connected toWMPHO's web services, but the search was only a brief trial, and had not been implemented live on either website. NWPHO did not reply.

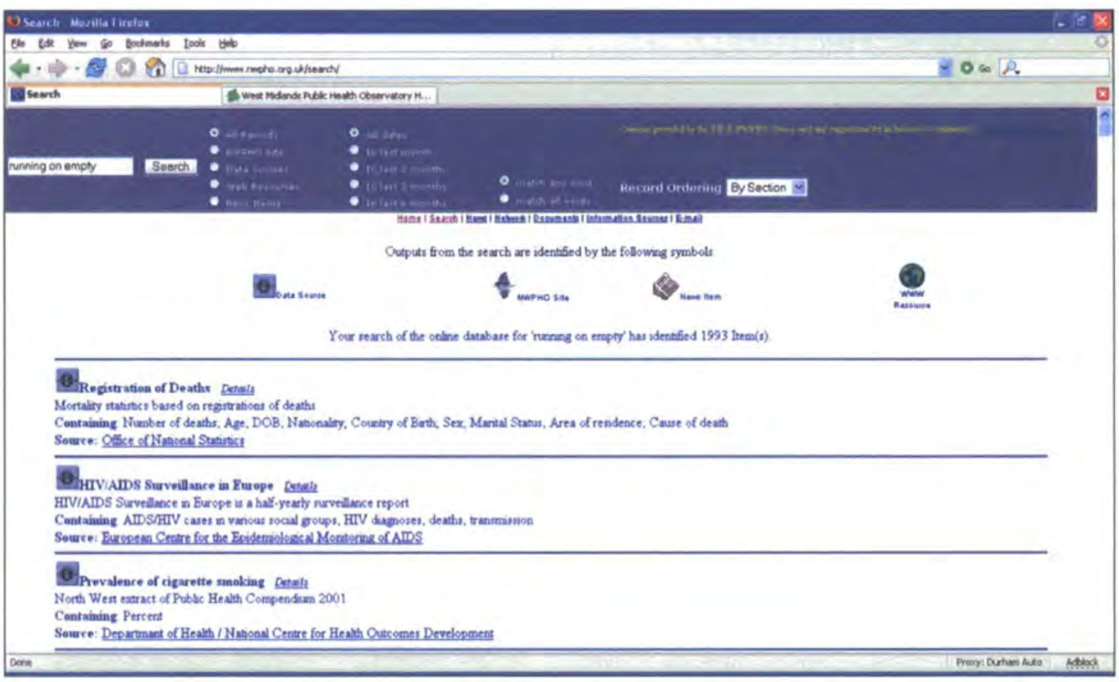


Figure 36 NWPHO search results aren't encouraging – 1,993 possible matches, none of which include WMPHO's report.

6.4.2 ERPHO - SEPHO

By June 2005, the Eastern Region PHO had developed interoperable searches between ERPHO, SEPHO and NWPHO. Interoperability between ERPHO and SEPHO's website was demonstrated after SEPHO joined the consortium in 2003. This functionality was unavailable for testing in June 2005 because of work being undertaken to implement the new site being developed by the consortium. This functionality would be restored when the new version of PHORMS was released.

6.4.3 ERPHO - NWPHO

The June report claims NWPHO and ERPHO are interoperable at an 'intermediate' level. This could not be verified during testing. The ERPHO website included an option to search NWPHO, however searches returned no results. ERPHO confirmed that this functionality was broken and would not be fixed while development work on PHORMS three was being carried out.

Figure 37 NWPHO's website lists a report titled "Drug treatment in the North West of England, 2003/04"

Figure 38 Search for 'Drug Treatment' returned no results.

6.5 Chapter Summary

Despite APHO agreeing standards for most metadata fields, developing interoperability between organisations has been slow due to political decisions within organisations, and because organisations are busy developing their own solutions. Keyword tagging is a typical example: Instead of using a common way of classifying documents, some organisations will use the National Public Health Thesaurus, others will continue to use the Public Health Information Tagging System, while another organisation has decided not to use any form of tagging language.

Five different solutions are being developed. Each solution has used slightly different metadata semantics. A number of PHOs have formed a consortium to develop interoperability, with a solution based upon a metadata caching mechanism built specifically for the consortium. Interoperability in the report presented to directors in June 2005 are not live implementations. Ironically, the solutions being developed are aimed at improving communication and data sharing between the organisations, yet the project could benefit from clearer communications between organisations.

Chapter 7 Results and Evaluation

7.1 Chapter Abstract

This chapter evaluates the work done as part of this thesis to make the NEPHO website comply with the list of requirements created in section 4.5. The chapter then evaluates the interoperability project from a national perspective based upon work undertaken by June 2005 by each PHO.

7.2 Overview of work done to NEPHO site

An action list was created in section 4.7. The action list contained five stages of development that should be followed in order:

- Action One: Develop online management for metadata elements
- Action Two: Develop online resource management
- Action Three: Create interoperable methods to access metadata
- Action Four: Develop a front-end for viewing metadata-based resources
- Action Five: Work with other observatories to develop interoperability

NEPHO's new metadata structure and metadata management function was developed in late 2004 (action one). Resource management functions (action two) and a web service to retrieve metadata (action three) were developed by February 2005. Existing resources were successfully imported into the new data structure.

The redeveloped NEPHO website front-end (action four) was launched in March 2005 for testing, and in April 2005 the new site replaced the existing site. NEPHO have worked with some other observatories to develop web services that search for metadata (action five) – however there is more work to be done.

7.3 Action One: Metadata Element Management and Data Structure

Section 5.2 describes the database structure developed for storing metadata. A web-based administration function was developed to manage the metadata elements. The online management function for defining metadata elements worked well, developing generic solutions throughout the project.

Element definitions were used to automatically:

- Create metadatabase structure (database tables and fields)
- Create administrative user interface for each element based on element type and cardinality
- Describe elements in WSDL
- Retrieve metadata about a resource using syntax from Dublin Core, e-GMS and APHOMS

Despite the best efforts of e-GMS and APHOMS, the boundary of what constitutes acceptable metadata is not very clear. Often there is scope for metadata elements to be interpreted in a number of different ways. Each differing solution increases the semantic complexity of making one metadata schema be understood by another implementation. Examples provided with e-GMS suggest that as much detail as possible should be provided in the <meta> tags – however, e-GMS leaves the format of the publisher element open for different implementations.

The metadata structure used by NEPHO is suitable to describe metadata to e-GMS standards. Future steps in interoperability, including the development of the semantic web will require use of an ontology language such as OWL [81] to semantically describe the metadata. For example, the NEPHO website stores annual reports published by each PCT in the North East. In the original metadata structure, the 'creator' element is stored as a single text field, but this would be more useful to provide more detailed information. A user may wish to view other documents created by a specific PCT, visit their website, email their director or find a GP in the area. Some of these examples will require intelligent agents as described by Tim Berners-Lee in outlining the semantic web. In the future, the use of ontologies and the semantic web may help the user achieve such a task. It is recommended that metadata should:

- be classified as accurately as possible by the use of specific fields
- common standards should be used and competing standards should be mapped

The database structure has remained very stable since it was introduced. While the front-end has been modified after feedback, the database has remained unchanged, with the occasional addition of new metadata elements. One change was considered to reduce the number of queries required when selecting information from linked metadata tables. The database structure requires a query to be executed for each linked table. As further linked element types are introduced and resources are more thoroughly tagged, the number of queries to be executed will be increased which may degrade performance. A more efficient solution could be developed once the service provider installs an update to the database server software.

7.4 Action Two: Administration functions

The new browser-based administration functions are easier to use than the previous mix of html editor program, ftp, and web-based content specific administration tools. Administrative functions have been provided through a standardised interface. The number of resources stored on the site which require a separate metadata entry has risen as linked files and thumbnail images require separate metadata in addition to just the report they are linked to. Very few resources meet the e-GMS and APHO standard for metadata tagging. Administrators have often neglected to complete metadata fields, while the existing resources that were imported to populate the site lacked a number of mandatory fields. Only thirty resources have been tagged using the NPHL; this leaves a vast majority of resources with no associated keywords. A majority of resources simply contain data in the title and description fields. Further work on the administrative function may improve the quality of metadata provided.

7.4.1 Incomplete metadata: e-GMS and Keyword tagging

Administrators often completed only the title and description metadata fields. It proved to be difficult to make administrators add full metadata for resources, while not making the task of adding a resource so laborious that administrators were discouraged from adding any content. This is a problem that has been observed by other observatories (section 6.2.2). Considerable effort was invested in developing a generic metadata element administration function, however the generic interfaces may require further work to ensure the quality remains comparable with custom-built interfaces.

Administrators at NEPHO are unsure who should be responsible for adding resources and day-to-day management of the website. The new system allows any member of staff to add or edit resources. There is no policy in place to ensure the author or the person who uploads a report completes the tagging exercise. Report authors see metadata tagging as a web-development issue, while web developers and site administrators often lack the domain knowledge necessary to correctly and consistently label reports about unfamiliar subject material.

In response to feedback from administrators, a number of improvements have already been implemented. For example, the creation of a new report has been improved by implementing a single page to edit metadata of all related files. A resource-specific 'wizard' was created to guide administrators through metadata entry for the organisation data type. This has proved to be successful, and should be developed to help guide administrators through metadata entry for other resource types. NPHL tagging, in particular, has proved difficult.

The vast majority of NEPHO's resources have not been keyword tagged, and consequently the majority of tags are not associated with any resources. Staff are unfamiliar with the tagging language. Administrators lack enough knowledge of the language to ensure they have selected all the appropriate tags for a resource.

The tagging system was improved following feedback from site administrators – as described in section 5.3.4. Crystal [60] observed that administrators "need better guidelines for determining the granularity of complex elements such as subject descriptors [keyword tagging]". Guidelines are clearly needed by all PHOs. Guidance should be created at an APHO level, thus ensuring the consistent application of metadata throughout all organisations. This will require training. When the NPHL was launched in 2004, the South East PHO agreed to provide NPHL training and guidance. They had played a large part in developing NPHL and they employed suitably skilled staff who could teach staff from other PHOs how to tag. Such training has never been provided because SEPHO were too busy in developing their new site. This is disappointing for organisations such as NEPHO, who have implemented the tagging system, but are still awaiting training in how to use it correctly.

A consequence of the continued lack of tagging training is that while browsing the NPHL on NEPHO's website, many terms are not linked to resources. The home page includes links to the ten NPHL top-level categories, but four of these categories don't have any resources tagged with the term. Asking users to browse a practically empty thesaurus will only damage the users perception of the project and discourage them from using the service in the future. Until resources are correctly tagged the NPHL top-level terms with no related resources should be removed from the homepage.

7.4.2 Rich-text edit function

The rich-text edit function worked very well however the chosen solution had some peculiar behaviour issues on different browsers. On gecko-based browsers such as Netscape and Firefox, the 'view source' function failed to insert any line breaks (Figure 39) which made editing the source code very difficult.

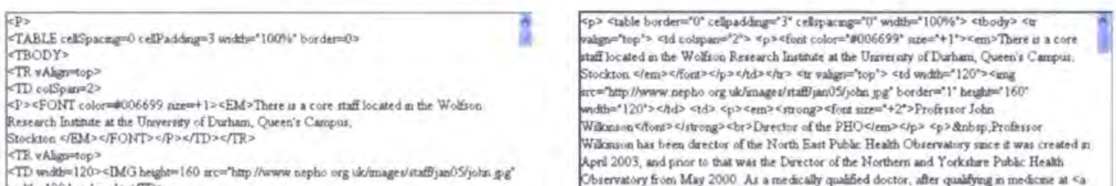


Figure 39 Difference in source code view between Internet Explorer (left) and Gecko (right)

A second problem was observed when pasting from a Microsoft Word file into a rich-text box. The word-file retained its formatting, which satisfied the requirements, however the source-code generated included a large amount of unnecessary formatting instructions. For example, a 'vacancy' page was created by pasting text from a Microsoft Word file into the 'description' field. The pasted text totalled 3,022 characters, including HTML tags to format headings and links. The same text pasted as plain text and formatted by using the tool was 1,955 characters. Furthermore, the additional tags included specific formatting which over-rode the default page style, introducing different font typeface and size. Neither issue was raised as being problematic by NEPHO – however it is seen as being somewhat annoying from a web development point of view.

7.4.3 Future Plans / Recommendations

NEPHO plan to appoint a new data analyst in October 2005 who will have specific responsibilities for website maintenance included in their contract. As such, they will be expected to assume overall responsibility for learning the NPHL and ensuring that NEPHO's resources are adequately tagged.

One method of improving metadata creation is to include a table of metadata on the inside front or rear cover of published reports. This practice has been implemented in some government reports, such as the e-GMS [47], however many publications such as the Department of Health's whitepaper 'choosing health' [82] still lack such information. NEPHO should ensure that future published reports include a table of metadata. This helps web administrators to correctly tag new resources because it makes the report authors provide basic metadata which would take a considerable time for a site administrator to find – for example, explicitly specifying the date range of a survey in a metadata table is a lot easier than having to scour the report for date ranges.

The South West PHO are developing an experimental automatic tagging function, which will be of interest for organisations like NEPHO who may be able to adopt the technology.

7.5 Action Three: Use of web services and XML feeds

Web services were used to deliver contents for each page on the NEPHO site and to provide a search function. XML feeds were created for the latest news, events and reports. These feeds were used to create the homepage, which used all the feeds to automatically display the latest resources (Figure 28). While XML feeds were not in the APHO interoperability criteria, they have proved a very simple method to allow resources to be syndicated in a common format by a large number of other organisations.

As the feeds comply with the RDF standard, the feeds can easily be viewed on devices such as handheld computers and mobile phones, or displayed in a similar fashion to email in a news reader application such as Mozilla Thunderbird (Figure 40).

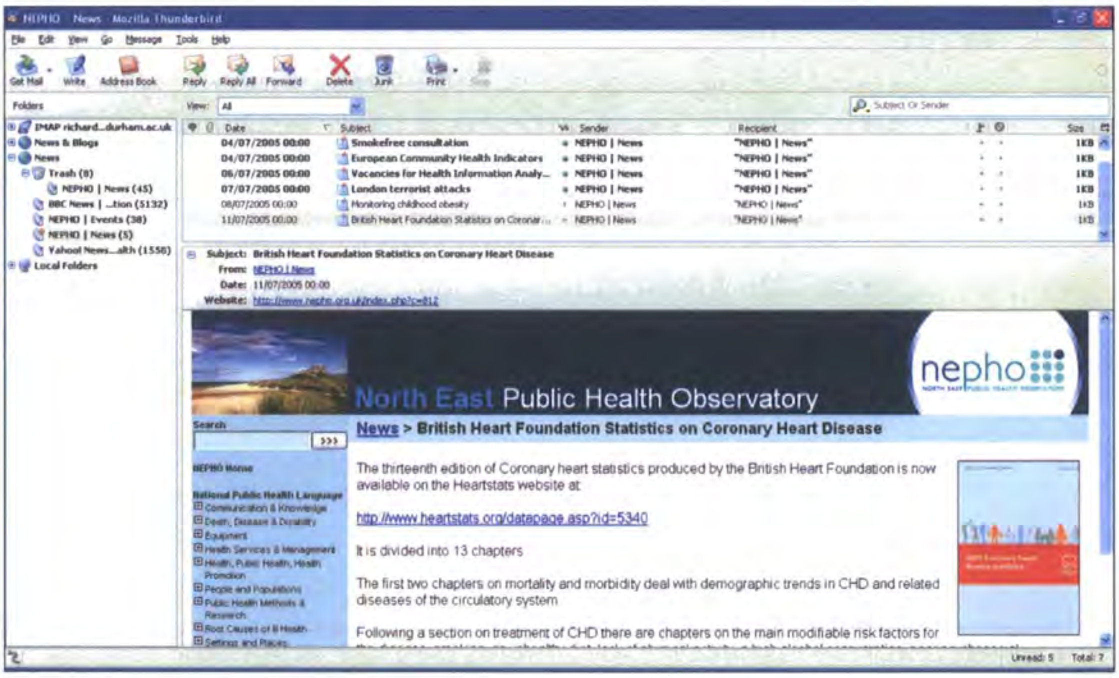


Figure 40 NEPHO's RSS news feed displayed in Mozilla Thunderbird

RSS feeds lack some advanced functions required for interoperability, however RSS is a common standard used for syndicating web-based content which has made the syndication of new feeds very easy. Work should be undertaken to support and develop further RSS feeds between organisations as a relatively easy first-step towards developing interoperability. NEPHO's RSS feeds should be publicised in places such as the National Electronic Library for Health's new RSS directory [83].

The getMetadata and searchMetadata web services have been used since the site was opened with no problems. While the web services automatically update to reflect changes to the metadata element definitions, further work will be required to allow refinements to searches.

7.6 Action Four: Front End, Common Access Control Policy and User Information

The new front-end developed for the NEPHO website is described in section 5.5. The site successfully uses metadata to display resources. Further work needs to be done to separate the data and design elements of the page. PHOs agreed an access control policy as some organisations planned to restrict content to a group of registered users.

7.6.1 Separation of data and design

At the start of the project, PHOs were not sure whether NHS web guidelines should apply to their websites. The guidelines stipulate backward-compatibility with old browsers which were unable to display style sheets. As the e-government push towards more advanced IT systems continues, such backward compatibility is seen as limiting the project and more modern standards should be adopted to further separate data from page design.

7.6.2 Access Control

The APHO interoperability criteria stated that PHOs would use a common access control policy to ensure the same level of access to restricted resources across interoperable websites. NEPHO developed a user authentication system to allow access to the administration functions. The project considered the possibility of restricting some resources to groups of users – for example, security was a major consideration when evaluating methods of file storage. A user registration procedure has been developed, however this has not been publicised because NEPHO do not plan to store restricted information on the website. The APHO access control policy will allow users to view metadata about restricted resources hosted on other sites, however they will not be able to download the resource without registering with the organisation who hosts the resource.

7.6.3 User Information

Each time a page is loaded, a 'hit' is recorded in the database. In June 2005, over 1,000 hits were received from users who had been referred to the NEPHO website from a search engine. If resources were only available to authenticated users, search engines would be unable to catalogue sections of the website and users would find if more difficult to find appropriate resources.

The main disadvantage of leaving the site open access to all users is that NEPHO do not collect much information about the site's users such as their profession, organisation, and research interests. This is useful for auditing website usage and has provided useful information for those PHOs who have implemented such systems. The system developed for NEPHO included a user registration page. This could be used for an optional user registration – despite the lack of any restricted resources. ERPHO have had some success in encouraging registered users to create their own resources, however, these resources still require moderation.

7.7 Action Five: Develop interoperability with other PHOs

While the previous action points have been implemented by NEPHO, the final action point requires work in association with other organisations. Progress by June 2005 was limited because other organisations were not in a position to develop web services with other organisations. There was a drive to implement a new metadata-based website by each organisation, often with little focus on how the final stage of interoperating between each metadata set would work. Funding for the interoperability project was quite complicated. The cost of developing an interoperable website was estimated based upon the cost of new hardware, software, implementation and staff training at a single observatory. This value was multiplied by the number of observatories, and the money was provided by the Department of Health.

Instead of money going into an APHO interoperability project central fund, money was distributed to each organisation, based proportionally on the population of the region. No money was provided for national project management. This funding formula has encouraged a number of different approaches to be taken. But the lack of national project management can be seen as a hindrance because of a lack of clear standards and goals.

In 2003, the South-East PHO adopted the online resource management system commissioned by the Eastern Region PHO, and developed by an external web development company. The two PHOs implemented a web service to share information between the two sites. Funding for the national project was based upon the costs of this implementation. The solution was not 100% e-GIF compliant, and the live-search of the SEPHO site set a precedent for the type of interoperability that would be developed by observatories. The APHO interoperability success criteria was based upon this model of using live web services. There is a problem with this approach, in that it worked fine for querying a single additional organisation, but the approach is difficult to scale to work with a large number of sites. Eight live searches would be required to search the resources hosted on all other observatories. Results from each observatory would have to be amalgamated and then re-ranked using some form of ranking criteria.

The plan to implement real-time search using web services for each PHO introduces a scalability problem termed 'quadratic complexity' – a specific interface from each implementation to each other implementation requires N × (N-1) possible connections between N applications (Figure 41). This means that for nine PHOs to implement distinct connections with each other PHO will require a total of 72 connections. If all 302 PCTs embarked on a similar interoperability project, a total of 90,902 interfaces would be required. This scalability problem was not thoroughly considered when the project was first initiated based upon the live web-service between ERPHO and SEPHO.

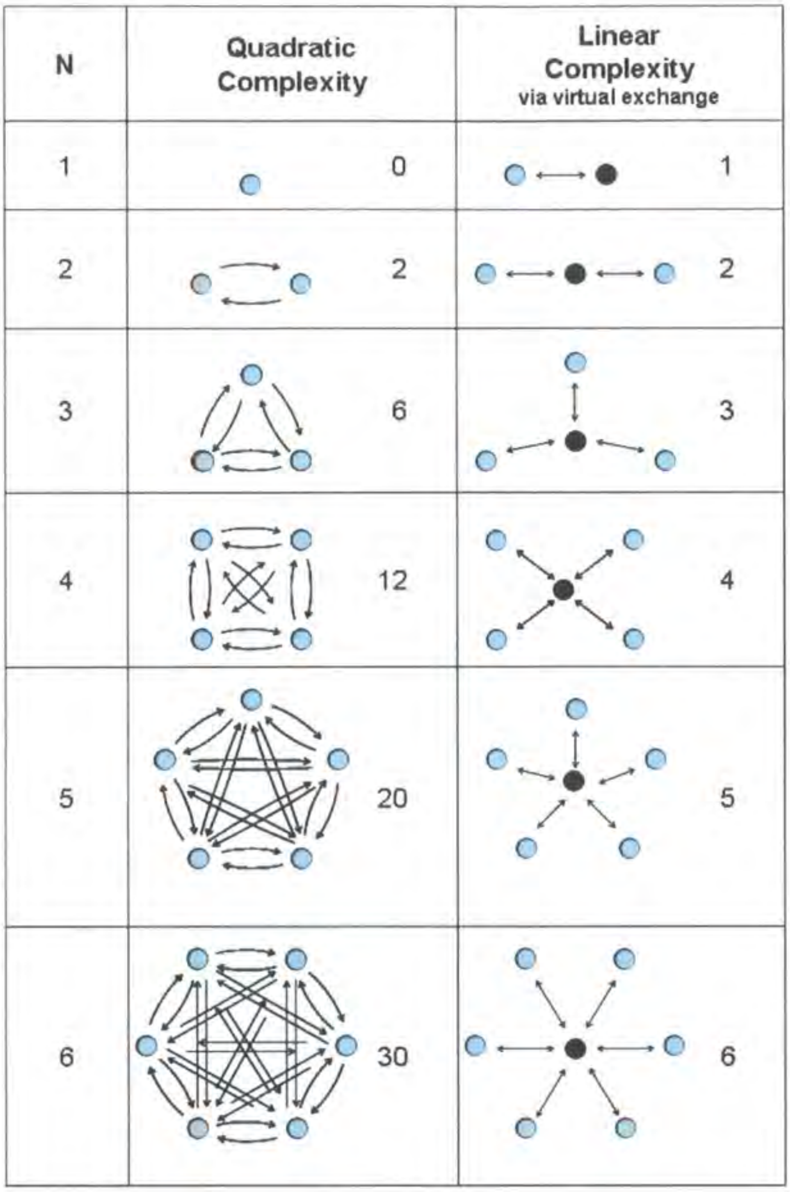


Figure 41 Quadratic vs. Linear Complexity

In addition to the scalability problem of integrating technical interfaces for performing the search, there is also a very important issue of developing a semantic understanding of the metadata format used by each organisation. Despite the best endeavour of e-GMS and APHOMS, metadata elements are still open to different interpretation. Graham et al [42] observe that 'web services technologies are great for discovery, description, and messaging of information, but they currently fall short of addressing the core semantic mapping problem'. Technical solutions to this problem remain research projects. The semantic web offers one vision; but it will take 'decades' to become reality.

PHOs are in the process of implementing a variety of e-GMS compliant web sites, however the semantics of each implementation will differ. PHO directors look upon the launch of a new metadata-compliant website as the primary objective for each organisation – however, as Graham et al [42] observe "The hype surrounding XML has made some believe that semantic mapping is no longer a big issue They're bitterly disappointed to find that, as always, a significant proportion of the integration effort is spent in this area". PHO directors may be equally disappointed to find that their e-GMS compliant websites are not automatically interoperable without further work to address the complexity issue of actually integrating data from the nine separate databases in any form of common way.

By contrast to web services, the much more simplistic RSS feeds were very easy to implement because they use relatively simple metadata and a standard message format. Without the semantic mapping concerns of the more complex data formats used in web-services, the number of RSS feeds available will continue to grow much faster than more complicated web services.

7.7.1 Caching Metadata

In this thesis, web services were developed to search one organisation's database from another organisation. A search of a single database produced acceptable results; however concerns were raised about the scalability of live searches across multiple websites.

A solution favoured by the PHORMS consortium was to store a local cache of partner's metadata and to update this information during a nightly update routine.

The advantages of caching data in general are described in section 2.5. To summarise, caches reduce the costs involved in performing an action. The same search algorithm is applied to all cached metadata from multiple organisations. This ensures consistent results from all organisations, as opposed to using different search algorithms implemented by other organisations.

Caching metadata is more complicated than a normal web-cache. They don't automatically solve the quadratic complexity problem as each local standard has to be converted into a format suitable for the cache. Web pages are catalogued by their unique URI and the plain text from the web page. Metadata needs to be catalogued by many different fields – titles, authors, publisher, dates, geographical coverage and so on. The metadata produced from the PHORMS consortia websites is in a uniform format which makes caching easy. This does not solve the semantic mapping problem, it negates it by using a common standard. Other organisations have developed different metadata structures. This requires a semantic mapping of each metadata element, thus making a uniform solution much more difficult to achieve.

Organisations implementing PHORMS plan to download the entire contents of each organisation's metadata database each night. This will use a lot unnecessary bandwidth as only a small proportion of items are added or modified each day. Resources are rarely removed from the database, therefore the amount of metadata will continue to rise over time, thus increasing the size of the nightly downloads. New resources will not be immediately available on other organisations websites. Sites will contain inconsistencies until the overnight update routine runs. A search on one site will not find a recently published report on another site. A resulting problem is that other observatories may create a duplicate entry on their local site. After the overnight update, both copies will be distributed to partner sites. A protocol to tackle duplicates is required to ensure the inconsistency is corrected by deleting one of the resources – however which organisation deletes the resource, and how this is reflected on the other sites is undefined – there is the potential for further inconsistencies coming from different organisations realising the mistake and deleting different resources.

7.7.2 Keyword Libraries

Organisations followed a number of different approaches towards the use of keyword tagging. Most sites chose to tag with the NPHL, one chose to retain their existing tagging system using PHITS and another chose not to use any keyword tagging system. Each organisation's decision has been outlined in Chapter 6. The choice of different languages affects the functionality and quality of the solution. In the following section, the user experience will be evaluated for searching between sites with different languages and between an NPHL tagged site and a site that does not use any tagging language.

7.7.2.1 NPHL – PHITS searches

The new NPHL language was adopted by organisations who were developing new sites or adding tagging to existing sites. PHITS had already been implemented by WMPHO who had invested considerable effort in tagging their grey literature library with PHITS tags.

As the NPHL was built from the combination of the HDA's HealthPromis thesaurus and the existing PHITS tagging system, about 95% of PHITS tags are included in the NPHL. Sites developed based on the PHITS system should be able to retrieve resources which use the corresponding tag from the NPHL. For example, both languages contain the term 'lung cancer' so searching for the keyword across both sites would prove straight forward. As the NPHL provides a larger number of terms, many resources are tagged differently. For example, a resource about stomach cancer would be tagged using 'stomach cancer' in the NPHL, but the most appropriate term in PHITS would be 'Other Cancers'. Searching based solely on matching keywords would fail for both the NPHL based and PHITS based site.

Possible solutions for this problem would be to introduce a mapping between every tag on each system, however to manually match tags would be time-consuming and would not lead to much greater reliability because there simply isn't a PHITS term referring to 'stomach cancer'. A second option is to perform a variety of different searches using individual words as keyword. For example; searching where keywords include 'stomach' and 'cancer', but not just 'stomach cancer'). The problem remains, however because there are no PHITS terms with the word 'stomach'. A search may return items which match the 'cancer' term, but would include resources about 'lung cancer' and so on. In the case where equivalent terms cannot be found between the languages, the search engine may be forced to resort back to using the keywords in a plain text search on other metadata fields. This solution is discussed below.

7.7.2.2 NPHL v No Controlled Vocabulary

One organisation participating in the interoperability project decided not to implement a keyword tagging system.

Websites which do not classify resources with keywords selected from a controlled vocabulary perform searches based on the plain text metadata fields associated with resources. This is typically a search of the 'title' and 'description' elements, which can often match some terms, but the quality of the result is likely to be reduced – for example, a report about "smoking" may not mention the NPHL keyword "tobacco consumption", despite the obvious relationship between the two terms. NWPHO have developed a web service which accepts keywords and returns appropriate resources, however NWPHO itself does not wish to keyword tag its resources so the keywords have to be matched against plain text fields, such as the resource's title and description.

7.7.2.3 Use of differing keyword libraries on cached metadata

Section 7.7.1 discusses the development of metadata caching mechanisms as opposed to live searches. Resources from other organisations are likely to present different results from the same search criteria depending on whether the search is performed using the original organisation's customised web service compared to the generic approach of a cache. If NWPHO resources are downloaded into the metadata cache, they would not have any associated keywords and the cache would not invoke NWPHO's keyword search web service. This would make it very difficult to find the resources by using the cache's default programming.

7.7.2.4 Evaluation of keyword tagging

Eight out of nine observatories plan to use some form of keyword tagging system. Observatories who have implemented keyword tagging have not had to re-classify all their resources when a new tagging system was developed, however as there are more specific tags, some resources are likely to benefit from additional tagging.

Some organisations who have implemented a tagging system have found it difficult to get administrators to tag resources, despite a number of improvements to the tagging mechanism. Administrators would benefit from training in how to tag resources correctly. As a minimum, e-GMS mandates use of the Government Category List for tagging resources.

The NPHL includes 80 e-GMS terms, so tagging doesn't have to be carried out twice. NWPHO plan to set the e-GMS keyword criteria to the generic "Health" keyword for all their resources. Technically, this complies with the e-GMS requirement; however the approach is unhelpful in terms of developing a useful search tool between all PHOs. Without central PHO project management, there is little action that can be taken to redress this issue.

7.7.3 Evaluation of national interoperability

The APHO interoperability criteria simply state that there is an ability to search another PHOs website and comply with e-GIF requirements. This ability is provided through live webservices. The use of live webservices for searching is analogous to early days of networking where every node was required to be connected to every other node. As networks expanded and more nodes were added, it became impossible to connect everything directly to everything else.

A virtual exchange (Figure 41) is a hub-and-spoke integration architecture where all data formats within a domain are mapped to a super vocabulary. They offer an architecture that scales better with the number of applications. In the architecture, all applications know how to communicate with a standard vocabulary that the exchange provides. This leads to linear complexity for the integration problem as a whole, but Graham et al [42] observe that the initial cost of integrating the first few applications are radically increased. Information exchange can't begin before a standard vocabulary is agreed on. Even within narrow vertical industry segments, this can take years to develop. It must cover the needs of all exchange participants. The architecture also poses significant change management challenges; it becomes difficult to evolve a vocabulary that so many applications depend upon.

7.7.4 Interoperability with other organisations

NEPHO have worked with regional organisations such as the North East Regional Development Partnership (NERIP) on early work to develop interoperability at a regional level. Early implementation work will focus on using the RSS feeds developed for news, events, and reports which is expected to the development of a regional cache of organisational metadata. At a national level, NEPHO have continued to work with organisations such as NICE on a national public health knowledge network, however progress has again been slowed by ongoing political uncertainty, such as the merger between HDA and NICE and plans to merge large numbers of primary care trusts by mid 2006. As a relatively stable group of organisations, Public Health Observatories should find themselves in an ideal position to provide support to new and merged organisations who wish to embark on similar projects to develop e-Gov compliant websites that could potentially interoperate with PHOs.

7.8 Chapter summary

This chapter contained an evaluation of the work undertaken to develop the NEPHO website in order to comply with the interoperability requirements described in chapter four. Work to make the NEPHO website comply with APHO and e-Gov requirements has been completed, however administrators need to ensure that appropriate metadata is provided for each resource hosted on the site. The final action point – to develop interoperability with other organisations has not been completed as this action relies upon other organisations who are not yet at a similar developmental position.

Chapter 8 Conclusions and Further Work

This chapter contains conclusions in the form of improvements and benefits gained by NEPHO, and any issues raised during the project. The chapter also refers back to the action list and APHO interoperability criteria defined in chapter four. As the APHO interoperability group had not completed the project by June 2005, a list of outstanding work is provided, along with a brief description of future work that follows on from the knowledge gained while implementing this project.

8.1 Project Successes

While there is still work to be done to make all the PHOs interoperable at the 'advanced' level criteria, this project should be considered a success. NEPHO have gained the following benefits:

- New e-GIF compliant website
- Recognised lead on developing interoperability between public sector organisations in the North East of England (section 7.7.4)
- Increased efficiency as most staff can upload reports without having to email web developer, use HTML editor and FTP package
- Content management system automates a number of metadata fields and ensures users don't accidentally delete files as they are backed up
- The file-upload function and administration function has been developed on another NEPHO project to allow remote users to upload confidential data sets without making the data publicly available. In other regions, data has to be securely couriered between offices.
- Use of the National Public Health thesaurus helps users find the resources they require, additional metadata fields help users browse for related items
- New site launched at least six months before other PHOs despite smallest budget based upon region size.

8.2 Summary of APHO Interoperability criteria

Chapter four described the 'advanced' level of interoperability (summarised in Table 4). NEPHO's progress towards achieving these criteria is summarised in Table 7, below.

Advanced level criteria	Advanced	NEPHO Progress
Ability to use PHITS or NPHL for tagging and retrieval	✔	✔ NPHL tagging
Resources keyword tagged with PHITS or NPHL categories	✔	✔ Some resources have not been tagged yet
Use common access control	✔	✔ See Note 1
Have a dynamic website (i.e. using a database)	✔	✔ See Note 2
Is fully e-GIF compliant	✔	* See Note 2
Use APHO standard for type encoding (PHRTES)	✔	✔ Used in tagging system
Use APHO standard for minimum metadata set	✔	✔ However legacy resources may not be fully tagged
Use APHO standard for common XML Schema	✔	✗ See Note 3

Table 7 Summary of NEPHO progress against APHO 'advanced' level interoperability criteria

Note 1 – The APHO group agreed a common access control policy for sites who restrict access to some resources. NEPHO do not impose such restrictions, but agreed to the policy which allows NEPHO's users to see metadata about restricted resources, but not the resources themselves.

Note 2 – As e-GIF continues to develop, full e-GIF compliance is a moving target. Both the e-GMS and e-GIF are currently being reviewed and new versions will be published in 2006 [84]. At present, e-GMS complaint metadata administration functions are available, however, metadata tagging is incomplete – therefore many resources do not comply with the mandatory requirements defined in e-GMS. This is simply a question of dedicating time and resources to tagging the existing resources on NEPHO's website. The thesis requirements to deliver an e-GMS / e-GIF complaint administration function have been completed – however it will be an ongoing task for administrators to ensure that both new and existing resources are fully metadata tagged using the system developed.

Note 3 – As part of the requirements analysis stage of this project, NEPHO proposed an APHO-wide common XML schema, as required by the 'advanced' level criteria. This schema was discussed in June 2004 at a meeting of the APHO interoperability group. Members of the PHORMS consortium agreed in principle that a common metadata schema would be used and requested to adopt the schema for PHORMS-specific metadata requirements before it could be agreed nationally. This proposal has since stalled as the PHORMS consortium produced a PHORMS-specific metadata schema. This is unacceptable to non-PHORMS PHOs who have not been able to contribute to the standard.

8.3 Summary of action list progress

In order to successfully comply with the 'advanced' criteria, a list of NEPHO-specific requirements and actions was drawn up. Chapter five describes how each of these actions has been implemented, and chapter seven provides an evaluation of each stage. By June 2005, the first four actions have been completed, however developing interoperability with other PHOs (action five) proved to be difficult, as described in chapters six and seven. Progress against implementing the action list is described in Table 8, below.

Action	Has action been completed?
Action One: Develop online management for metadata elements	Yes
Action Two: Develop online resource management	Yes
Action Three: Create interoperable methods to access metadata	Yes
Action Four: Develop a front-end for viewing metadata-based resources	Yes
Action Five: Work with other observatories to develop interoperability	Partially - trials have been carried out with SWPHO and WMPHO, but further work relies upon other observatories

Table 8 Progress achieved against action list

8.4 Discussion

This thesis has documented one organisation's implementation of a set of requirements laid down by the Department of Health, PHO directors and the APHO interoperability group. While NEPHO have successfully implemented these requirements this did not mean the end of the project, as interoperability with other organisations is still to be completed. Many of the issues and problems raised during the implementation of this project can be tracked back to when the project was first initiated.

Chapter four describes how the project was based upon an early trial developed using web services between the ER and SE PHOs to query each other's database. The APHO interoperability project was largely based on scaling this early example to a national level. There are two major problems with this approach:

1. The ER-SE demonstration uses the same code on both websites. Metadata was stored in a uniform format, and identical webservices were used on both sites. This presents an overly simplistic model for interoperability that ignores problems associated with different implementations. The solutions subsequently implemented by differing organisations demonstrated difficulties in creating semantically compatible solutions.

2. The live web-service demonstrated by ER-SE ignores the complexity of adding further nodes to the search. Initiating one web service is acceptable. Initiating eight web services is much more complicated as results from each partner organisation must be collated and ranked to some uniform criteria before the results can be displayed. A live solution would be unimaginable if the interoperable search were extended to all 302 PCTs.

If the project aim was solely to share metadata about resources between PHOs, a much more efficient solution would be to invest in the APHO website and implement a single database of all resources, hosted on a national site. PHOs could then create region-specific sub-sites. This solution would not require the use of web-services or XML feeds, and was not seriously considered – especially as many observatories had invested considerable resources into developing their existing websites and chose to protect the regional resources they were building up as opposed to supporting a national solution.

The lack of clear national governance was noted in chapter four, which describes how the APHO interoperability group chair was appointed in order to stop discussions getting too technical. As Sommerville [80] warned, the requirements, written at a broad overview level were too vague to sufficiently describe all the issues involved in implementing the project. When the interoperability group did meet, key issues such as semantic interoperability were raised, but were poorly understood, with the chair believing that semantic interoperability had been resolved by the decision to implement a cut-down version of e-GMS and to create the 'APHO metadata standard'. While this was a step in the right direction, further work was required and this issue is beginning to resurface as PHOs start to launch their new websites and start working with other organisations who have implemented different schemas. The lack of national project management has been evident throughout this project – chapter six described how SEPHO agreed that they would provide training for the National Public Health Language. A year later, and SEPHO are still working on their new website, and therefore have not provided any training. Consequently, this affected all the other projects dependant upon such training to ensure resources are correctly and consistently tagged. NEPHO found they lacked the training required to tag resources with the new ontology.

A question should be asked whether the primary goal was to develop "interoperability" or to secure more funds for individual web development. Many decisions were taken to duplicate effort within each region. APHO appears to be a group of rival organisations attempting to win a 'best website' contest, as opposed to focussing on their primary goal – that of providing the best possible set of health resources for each region. From a user's perspective, there are few benefits in having nine regional websites as opposed to one national website, which could subsequently host regional sub-sites. This region-based focus can be seen in much of the other work developed for PHO websites, leading to inequalities in the data available for each region. For example, a user can visit the NEPHO website and use an online mapping application to display health-related data by electoral ward – but only for the North East of England. A user can compare cancer rates between Newcastle and Middlesbrough, but not between Middlesbrough and York. This is the sort of tool that should be supported through interoperability – mapping tools could pull data from a variety of sources to aid mapping – whether that be the Yorkshire and Humber PHO, the Department of Health or the National Institute for Mental Health.

The national project funding provided by the Department of Health seems ill-conceived. The costs involved in implementing the ER-SE trial were simply multiplied by nine, and then the total money was distributed to each PHO based upon the size of the population based in the region. PHOs covering a small region substantially lost out. The apparent willingness to provide PHOs with money for this project before a national plan was agreed upon can be viewed as a central cause of the later decisions to develop differing solutions, each increasing the quardatic complexity of the final solution, the amount of time, money and effort required to deliver interoperable and e-GMS compliant websites to each region.

A likely step to complete the interoperability project is each PHO's integration with the metadata caching function developed by the PHORMS consortium. While a central metadata cache could have been developed on a national basis in conjunction with all observatories, the consortium took the strategic decision to abandon the goal of using live web-services. The cache may offer the centralised solution that the project needs – however it is unfortunate that this requirement could not have been developed nationally – incorporating feedback from observatories who are not members of the consortium.

8.5 Overview

Since large-scale computer applications were developed in the 1960s, interoperability has often been a goal of many system developers; however technology has often been blamed for making interoperability difficult by incompatible hardware and software environments. Major milestones towards developing interoperability were investigated in chapter two. This described the current research foci about web-services, XML and the vision of a 'semantic web'.

e-GMS is a generic framework that encourages local implementations to be developed. Consequently, the APHO interoperability group agreed a suggested set of metadata elements (termed 'APHOMS'). While this step was useful, the content of each element was left open to different interpretation. This led to observatories implementing semantically different versions of metadata elements – for example, one organisation has implemented the 'publisher' element as a text string, while NEPHO have implemented the element as the concatenation of a number of smaller, more specific sub-elements – such as publisher name, publisher address, publisher telephone number, publisher email address – and so on. The semantic mapping of one metadata scheme to another will take time to implement, and will usually result in the adoption of the most basic of the two metadata schemes. This is a time-consuming process that quickly becomes impossible to manage without the use of standards. The integration problem is quadratically complex – for N nodes, there are N × (N-1) interfaces required. The use of a common standard could make this problem easier to implement because complexity would remain linear.

Work still remains to be done in implementing interoperability between organisations. While interoperability in the past was often blocked because of technical inabilities, these problems have now been bridged by common standards such as XML-based web services. The excuse that it is not technically possible is no longer valid to hide political differences behind. The solutions developed by different PHOs demonstrate that even within a narrow industry segment, many different semantics can be derived of the same metadata.

8.6 Thesis Aims Revisited

Chapter 1.4 sets out four aims of this thesis. Each of these aims is evaluated in the following sections

8.6.1 Aim One: To investigate interoperability, metadata and standards developed for the UK's electronic government programme in suitability for use by NEPHO

Chapter Two descried technical interoperability – the development of high-speed computer networking, the birth of the internet, and the emergence of markup formats such as XML. The chapter went on to describe the concept of data caching, provided a detailed description of web services, and concluded with an overview of the semantic web. Chapter Three described electronic government initatives such as the e-GIF, the e-GMS and associated semantic issues such as the use of various controlled vocabularies. Subsequent chapters refer back to the research contained in chapters two and three while describing the work undertaken, thus accomplishing Aim One.

8.6.2 Aim Two: To analyse the existing NEPHO website and create an action list of tasks necessary to develop a site that would comply with requirements defined by the APHO interoperability group

Chapter Four described the formation of the public health observatory network, and the rationale behind the interoperability initiative. The chapter describes three levels of interoperability. A list of project requirements was drawn up in section 4.5. Subsequently, the existing NEPHO website was analysed, and a five-point action list of tasks was created which, when complete, would ensure the NEPHO website complied with APHO interoperability standards. Chapter Four therefore successfully completes Aim Two.

8.6.3 Aim Three: To design, implement and review the new interoperable NEPHO website.

Aim Three is successfully completed in Chapters Five and Seven. Chapter Five described the design and work undertaken to complete each action required to create the the new interoperable NEPHO website. As Chapter Seven evaluates, the new NEPHO website successfully complied with the APHO interoperability criteria. This did not mean the new NEPHO website was immediately able to interoperate with the other public health observatories, however.

8.6.4 Aim Four: To describe the approach taken by other PHOs to comply with the same requirements, to develop interoperability with these PHOs and to critically examine the successes and failures of the project at a national level.

The APHO interoperability project was not completed within the scope of this thesis. By June 2005, NEPHO had developed a website that complied with the APHO interoperability criteria, however other observatories had interpreted the requirements in a number of different ways. The approaches taken by each observatory are examined in Chapter Six. The implications of the approaches taken by other PHOs are examined in Chapter 7.7. Chapter 8.4 critically reviews the interoperability project, highlighting errors when the project was first envisaged such as scaling up a small prototype to a national model. The national project management is also criticised – there was little co-ordination between organisations, and the requirements fail to provide sufficient technical detail to sufficiently describe all the issues involved in implementing the project. Despite these concerns, the work described successfully completes Aim Four which is to describe and examine the project, not to deliver a final product.

8.7 Further Work

This section provides an overview of further work required to complete the interoperability project as envisaged by the APHO interoperability group. Work beyond this criteria is also described in section 8.7.2

8.7.1 Work required to complete interoperability project

The following work remains outstanding in order for NEPHO to complete the interoperability project:

- Complete metadata tagging for all existing resources
- Develop interoperable web service request / response for each of the following observatories:
  - North West
  - West Midlands
  - South West
  - PHORMS consortium (who will use a common interface):
    - YHPHO, EMPHO, ERPHO, SEPHO, LHO

Interoperability will ultimately be developed to include other organisations such as the new observatory functions in Wales, Scotland and Ireland – however such criteria go beyond the original interoperability brief, and are therefore beyond the project requirements. Such further work is described below.

8.7.2 Work beyond APHO interoperability criteria

The organisational characteristics of PHOs lend themselves to the early adoption of interoperability standards. There is a PHO in each English region, however there is very little national management to link each observatory. Interoperability was required to ensure that nationally significant reports published by one region could be found by visitors to another regional website. On a national scale, there are many organisations who collect, analyse and publish data about public health concerns. Typically, users have to visit many different websites to find resources about specific health concerns. The ultimate aim of the interoperability project will be to propose the way for the development of an interoperable public health knowledge base. For this to become a reality, common standards should be implemented, and the use of a virtual exchange or a single metadata cache should be explored.

Web services can do much more than simply search for resources. Part of a PHO's function is to provide health-related data. To date, much of this data has been in the form of published reports, stored as PDF files. As described in section 8.4, a number of PHOs have begun initiatives to develop online data analysis tools. NEPHO have developed online mapping tools that display health-related statistics in an interactive choropleth map (Figure 42).

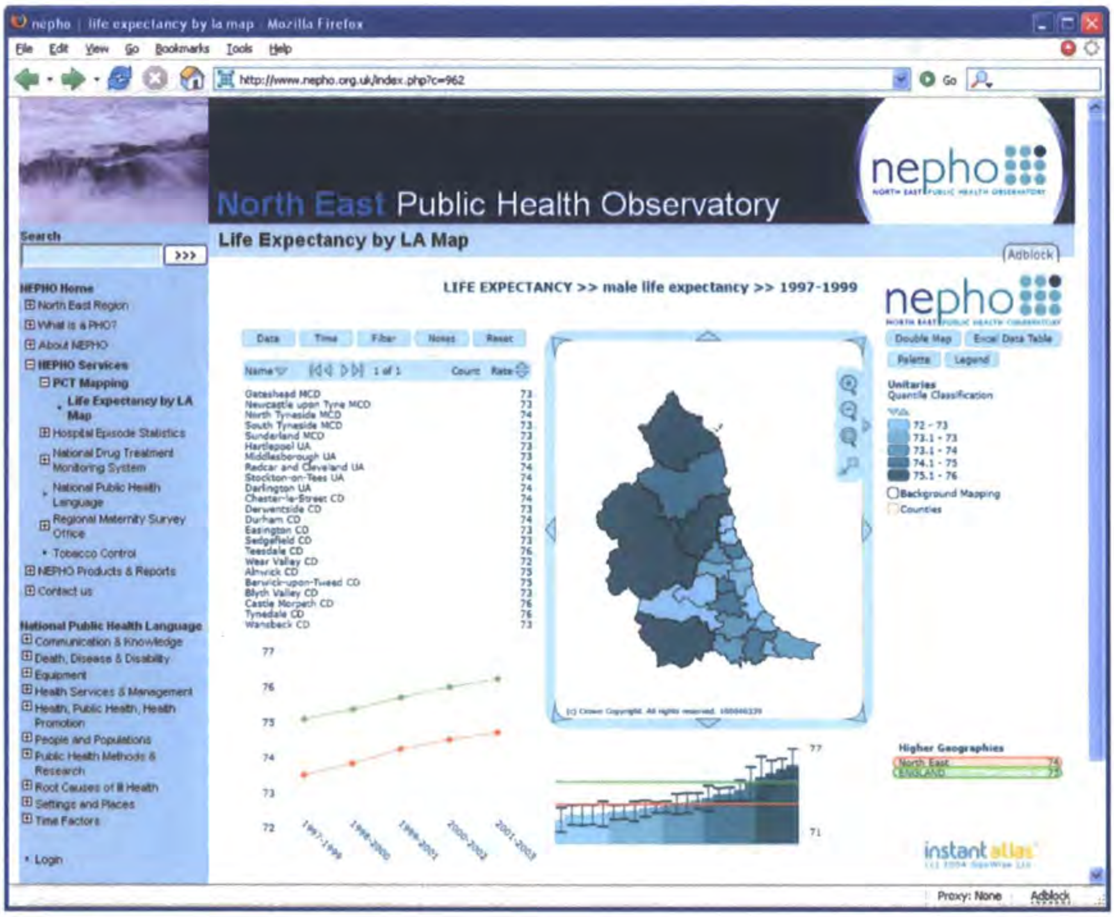


Figure 42 Online Choropleth Mapping

These new data analysis tools are not uniform across the country, leading to data inequalities between English regions. Future use of web-services and XML feeds could allow these interactive tools to include information stored by other organisations, such as the Office for National Statistics, National Institute for Mental Health, or observatories with a specific lead area. The knowledge to implement such a solution has been gained during this interoperability project. If issues of project planning and clear communication between observatories can be addressed, PHOs should be well-placed to continue developing interoperability between public health related organisations in the future.

Chapter 9 References

1. Department of Health – Saving lives: Our Healthier Nation – The Stationary Office, July 1999 – Last visited 21 August 2006 http://www.archive.official-documents.co.uk/document/cm43/4386/4386.htm
2. APHO Executive Group Meeting: 3rd March 2003 – Paper 7.2.2 – Report to PHO Directors from the APHO Interoperability Group.
3. Wolfe-Coote, V and Cotton, B - Achieving Interoperability through e-GIF – Findings and Recommendations for NICE, Oyster Partners Ltd – April 2005
4. Dietel H. M., Dietel, P. J., DuWaldt B. and Trees, L. K. - Web Services: A Technical Introduction – Prentice Hall, 2003
5. Object Management Group – CORBA – http://www.corba.org/
6. Sun – RMI – Last visited 21 August 2006 http://java.sun.com/docs/books/tutorial/rmi/
7. Horstmann, M & Kirtland, M. - DCOM Architecture – Microsoft Developer Network – July 1997 – Last visited 21 August 2006 http://msdn.microsoft.com/library/en-us/dndcom/html/msdn_dcomarch.asp
8. Microsoft – COM: Component Object Model Technologies – http://www.microsoft.com/com/default.mspx – Last visited 21 August 2006
9. Lau, C – Object Oriented Programming Using SOM and DSOM – John Wiley and Sons, March 1995
10. Wikipedia – CORBA – Last visited 11 August 2005 http://en.wikipedia.org/w/index.php?title=CORBA&oldid=20765430
11. National Center for Supercomputing Applications – NCSA Mosaic Homepage – Last visited 21 August 2006 http://archive.ncsa.uiuc.edu/SDG/Software/Mosaic/
12. National Center for Supercomputing Applications – NCSA Mosaic History – Last visited 21 August 2006 http://www.ncsa.uiuc.edu/News/MosaicHistory/
13. Gillies, J. and Cailliau, R. – How the Web was Born: The Story of the World Wide Web – Oxford University Press, 2000

14. NUA Internet Surveys – How Many Online – Europe – Last visited 11 August 2005
http://www.nua.ie/surveys/how_many_online/europe.html
15. Wikipedia – Cache - 04:41 11 August 2005 Revision – Last visited 11 August 2005
http://en.wikipedia.org/w/index.php?title=Cache&oldid=20750322
16. Harold, E. R. and Means, W. S. – XML in a Nutshell (Third Edition) – O'Reilly Media Inc.
17. Finkelstein, C and Aiken, P – Building Corporate Portals with XML – McGraw-Hill, 2000
18. Beznosov – Introduction to Web services and their security, Information Security technical report (2005) 10-14
19. Wahli, U., Tomlinson M., Zimmermann O., Deruyck W. and Hendriks D. – IBM Red book - Web Services Wizardry with WebSphere Studio Application Developer – Last visited 21 August 2006
http://publib-b.boulder.ibm.com/Redbooks.nsf/0/e8e74c4feadac98885256a560078c251?OpenDocument
20. Wang, H., Huang, J. Z., Qu, Y. and Xie, J. – Web services: Problems and future directions – Web Senamantics: Science, Services and Agents on the World Wide Web 1 (2004) p309-320
21. Vasudevan, V. – A Web Services Primer – O'Reilly Media Inc – Last visited 21 August 2006
http://www.xml.com/pub/a/ws/2001/04/04/webservices/index.html
22. Wikipedia – Semantic Web – 4th August 2005 Revision – Last visited 5 August 2005
http://en.wikipedia.org/w/index.php?title=Semantic_Web&oldid=20282967
23. Berners-Lee, T. Hendler, J. and Lassila, O. – The Semantic Web – Scientific American, May 2001.
24. Antoniou, G and van Harmelem, F – A Semantic Web Primer – Massachusetts Institute of Technology – 2004
25. Horst H. J. T. – Completeness, decidability and complexity of entailment for RDF Schema and a semantic extension involving the OWL vocabulary – Journal of Web Semantics, June 2005

26. Parliamentary Office of Science and Technology – Electronic Government: Technologies and the Citizen – February 1998 – Last visited 21 August 2006
http://www.parliament.uk/post/egov.htm
27. Irani Z, Love, P.E.D., Elliman, T, Jones S & Themistocleous, M – Evaluating e-Government: learning from the experiences of two UK local authorities – Information Systems Journal (2005) 16, 61-82
28. Chenery J – Seamlessly serving citizens. The business of public sector procurement – Summit, March 2001 – Last visited 21 August 2006
http://www.findarticles.com/p/articles/mi_qa3993/is_200103/ai_n8937213/print
29. Office of the Deputy Prime Minister - The National Strategy for local e-government - November 2002 – Last visited 21 August 2006
http://www.localegov.gov.uk/images/The%20National%20Strategy%20for%20local%20egov%20-%20Nov%2002_198.pdf
30. Office of the Deputy Prime Minister – Two Years On: realising the benefits from our investment in e-Government – March 2005 – Last visited 21 August 2006
http://www.localegov.gov.uk/images/2%20years%20on%20-%20Realising%20the%20benefits%20from%20our%20investment%20in%20e-gov_227.pdf
31. Horrocks, I & Hamberley, N – The 'webbing' of British local government – Public Money and Management, April-June 1998 – Last visited 21 August 2006
http://www.blackwell-synergy.com/doi/pdf/10.1111/1467-9302.00114
32. BCC E-Business Taskforce – BCC's Priorities for E-Commerce 2001 – British Chamber of Commerce, November 2001
33. Department of Trade and Industry – Our Competitive Future: Building the Knowledge Driven Economy – HMSO, 1998
34. Cabinet Office – Modernising Government – HMSO, March 1999 – Last visited 21 August 2006
http://www.archive.official-documents.co.uk/document/cm43/4310/4310.htm
35. Government to speed up introduction of online services – Downing Street Press Notice, 30 March 2000 – Last visited 21 August 2006
http://www.number-10.gov.uk/output/Page2795.asp

36. Cabinet Office – E.Gov: Electronic Government Services for the 21st Century – Last visited 21 August 2006 http://www.strategy.gov.uk/downloads/su/delivery/e-gov.pdf
37. IDA Frequently Asked Questions – Last visited 21 August 2006 http://europa.eu.int/idabc/servlets/Doc?id=17311
38. Decision 1719/1999/EC: Decision of the European Parliament and of the Council of 12 July 1999 on a series of guidelines, including the identification of projects of common interest, for trans-European networks for the electronic interchange of data between administrations (IDA) http://eur-lex.europa.eu/LexUriServ/site/en/oj/1999/l_203/l_20319990803en00010008.pdf – Last visited 25 September 2006
39. Decision 1720/1999/EC of the European Parliament and of the Council of 12 July 1999 adopting a series of actions and measures in order to ensure interoperability of and access to trans-European networks for the electronic interchange of data between administrations (IDA) – Last visited 25 September 2006 http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:1999:203:0009:0013:EN:PDF
40. European Commission - Electronic interchange of data between administrations: IDA programme – Last visited 25 September 2006 http://europa.eu.int/scadplus/leg/en/lvb/l24147a.htm
41. Cabinet Office e-Government Unit - e-Government Interoperability Framework – Version 6.1 – March 2005
42. Graham, S. et. al. – Building Web Services with Java – Making sense of XML, SOAP, WSDL and UDDI – Second Edition – Sams Publishing – 2004
43. Dublin Core Metadata Initiative - http://dublincore.org/ – Last visited 21 August 2006
44. NCSA – The Second International Web Conference '94 http://archive.ncsa.uiuc.edu/SDG/IT94/IT94Info-old.html – Last visited 21 August 2006
45. Dublin Core Metadata Initiative – History of the Dublin Core Metadata Initiative – http://dublincore.org/about/history/ – Last visited 21 August 2006

46. Wielel, S and Miller, E – An Introduction to Dublin Core – O'Reilly XML.com - October 2000 – Last visited 21 August 2006 http://www.xml.com/pub/a/2000/10/25/dublincore/
47. Cabinet Office E-Government Unit – *e-Government Metadata Standard – Version 3* – April 2004
48. Office of the e-Envoy – e-Government Metadata Standard Version 3.0 for websites – Last visited 21 August 2006 http://www.govtalk.gov.uk/schemasstandards/metadata_document.asp?docnum=806
49. Office of the e-Envoy – Electronic Government Metadata Framework – May 2001 – Last visited 21 August 2006 http://www.govtalk.gov.uk/schemasstandards/metadata_document.asp?docnum=860
50. Cumming, M – Cabinet Office – e-Government Unit – Tomatoes are not the only fruit: a rough guide to taxonomies, thesauri, ontologies and the like Version 2.1 Last updated April 2005 http://www.govtalk.gov.uk/schemasstandards/gcl_document.asp?docnum=681
51. Caselaw – U.S. Supreme Court – Nix v Hedden, 149 U.S. 304 (1893) – May 10, 1893
52. Cabinet Office – E-Government Unit – Implementing the GCL v2 – August 2004
53. Cabinet Office – E-Government Unit – IPSV/GCL Help http://www.govtalk.gov.uk/schemasstandards/gcl_help.asp – Last visited 21 August 2006
54. Dorst L., Pocetta, G. and Karlström, K. – European multilingual thesaurus on health promotion in 12 languages – NIGZ Netherlands Institute for Health Promotion and Disease Prevention – March 2001
55. Dorst L. – Restoring the Tower of Babel: building a multilingual thesaurus on Health Promotion – Health Libraries Review, Vol 15, Issue 1 – March 1998
56. United States National Library of Medicine – UMLS Metathesaurus Fact Sheet – November 2004 – Last visited 21 August 2006 http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html
57. Association of Public Health Observatories - Public Health Observatories in England: Progress and Prospects 2001/02

58. Association of Public Health Observatories - Public Health Observatories in England: Progress and Prospects 2000/01
59. Mindell, J. – Report on PHO Interoperability for the National Governing Board – 21 November 2003.
60. Crystal, A., School of Information and Library Science, University of North Carolina at Chapel Hill – Interface Design for Metadata Creation – Poster presented at CHI 2003: New Horizons.
61. phpBB – What is BBCode?
http://www.phpbb.com/phpBB/faq.php?mode=bbcode#0
Last Accessed 2 June 2005
62. Mozilla.org – Midas Specification – Last visited 29 August 2005
http://www.mozilla.org/editor/midas-spec.html
63. Microsoft Developer Network – Editable Regions (IE 5.5) – Last visited 29 August 2005
http://msdn.microsoft.com/archive/en-us/samples/internet/ie55/EditRegions/default.asp
64. Roth, K – Cross-Browser Rich Text Editor – Last visited 21 August 2006
http://www.kevinroth.com/rte/demo.htm
65. Pelzer, T – DBA Zine – Picture This: BLOBs – April 2005
http://www.dbazine.com/db2/db2-disarticles/pelzer1 – Last visited 21 August 2006
66. Atkinson, L – Storing files in a relational database – October 2001
http://www.zend.com/zend/trick/tricks-sept-2001.php?article=tricks-sept-2001&kind=tr&id=358&open=1&anc=0&view=1 – Last visited 24 August 2005
67. Waterson, K – An Introduction to BLOBs – Storing images in MySQL
http://phpro.org/mysql-image.php – Last visited 24 August 2005
68. NuSOAP – SOAP Toolkit for PHP – Last visited 29 August 2005
http://sourceforge.net/projects/nusoap/
69. Pilgrim, M – What is RSS? – O'Reilly XML.com – December 2002
http://www.xml.com/pub/a/2002/12/18/dive-into-xml.html – Last visited 26 August 2006
70. National Electronic Library for Health – What's New? RSS Feed
http://www.library.nhs.uk/RSS/xml/WhatsNew.aspx – Last visited 22 August 2006

71. National Electronic Library for Health – Document of the Week RSS Feed http://www.library.nhs.uk/RSS/xml/DocumentOfTheWeek.aspx – Last visited 22 August 2006
72. National Electronic Library for Health – Hitting the Headlines XML Feed - http://www.library.nhs.uk/RSS/xml/HittingTheHeadlines.aspx – Last visited 22 August 2006
73. National Electronic Library for Health – Hitting the Headlines Help - http://www.nelh.nhs.uk/hth/help.asp – Last visited 22 August 2006
74. Seidl, F. A. – Myst Technology Partners – RSS: Disruptive Technology Hiding in Plain Sight – September 2003
75. West Midlands PHO – Library of Grey Literature http://www.wmpho.org.uk/greylit/ – Last visited 22 August 2006
76. Mindell, J – Report to PHO directors of Interoperability Group – 6 July 2005
77. Microsoft – SharePoint Products and Technologies – http://www.microsoft.com/sharepoint/default.mspx – Last visited 22 August 2006
78. UK Atomic Energy Agency – Procurement Guidelines: Official Journal of the European Union (OJEU) (Last visited July 2005) http://www.ukaea.org.uk/contract/ojeu.htm – Last visited 22 August 2006
79. Health (Wales) Act 2003. Chapter 4. The Stationary Office. 2003. http://www.legislation.hmso.gov.uk – Last visited 22 August 2006
80. Sommerville, I – Software Engineering (6th Edition) – Addison Wesley Publishers Ltd - 2001
81. Web Ontology Language (OWL) – W3C Semantic Web Activity http://www.w3.org/2004/OWL/ – Last visited 22 August 2006
82. Department of Health – Choosing Health: Making Healthier Choices Easier – November 2004
83. National Library for Health – RSS Directory (In development) http://dev.library.nhs.uk/rss/directory/ – Last visited 12 October 2005
84. Office of the e-Envoy – Interoperability News – Last visited 6 October 2005 http://www.govtalk.gov.uk/schemasstandards/egif_document.asp?docnum=921
85. Wikipedia – *Internet* – 12 January 2007 Revision http://en.wikipedia.org/wiki/Internet Last visited 12 January 2007

86. Wikipedia – *Internet Protocol Suite* – 12 January 2007 Revision
http://en.wikipedia.org/wiki/Internet_protocol_suite
Last visited 12 January 2007
87. Sun Microsystems
http://www.sun.com/br/government_529/feature_uk.html
Last visited 20 January 2007
88. UK Cabinet Office – *e-Commerce@itsbest.uk: A Performance and Innovation Unit Report* - 1999
http://www.cabinetoffice.gov.uk/strategy/downloads/su/ecomm/ec_body.pdf
Last visited 20 January 2007
89. UK Office of the e-Envoy – *About Us*
http://www.e-envoy.gov.uk/about-oee/$file/aboutus.htm
Last visited 20 January 2007