Llywodraeth Cymru
Welsh Government

www.cymru.gov.uk

Digital Continuity: Record Classification and Retention on Shared Drives and Email Vaults

A 'proof of concept project' 2010-2011

Final report

ISBN 978 0 7504 6591 5

WG13492

PUBLISHED BY
Welsh Government
Cathays Park
Cardiff
CF10 3NQ

Acknowledgement
This report has been produced by
Dr S. Vidalis, Dr O. Angelopoulou and Mr L. Emanuel
Centre for Information Operations
University of Wales, Newport.

A special word of thanks to **Guidance Software** for the use of their eDiscovery Suite EnCase® and to Mr F. Wise from Guidance Software for his support and expertise.

ISBN 978 0 7504 6591 5

Information about this publication is available from:
The Publications Centre
Room 3.022
Welsh Government
Cathays Park
Cardiff
CF10 3NQ

Tel: 029 2082 3683
Fax: 029 2082 5239
E-mail: wag-en@mailuk.custhelp.com

Welsh Government Switchboard Service: 0845 010 3300
This document is available on the Welsh Government website: http://www.wales.gov.uk/

Executive summary

In 2007 the UK government identified several objectives for improving the storage of public sector information. In particular, and of direct relevance to this project, it wanted to:

improve the responsiveness to demands for public sector information
ensure the most appropriate supply of information for reuse
improve the supply of information for reuse
promote the innovative use of public sector information.

The aim of this project was to mine, categorise and classify information from a heterogeneous large-scale computer infrastructure and then store the search results in a forensically sound manner. Duplicate information was to identified for destruction and the process designed so that it could be implemented without disrupting staff operations.

The test data was a a 217Gb (810,000 files) sample taken from the Welsh Government (WG) shared drives and email vault. The records concerned largely related to the work of the Department of Education and Skills though 25% of the sample were taken from the wider organisation in order to ensure that the classification system used were useful over a broad range of subjects. The test data was stored in an isolated test environment with virtualised structures. All development work within the project occurred within the test environment.

De-duplication of the test data was achieved. Some **35.88%** of the files were identified as duplicates. Removing these files resulted in a saving of **29.49%** of physical space. After one pass of the data, it was possible to generate usable metadata for **75.7%** of the de-duplicated data set. This became the rich data set. The retention policies of the WG were used to design queries and rules for analysing the rich data set.

It was possible to extract **65%** of the files in the rich data-set for long-term retention together with their metadata in a format that would allow transfer to the WG Electronic Document and Record Management System (ERDMS Know as iShare within the WG). This translates to **55%** of the de-duplicated data set. Further analysis of the rich data set would have produced a better extraction rate. This would have been further facilitated by the use of knowledge extraction applications such as Pingar.

The data acquisition took **24 hours and 3 minutes** for 211.9GB. That is 150.371MB/min, which is within the lower range of the network performance based on performance tests. Projecting that to the whole infrastructure of WG, it is estimated that a straightforward data acquisition through the eDiscovery Suite would take **290.5** days. If it is possible to get maximum performance from TCP, then this estimate would fall to **60.6 days**. Of course, even this is not practical, hence it is recommended that in any follow on work operators **fragment the data set and parallelise the operation**.

The de-duplication process took **5 hours** for 211.9GB. Projecting to the whole of the WG infrastructure it is estimated that a full de-duplication would take approximately **60.4 days**. As this is not practical, it is recommended the process should follow the acquisition fragmentation, by de-duplicating the fragments and further parallelising this operation.

The indexing process took 5 days for 149.4GB. Using the **35.88%** duplication figure, some **39395GB** would need to be indexed. This would take an estimated **1318.451** days. With the suggested fragmentation of the data set and the parallelisation of the operation this time would be reduced.

There have been some technical (and non-technical) issues that affected the operations of the investigators.

> The virtualisation of the e-Discovery components was problematic, as virtualising within a virtual environment caused instabilities to the majority of the eDiscovery components.
>
> Legacy data types created in FAT32 systems do not hold rich metadata. This meant that the e-discovery process does not produce metadata to The National Archives standards. The retrieved metadata was not sufficient to answer all the classification queries. Interviews with WG personnel had to be performed in order to collect additional primary data about the current practice of classifying documents in WG.
>
> The lack of an isolated network and dedicated hardware resources greatly affected the performance of the eDiscovery Suite components. The acquisition, hashing and indexing operations were most affected.
>
> Towards the end of the project, there was insufficient memory to load the case and initiate the keyword searches to analyse the residual data further. It is imperative that high-spec computers with adequate processing power and memory capacity are used to host all the eDiscovery Suite components.

Despite the above problems the test data set was preserved in a forensically sound manner for the duration of the project. The hashing and indexing operations were conducted automatically and transparently by the eDiscovery Suite with minimal human intervention. There is an audit trail for all of the data manipulation activities through the e-Discovery Suite.

Given the problems encountered during the project, it is recommended that for future development work dedicated hardware resources (including networking resources), a secure 'classification environment' with root access to the whole of WG's ICT infrastructure and all human resources participating in the classification operations (network engineers, IT support personnel and investigators) should be based in this secure environment. The capability of easily isolating the classification environment from the rest of the infrastructure should be considered.

For efficiency purposes It is recommended that data is fragmented during the acquisition and hashing operations. The de-duplication operations will not be affected by this fragmentation. Several servers with appropriate computing power and memory capacity should be based in the classification environment. These servers should be used for running the software applications required for the analysis and classification of data as well as for temporarily storing the data under examination. After the successful classification of the data, the records will be exported to the predefined data repositories in the normal WG ICT infrastructure and their logical evidence files will be deleted from the classification environment. The servers should be connected to a number of computers running the Examiner modules. The Examiners can be virtualised so the host computers can run a number of virtual Examiners according to the requirements of the classification operations.

Regarding the applications used for the analysis of the data it is recommended that eDiscovery Suite from Guidance Software and the Pingar API from Pingar are used. Pingar API would allow some classification of the residual files that lacked appropriate metadata for categorisation.

Table of contents

List of tables

List of figures

Glossary and definitions

Case: a group of jobs.

Classification or **taxonomy**: Taxonomy is the practice and science of classification. A taxonomic scheme is a particular classification ("the taxonomy of ...") arranged in a hierarchical structure.

Condition: A single criterion or a combination of criteria combined using Boolean logic, to be applied to items coming from any data source, or a particular data source.

Criteria: a group of conditions, keywords, and matching files sets. A criteria set is applied during the collection or processing job.

Custodian: the owner of a given target.

Data: Individual observations, measurements and primitive messages from the **lowest level**. Human communication, text messages, electronic queries, or scientific instruments that sense phenomena are the major sources of data.

Digital Continuity: The ability to use your information in the way you need, for as long as you need.

Digital forensics: a branch of forensic science encompassing the recovery and investigation of material found in digital devices, often in relation to computer crime.

Digital media: a form of electronic media where data is stored in digital (as opposed to analogue) form.

ERDMS: Electronic Document and Record Management System.

File or **computer file:** A block of arbitrary information, or resource for storing information, which is available to a computer program and is usually based on some kind of durable storage. A file is durable in the sense that it remains available for programs to use after the current program has finished.

Indexing: Assigning a unique document identifier.

Information: Organised sets of data are referred to as information. The organisational process may include sorting, classifying, or indexing and linking data to place data elements in relational context for subsequent searching and analysis.

iShare: The internal name of the Welsh Government Electronic Document and Record Management System. (The commercial product is known as Objective).

Job: a group of targets, along with the criteria that will be used to scan them, and the output locations where responsive files will be stored.

Record: Information that has been filed.

Risk: Risk is the probability that a threat agent (cause) will exploit a system vulnerability (weakness) and thereby create an effect detrimental to the system.

Target: A location on a data source.

1.0 Project scope

1.1 Introduction

This report is submitted in fulfilment of Deliverable 1.2 (see Appendix A) of the Welsh Government Digital Continuity project. The project was commissioned from the Centre for Information Operations (CIO) of the University of Wales, Newport (UWN) by the Knowledge and Information Management Division (KIMD) of the Welsh Government (WG). Dr Vidalis is a senior lecturer at the University of Wales, Newport and Head of the Centre for Information Operations. Dr Vidalis was the lead investigator. Two other investigators participated in the project: Dr Olga Angelopoulou and Mr Les Emanuel.

The report contains the findings of the investigators of a study into the classification of unstructured electronic records using e-discovery techniques. The investigators used the eDiscovery Suite developed by Guidance Software to classify and structure the records identified in the data set provided to the CIO by the KIMD for the purposes of this project.

EnCase and eDiscovery are registered trademarks, and Guidance Software and EnScript are trademarks of Guidance Software, Inc.

1.2 Scope

The Welsh Government (WG) faces the ever-growing challenge of managing the risks associated with the storage of digital information if it is to ensure that the information it holds remains accessible over time. This requires the organisation to establish systems for the comprehensive capture and management of digital records. These systems should be designed so that WG records are available to support business activities as and when they are required. A failure to secure and provide proper access to digital information could result in the Welsh Government being unable to support the work of its administration and its staff or meeting the information requirements of the public it serves.

In 2007 the UK government identified several objectives for improving the storage of public sector information. In particular, and of direct relevance to this project, it wanted to:

improve the responsiveness to demands for public sector information
ensure the most appropriate supply of information for reuse
improve the supply of information for reuse
promote the innovative use of public sector information.

The primary risks which need to be addressed through a digital continuity strategy are machine dependency, technological obsolescence and the fragility of carrier media. There are also some emerging risks associated with record management in a digital environment. These new risks mainly arise from UK legislation and record management guidance. They concern the effective management of third party records created in the course of collaborative working, the effective management of the transfer of digital records into or between government systems as part of changes in the machinery of government, and possible penalties under the enforcement mechanisms available to the Information Commissioner and others.

1.3 Project aims and success criteria

The project examined the means of managing digital records in the WG legacy digital stores (shared drives and e-mail vaults). The goal was to ensure that high-value material can be migrated to secure long-term storage systems and that unwanted material can be identified and destroyed in a manner which conforms to UK record management guidance. The project was part of a wider digital continuity project within the WG.

This was a proof-of-concept project. The aim was to test whether the software could meet the following 15 tests for functionality and usability.

1. De-duplicate, classify and automatically extract digital records from WG's digital stores.

2. Generate metadata to The National Archives standards.

3. Allow record service staff to assign retention and disposal periods for groups of records with similar content (as classified by the software).

4. Extract records scheduled for disposal and generate disposal lists.

5. Extract digital records scheduled for long-term retention together with their metadata in a format which would allow transfer to the WG Electronic Document and Record Management System (iShare).

6. Identify groups of information (such as Education and skills » Post 16 » Further Education Student Financial Support » Individual Learning Accounts Wales (ILA)).
7. Identify types of data.
8. Identify personal data.
9. Identify file extensions and other metadata.
10. Interrogate legacy file formats.
11. Work within the constraints of the Government Secure Intranet (GSI).
12. Be compatible with Welsh Government file systems and storage methodologies.
13. Be compatible with Welsh Government technical infrastructure where required.
14. Be able to process the data provided in a timely fashion that would scale to the whole estate.
15. Prove to be safe, secure and reliable, requiring minimal human intervention.

1.4 Summary of achievements

1. The software successfully de-duplicated the test data; the results can be seen in **Section 5.1**. The test data was classified based on the methodology described in **Section 2**. The results of the classification can be seen in **Section 5.5.** The identified digital records were automatically extracted from the testbed into a predefined directory structure. The results of the extraction can be seen in **Section 5.6**.

2. Usable Metadata was generated for the majority of the identified digital records. This is described in **Section 2** and the results can be seen in **Section 5.2.**

3. Retention policies were used for designing queries and rules that were then used to analyse the test data. The rules can be seen in **Section 2** and a statistical summary of the results can be seen in **Section 5**. The actual results are contained within the logical evidence files generated by the eDiscovery Suite. These files currently reside on the testbed that was used for the project experiments.

4. Records were extracted and scheduled for disposal based on the retention policies. A sample of the software-generated reports is provided in **Appendix E**. The full audit trail is contained within the logical evidence files generated by the eDiscovery Suite. These files currently reside on the testbed that was used for the project experiments.

5. Digital records together with their metadata were extracted for long-term retention in a format that would allow transfer to the WG Electronic Document and Record Management System (iShare). The operation is described in **Section 2** and the results can be seen in **Section 5.6**. This operation used keywords and keyword lists based on input from KIMD employees. It did not use pattern-based analysis.

6. The software could identify groups of information. The methodology followed is described in **Section 2** and the actual results can be seen in **Section 5.5**.

7. The software could identify the types of data contained within the test data set. The results can be seen in **Section 5.3.**

8. The software could identify personal data. The results can be seen in **Section 5.4**.

9. The software could identify file extensions and other metadata. The results can be seen in **Section 5.2.**

10. The test data set contained legacy file formats.

11. The software worked within the constraints of the WG security protocolsI.

12. The approach was compatible with Welsh Government file systems and storage methodologies.

13. The approach was compatible with Welsh Government technical infrastructure where required.

14. Data regarding the length of time of the operations and projections for the whole estate based on these figures can be seen in **Section 5.7**.

15. The digital records were preserved in a forensically sound manner for the duration of the project. The operations were conducted automatically by the software application with minimal human intervention.

1.5 Report structure

Section 2 describes the methodology used for acquiring and analysing the test data set. It sets out in detail the different phases of the analysis and the data manipulation operations that were conducted.

Section 3 discusses the technical architecture of the software application that was used for conducting the operations specified in Section 2, and for analysing and classifying the test data set.

Section 4 describes the testbed that was developed for running and, more importantly, containing the experiments and the operations specified in Section 2.

Section 5 presents the results of these experiments. It sets out the results of the operations that were conducted during the project in term of actual figures and statistics, and it details the projections made from those figures for operations on the whole digital record estate.

Section 6 discusses the quality checks that we were conducted on the provided data set and on the experiments and operations conducted during the project.

Section 7 discusses the problems that were encountered during the project. It lists the main issues and makes recommendations for future work.

Section 8 contains a series of appendixes in support of the results presented in Section 5.

2.0 Classification methodology

2.1 Classification requirements

Digital forensics is a generic term that covers all aspects of the examination and recovery of material that resides on digital devices. It is often associated with the investigation of computer crime, dealing with situations ranging from industrial espionage to damage assessment. However, digital forensics can be applied in any computer-based environment that requires the collection and analysis of data. It can be described as a specialised approach to data manipulation that allows the content of digital files to be examined in a forensic manner. The data is preserved and the actual content remains unaltered during this examination.

The evidence used in this examination is any kind of digitally processed information that is stored on any sort of digital media. Residual data on digital media can recover the digital trail of the media. It provides valuable information about the history of the system. Recovered data can also enhance the investigation process.

The application of digital forensics in this project played a key role. The analytical forensic approach enabled a more straightforward and speedier analysis of the data. It provided accurate and reliable results. As a starting point for this analysis, it was necessary to organise the electronic files contained in the testbed. The following sub-sections set out the methodology use to structure the files.

2.1.1 Digital continuity attributes

Several attributes influenced the design of the classification framework. First, it was necessary to consider the features of the existing system.

a) The WG has an existing classification system. An initial analysis of the testbed showed that files have been categorised, often inappropriately, under a directory structure. This is a weak approach for structuring data.

b) The testbed contained files that were created in obsolete file management systems, which were designed prior to the current The National Archives standard for metadata.

The possible interrelationships between the files are important consideration in designing the classification system. The system should be built in a way that makes it possible to identify, resolve and successfully manage these interrelationships. The initial research suggested that the interrelationships can best be managed by linking a file with the department from which it originated. Therefore, it was decided to classify files by the department and the author that initially created the file. This means that each file is directly linked with its source.

2.1.2 Design principles

A specific methodology is required in order to achieve a comprehensive and repeatable procedure. For the Digital Continuity project, this procedure should be able to identify and classify the WG's electronic files and to manage their interrelationships.

The methodology adopted for this project is divided into phases. Every phase represents a major set of procedures. In essence, a phase comprises procedures that are related with each other. When these procedures are completed, the outcome of the phase is created. The examination then can continue to the next phase. The naming of each phase is specific to the Digital Continuity project. The terminology used for defining each phase also reflects the proposed methodology.

The methodology is designed in accordance with the attributes and capabilities of the eDiscovery software.

2.1.2.1 Specific terminology

The **phases** are individual procedural components inside the methodology. Their names describe their purpose in the terms of the Digital Continuity project. A phase consists of several **processes** that enable a structured approach to the satisfaction of the input/output (I/O). However, on a lower level, the processes comprise **activities** – customised, focused classification guidelines that clarify the required actions set by the processes.

Every phase requires some form of **input**. This is then modified or examined by the processes present within that particular phase. The processes satisfy the needs that their preceding input or output processes require. In turn, the processes are built up from activities that satisfy the purpose of the corresponding process. Figure 1 describes the breakdown of the elements that constitute each phase.

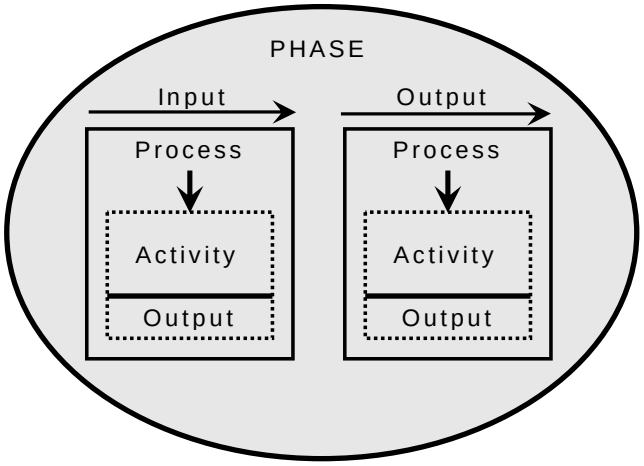


Figure 1: Representation of the methodology's elements

2.2 Methodology

This section describes classification methodology used in this project. It includes a description of each phase of the methodology, a figure of the required input and the produced output, as well as a graphical representation of the phase.

Figure 2 represents the initial contact with the structure of the WG file system. It illustrates the procedures that needed to be adopted for the project and how these could be translated into a systematic approach towards the file system. It shows the project requirements, from the collection of the test data to the evaluation of the classification results. These requirements informed the design of the methodology that is presented in this section of the report.

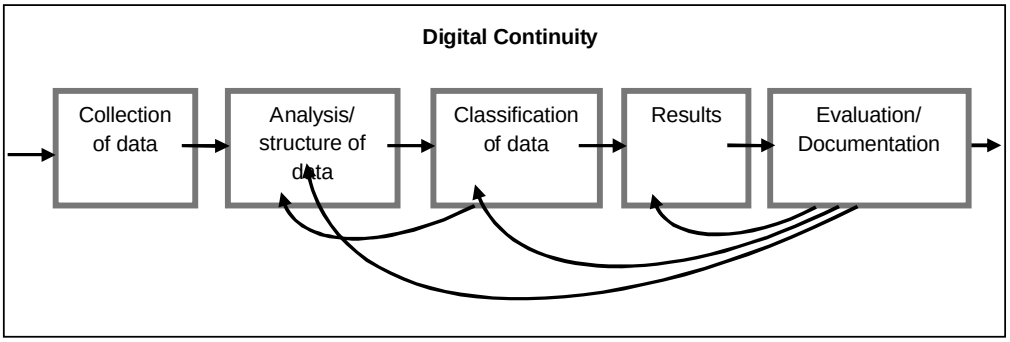


Figure 2: Classification methodology

2.2.1 High-level overview

It was decided that three phases could satisfy the objectives of the project. Each phase represents a specific and independent set of procedures. The interaction between the different phases is similar to that used by the waterfall model in software engineering – the approach is to proceed sequentially from one phase to the next, moving to the next phase only when the preceding phase is completed. However, in order to satisfy the needs of this project, the model needs to be rerun continuously. This should continue until there is no, or a minimum amount of, residual data. The model should also constantly check for any population of new data in order to determine any new interrelationships. The methodology needed several levels in each phase in order to achieve the required results.

Figure 3 shows the three phases used in the methodology. The preparation phase aims to identify the available data to be classified. The metadata analysis phase (or content phase) performs an initial sorting of the available files. The data extraction phase (or work phase) provides an analysis of the current obtainable data.

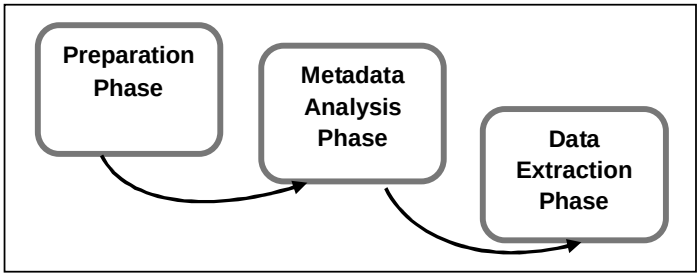


Figure 3: Digital Continuity ERMS methodology phases

Every phase requires an input and produces an output. The inputs and the outputs define the processing requirements of each phase and are integrated in the high-level framework. Their existence is imperative as they constitute the purpose of the phase and define the object and the subject of the examination.

The input is the object that needs to be entered into the phase. It is examined for a particular set of features, and the results form the output of the phase. This output is the subject of the phase, and it is also examined in order to provide further results. The importance of the output is verified in the subsequent phase, where it will be used as an input.

2.2.2 Methodology phases

2.2.2.1 Phase 1: Preparation

This phase allows for the continuous population of new data. An EDRMS (electronic Doucment and Records Management system) addresses the need for constantly classifying new data. For this reason, in a live system, the preparation phase is run continuously. It collects recently added data and prepares the system for the metadata analysis (or content) phase. In a live system, therefore, the preparation phase involves continuously acquiring data from active servers.

The preparation phase engages with the available files that need to be classified. Every records management system is populated with different file types. In this phase, the aim is to collect all electronic files, classified by their type (such as text documents, spreadsheets etc.), and create a controlled system of files that can be analysed at the content stage. This corresponds to the acquisition phase of the available digital media in a traditional digital forensics procedure.

The data set used in the pilot was a 217Gb (810,000 files) sample taken from the Welsh Government shared drives and email vault. The records concerned largely related to the work of the Department of Education and Skills, though 25% of the sample were taken from the wider organisation in order to ensure that the classification system used were useful over a broad range of subjects. The sample included records from the period 1997 to 2010.

. The contents of this data set was the system that was analysed and was used as an input to the next phase. The data set included e-mail communications that were linked with specific issues and with files. This necessitated working out the interrelationships of the files to eamilas. Any e-mail communication that is related with certain assignments as well as all data types that contain metadata was be included in this analysis.

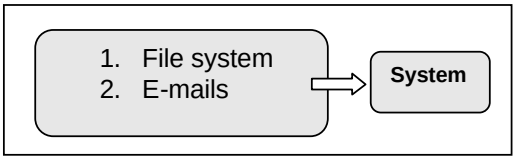


Figure 4. Preparation phase

All files were concentrated in one compilation of data. They were then sorted, based on the metadata they contain. Table 1 shows the elements that should be included in

an EDRMS as recommended by the record management metadata standard of The National Archives.

Deleted: Page Break

#	Metadata Element
1.	Identifier
2.	Title
3.	Subject
4.	Description
5.	Creator
6.	Date
7.	Addressee
8.	Record type
9.	Relation
10.	Aggregation
11.	Language
12.	Location
13.	Rights
14.	Disposal
15.	Digital signature
16.	Preservation
17.	Mandate
18.	Format
19.	Function
20.	Coverage

Table 1: Metadata elements

Each element builds up a 'record' of the file, and details its special unique characteristics. Both the file system and the e-mail communications should have contained most of the elements listed in Table 1. These will assist with the classification of the system.

During this phase, an initial assessment of the properties of the files is made in order to identify the actual content of the metadata they contain. For this purpose, it is sufficient to assess a random sample of files rather than the vast amount of acquired files. The aim here is to get a general overview of the metadata. There is a possibility that some files may not contain appropriate metadata. This issue will be resolved in Phase 2, where the system will be analysed and classified based on its metadata content.

For this pilot project, the supplied data is of limited volume. However, the initial assessment of this data should produce an indication of the existing metadata record keeping for the whole ERMS.

Deleted: ·····Page Break·····

Figure 5 illustrates the input/output for the preparation phase. The input are the available files supplied by the WG. The aiming of the preparation phase is to produce a collection of files that can be further analysed for the purposes of this project. This is the output or the system. This system will be further treated in Phases 2 and 3.

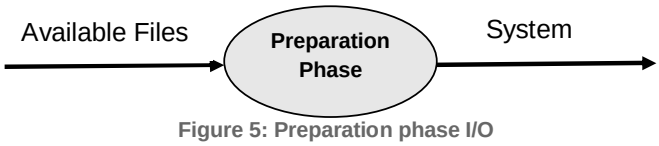


Figure 5: Preparation phase I/O

2.2.2.2 Phase 2: Metadata analysis

Phase 1 constructed the system from supplied data that will be analysed in this phase. It consists of a collection of files that should be possibly saved and archived based on The National Archives standards and in accordance with the WG's filing system. Phase 2 examines and analyses the available metadata in order to achieve the classification of the system. The purpose of this phase is to provide a basis for transferring the unstructured data into a file management system. This procedure will be initialised in Phase 2, where the files will be identified, and finalised in Phase 3 where they will be analysed and classified.

The system (the output from phase 1) contains all the available data acquired from the digital media. Some files will hold the required metadata in order to identify their content and purpose. However, other files may not include all the necessary metadata elements. In such cases, additional analysis may be required to provide further information before these files can be classified. This process should reduce amount of residual data – for which there is insufficient metadata – to a minimum.

Not all the metadata elements presented in Table 1 need to be present for this task. However, some are considered essential for classifying a file for the purposes of the Digital Continuity project. These are the Identifier, Title, Subject, Description, Creator, Date, Addressee, Disposal and Relation. These elements provide a description of the actual purpose of a file and assist the classification process by identifying the main values contained in the file.

In the initial manipulation of the files in Phase 1, it was noticed that the vast majority of files do not have a metadata structure based on The National Archives standards. The acquired metadata tends to have this structure.

1. Title
2. Company name, which usually refers to the department that the file is created in
3. Creator, which refers to the employee that created the file and is assigned to a specific department
4. Date

To extract metadata in an appropriate format and structure for classifying the system required specific manipulation of the data in the testbed. This was achieved by running custom EnCase scripts written in EnScript.

For the analysis of the system in Phase 2, two processes are required. Process 1 involves originals detection and Process 2 involves metadata examination.

Process 1: Originals detection

In order to proceed with the analysis and classification of the system, the data was cleansed by excluding any file duplicates. To do this n examination of the unique identifiers for the files was required. The duplicates werel be isolated from the system and stored in a separate archive folder. This allowed the actual number of files needing classification to be identified. It provided a system that only contains unique files.

The de-duplication process was based on a compare-by-hash analysis. Calculating a hash value (or hashing) is the process of taking an arbitrary length of data and calculating a value (the hash value) that can uniquely identify that the data has not been altered or changed in anyway. The compare-by-hash technique is used to discover identical blocks of files. By comparing the hash values for two inputs, it is possible to determine whether the inputs are definitely not the same, or that they are the same.

Hashing can be used to check that data has not been altered in any way from when the hash value was first calculated. If the data is modified, even by changing just one bit, then the hash value will change because of the routine employed by the hashing algorithm.

EnCase automatically creates hash values for all files contained in a data set. Therefore, after the acquisition of data in Phase 1, a hash table can be created listing the hash values of all files in the system. Multiple files with the same hash value are duplicates, and an original file can be retained with the others isolated from the system.

The output from this process is therefore a set of unique files. These form the testbed that is used for the classification of the system.

Process 2: Metadata analysis

Phase 2 then continued with the processing of the remaining unique files. These needed to be examined and classified based on their metadata.

The different data types in the testbed are also identified at this stage. The data types saved in the shared drives of the WG not only consist of files that contain metadata, but also system files, old and unusable file extensions and other file types that do not hold metadata. These were be highlighted and flagged according to the WG's retention policy.

The WG's corporate disposal schedule was applied in the metadata analysis phase. It set the parameters for isolating those files that have fulfilled their purpose and can be disposed of as they are no longer needed by the organisation. There are different recommended retention periods for digital records depending on their content. The current system used by the WG sets retention periods related to the date of the creation of the file. Some of the testbed data set contains records for which the content is unknown, so the WG's disposal schedule (which is based on the content of a file) cannot be applied. In this case, a general 10-year rule has been applied: these files are migrated from the system 10 years after their creation date.

Appendix D contains a flowchart diagram that shows part of the disposal schedule. It demonstrates a study that was undertaken in order to systemise record keeping practices.

The categorisation agreed with the WG attempted to sort the data among the WG's different divisions. The specified department was be the top level of the classification, followed by the division. In order to achieve this breakdown, the classification software needed to access the metadata that incorporates the required information. This was achieved by conducting metadata searches in the files.

There was a dual purpose to this process. First, it identifies those files in the system that contain the relevant metadata that allows them to be sorted. These can be then be classified. Second, its creates a set of files that do not hold enough relevant metadata. These require further analysis, which is the final stage of Process 2.

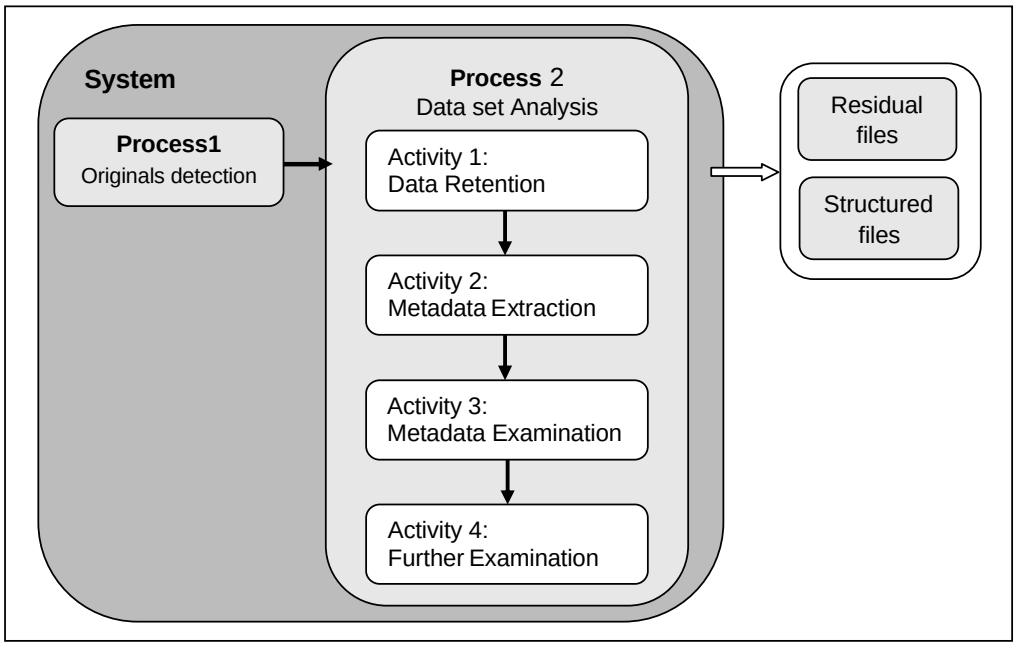


Figure 6: Metadata analysis phase

Figure 6 provides an overview of this process. The files were the subject of a metadata analysis, and they were sorted based on the current departmental structure. However, not all files can be sorted using the existing available information. There will be a residual set of unsorted files that will need to be examined further in the next phase.

The specific activities are customised to the needs of the current system. If the initial assessment of the available metadata obtained from looking at a sample of the testbed data had produced different results, then the approach and analysis would have been adjusted accordingly.

Activity 1: Data retention schedule

As explained above, the WG corporate disposal schedule cannot be applied to all files in the testbed because of incomplete metadata. However, by applying a general 10-year rule, some files with incomplete metadata can be migrated. An EnCase

query was written and run across the acquired data in order to isolate those files that are older than 10 years. The highlighted files are separated from the testbed and do not need any further analysis.

Activity 2: Metadata extraction

The remaining files after Activity 1 are then subject to metadata extraction. The extraction of the metadata is not a straightforward procedure that can be performed by the eDiscovery suite. An EnCase script in EnScript was written and run at the initialisation of this process. This extracts all available metadata from the testbed files. The collected metadata are then used in Activity 3.

Activity 3: Metadata examination

The examination of the metadata is an important part of Process 2. This analysis is carried out using the ECC Web Server (see Section 3). This allows extensive metadata analysis.

The analysis is broken down into three parts to allow all useful information to be used to inform the classification process. It takes into account the structure of the metadata in the current system. As described above, this is organised as follows.

1. Title
2. Company name
3. Creator
4. Date

In the majority of the files, the content of the title field appeared to be inadequate for classification purposes. The titles tended to be relevant to the content, but they do not provide sufficient information that could link the file with a specific department or give other details about the content that could assist the classification. It was therefore considered to be the least valuable item of the available metadata.

Much more valuable is the company name metadata. The vast majority of the files contained metadata related to the department in which the files were created. This information is descriptive and relevant, as it directly provides the required department name. It satisfies the needs of the classification. It was therefore decided to use the company name metadata field as the primary classification source.

However, there are instances where the company name field contained a description or title that does not match any of the current or past departments or divisions of the

WG. There are some other instances where the files have an outsource origin, such as a local council. The classification hierarchy accommodates these files according to their origin.

The files that provide metadata information that can be linked with the WG are then examined further. The creator, or author's name, is used as the secondary classification source. The author is the employee who created a specific file. Therefore, the author can be linked with a specific department. As with the company name metadata, there can be some issues with this data. There may be author names do not match any names on the WG staff list. These will need to be further examined or considered as residual files.

The employees of the WG tend to move between different departments over the course of their careers. In order to achieve a reliable classification of the files by creator, the date that the file was created will need to be taken into account. This is because the author may have subsequently moved departments, so the name needs to be matched against the WG staff list at the time the file was created. The date is the tertiary classification source. However, this raises a problem because the staff lists provided by the WG only cover the period 2006–2010. Given that the testbed contains some files dating back to 1998, there is an eight year period that cannot be covered. In order to a bypass this issue, the files from this period that contain an author's name will be classified by the name of the author.

The files that do not contain any of this metadata form a set of residual files. These are further examined in Activity 4.

Activity 4: Further examination

The analysis of the company, author and date metadata produced a data set of grouped files. It also produced a set of files that could not be classified because there was inadequate metadata. These were further examined to see if they could be managed and grouped in order to minimise the number of residual files. These remaining files are examined by their file extension (or file type) and their content.

First, the remaining files need to be sorted by file type. Some files may belong to file types that do not contain metadata needed for classification, such as system files, or may belong to a file type that cannot be classified under any structure, such as sound files, or may have an unusable file extension due to their age.

Second, it may be possible to apply retention schedules based on the file type and on anticipated future usage. For example, some image files may be usable in the future, while others may not be required to be retained for future reference. This requires a content analysis. This is a manual procedure that requires some input from the creators of the files for verification. Therefore, these are considered as residual files of the system. Further manual analysis of these files can occur after Phase 3.

Figure 7 illustrates the input/output for the metadata analysis phase. The input for this phase is the system that was the output of Phase 1. The output of this phase will be two sets of files: the files that can be structured and classified based on their metadata; the residual files that do not contain enough metadata to be classified, but which are sorted by file type.

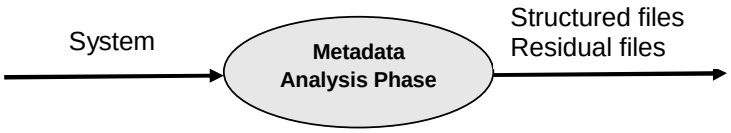


Figure 7: Metadata analysis phase I/O

2.2.2.3 Phase 3: Data extraction

Data extraction is the last phase of the data classification system. The two outputs of the previous phase will be used: the structured files as resulted from Activity 3 and the residual files as resulted from Activity 4.

Structured files

The files that have been classified in the previous phase need to have their relationships within the EDRMS verified. This was achieved firstly by the three-stage classification – by department < author < date – and secondly by the mutual characteristics of a set of files.

Residual files

This set of files has been organised by file type in Phase 2. The content of these residual files needed to be further examined in order to determine their relationship with the structured data.

Further analysis on the residual files can be undertaken by using search strings to uncover any personal data contained among the files. This analysis focuses on:

a. e-mail addresses – if governmental and corporate domain names are excluded, then all other e-mail addresses found by a search could be personal communication; a webmail script will also reveal any webmail accounts

b. credit card numbers – these could perhaps identify online purchases by users

c. a keyword list search across saved web pages – this list should include the names of retail shops, car companies, football teams etc.

This process can identify files that users could have accessed and stored for personal use. Of course, some e-mails may contain a mix of personal messages and business-related information. However, these e-mails are still going to be regarded as business related and should be treated accordingly.

Hash analysis

Before the file extraction is performed in Phase 3, an additional hash analysis should be undertaken in order to validate the results and ensure the existence of unique files. The categorisation of files during the metadata analysis could have created new duplicates. At this stage, the results need to be verified and evaluated.

Storing the results

Before the file extraction process is begun, a hierarchical folder structure is created to accommodate the classification results. This structure is designed to reflect the specific WG departmental structure and the features of the supplied system. The aim is to match all files contained in the testbed to a specific category, thereby grouping files with common characteristics.

The structure contains all the departments and divisions of the WG, according to the 2010 organisational structure, as well as the main types of external sources that engage with the WG in the course of its work.

Hierarchy Folder Structure	
1.	First Minister
2.	Economy and transport
3.	Counsrl General and Leader of the house
4.	Social justice and Local government
5.	Finance Public Services and delivery
6.	Health and Social Services
7.	Environment, sustainabilty and Housing
8.	Children Education And Lifelong Learning and Skills
9.	Heritage
10.	Rural Affairs
11.	AUTHORS
12.	COUNCIL
13.	CONSULTANT
14.	PERSONAL
15.	FILE TYPES

Table 2: Hierarchy folder structure

Table 2 shows the top level of the hierarchy folder structure. For each department (folders 1–10) there are subfolders that refer to the different divisions. Files that contain enough metadata to link them with a department but not with a specific division are extracted and placed in the top level folder.

For some files, the specific originating department cannot be identified because the metadata is insufficient. Sometimes the author's name is known but it cannot be linked to specific department due to the limitations of the staff lists. These files are listed under the author's name in folder 11. The folder contains subfolders for each author.

Councils (folder 12) contains those files for which the metadata indicates that they originated from a local authority. They are extracted and filed by the name of the council. Consultant (folder 13) holds files that originate from a consultancy firm or other supplier that provides services to the WG. They are extracted and filed by the name of the firm. Personal (folder 14) contains those files in the shared drives that are identified as containing employees' personal communications. File types (folder 15) contains the remaining files for which no information could be extracted from their metadata that would allow classification. These are extracted and filed by their file type for further analysis. Any remaining files should be placed in this folder.

In Phase 2, the metadata was analysed in the ECC Web Server. The ECC Desktop was used to tag a group of files and create logical sets of files from the data. In this phase, the logical files are extracted and filed in the hierarchical structure using EnCase. An evaluation of the created system was first undertaken to verify the authenticity and integrity of the results. This is achieved by digital forensic signature

analysis. Signature analysis is a process of comparing files, and their headers and extensions, with a known database of file headers and extensions in an attempt to verify all files on the storage media and discover any which may be hidden. EnCase can automatically verify the signature of every file it searches. It therefore verifies the source of each file or identifies any may mismatches.

Figure 8 shows the input/output of the data extraction phase. The input of this phase are the residual and the structured files from the metadata analysis phase. The output is a classified system, with a main emphasis on the WG's departmental structure. There is a complementary classification to take into account other types of files that have insufficient information or don't fit within the departmental structure.

Figure 8: Data extraction phase I/O

The product of this stage was a classified file system that reflects the departmental structure of the WG. The files are stored by the appropriate division within each department.

3.0 Technical specifications

3.1 Application overview

The software application used for implementing the methodology described in Section 2 was the EnCase Command Centre (ECC) and its accompanying toolkit.

This section describes this software. It details how resources are assigned to each system component, and specifies the resource requirements. It also includes recommendations for optimising the performance, integration and usage of these components. It outlines the user interface for the software.

3.2 Technical architecture

The software that is primarily utilised in this project is the EnCase Command Centre (ECC). This is the interface that supports the EnCase eDiscovery and EnCase CyberSecurity applications. The ECC consists of the several components.

ECC Desktop – this component allows users to define cases, set up jobs, and define values for sources, custodians and targets.

ECC Examiner – this performs the data collection and processing activities. These activities are resource intensive, so Guidance Software recommends deploying ECC Examiner on a server grade computer.

ECC Database Server – a database server, such as SQL Server or MySQL, is required to create and administer the global and case databases. The ECC Desktop and an ECC Examiner must communicate with the database server for full integration. If the project requires the ECC Web component, the global database must communicate with the ECC Web Server as well as with the ECC Desktop.

Output File Storage – for five jobs running concurrently, the minimum amount of storage for temporary files is approximately 25GB. However, Guidance Software recommends allocating 50GB of output file storage to avoid any possible issues with memory allocation.

ECC Web Server – the ECC Web Server must communicate with the global database, but it is not required to connect to the case databases, ECC Desktop or an ECC Examiner.

Table 3 shows the hardware, software, and system requirements needed to install and configure the ECC, the ECC Examiner, the third-party databases, the ECC Web Server and the ECC Web Client. The recommended and preferred resource values and configurations are noted where applicable.

EnCase Command Centre and EnCaseCyberSecurity	
Class	Desktop or server class hardware (64-bit)
Operating Systems	Windows 2003 Server - SP2 (64-bit) Windows 2003 Server R2 - SP2 (64-bit) Windows 2008 Server - SP2 (64-bit) Windows 7 (64-bit) Windows Vista (64-bit) (Administrator only) Windows XP Professional – SP3 (64-bit)
Processor (CPU)	Intel Quad-Core (for example, Intel Core 2 Quad) AMD Opteron
Memory (RAM)	8 GB or greater (>16 GB preferred)
Hard Drive Capacity	250 GB or greater in the temporary location. If running multiple examiners on the same machine, scale with the number of examiners. The amount of disk storage required is dependent upon the size of the original source data to be processed and which processing options are selected. A general guideline is to allow for 10 times the original data size.
Hard Drive Speed	7,200 RPM (10,000 RPM or faster preferred)
Number of Hard Drives Recommended	Three (Application, Temp, and LEF files should reside on separate physical drives)
Network Configuration	Gigabit Ethernet (GbE)

Table 3: The hardware, software and system requirements for ECC

A dongle is required to run any of the ECC components, including the ECC Web Server, ECC Desktop, ECC Examiner and the ECC Indexing service. These components can:

detect and use a dongle that is physically plugged into the computer hosting the ECC component

be configured to connect to a remote licence server (NAS) running on a separate computer with the dongle plugged into it.

A NAS allows several components to share a single dongle, which has the advantage that the system administrator does not need to monitor dongle deployment across several computers. For this reason, Guidance Software recommends using NAS for enterprise deployment of ECC.

Before launching the ECC Desktop, ECC Examiner or the ECC Indexing service, a dongle should be connected to the computer or EnCase needs to be configured to use a remote licence. The first time ECC starts, the ECC Examiner must be connected to the global database. After the first session, the ECC Desktop and ECC Examiner automatically connect to the global database.

3.3 ECC

The EnCase eDiscovery software runs within a framework known as the EnCase Command Centre (ECC). ECC uses Structured Query Language (SQL) databases to store the information required to search for, collect and potentially remediate live data across the network, and to process the collected evidence.

The EnCase Legal Hold, First Pass Review and Analysis components of eDiscovery run on a proprietary server known as EnCase Command Centre Web Server (ECC Web Server). Access to the First Pass Review and Analysis components are provided through a web browser (ECC Web Client). Figure 9 shows the interrelationships between the main components of the ECC.

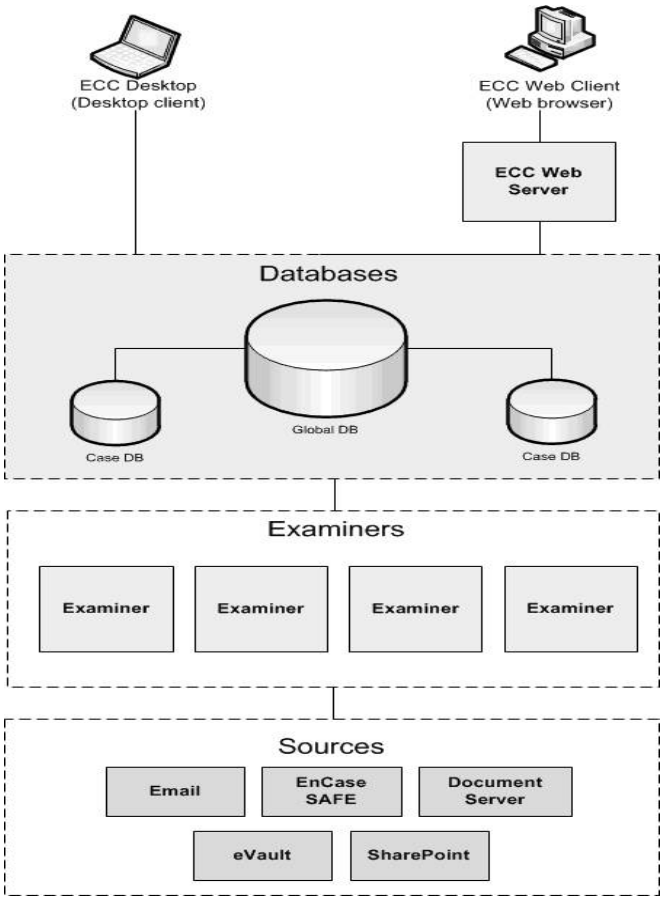


Figure 9: The main components of the ECC

3.4 ECC Examiner

The examiner component of the ECC performs the data collection and processing activities, including discovery, preview and data acquisition from target computers. Examiners are configured to connect to the global SQL database, which stores connections to the individual SQL case databases.

For the ECC Examiner to perform correctly, a data source may require libraries to be installed and properties to be configured. Multiple examiners can be configured for the same client software on the same source to increase processing speed.

The ECC Examiner component:

adds and lists the SAFE clients available on the network

provides log-on access to the SAFE for these clients.

adds and lists the network devices connected to each of the SAFE clients.

When using the Web Server component, First Pass Review and Analysis features of version 4.1, Guidance Software recommends using the Examiner 64-bit Indexing service, a specialised type of Examiner. The service needs to be installed on a dedicated computer running a 64-bit version of Windows. Furthermore, ECC Examiners must be able to communicate with the sources where data will be searched and with the global database.

3.5 ECC Web Server

ECC Web Server is a web application that is integrated within the EnCase eDiscovery process. To satisfy requirements for preserving data, the ECC Web component can be utilised with a browser to perform a variety of tasks that place restrictions on the use of documents and other electronic evidence.

By using the EnCase eDiscovery software with ECC Web component an organisation can:

identify and notify custodians of the need to preserve important data when litigation is anticipated or pending

receive acknowledgements from custodians that they recognise a hold is in place

interview custodians about their data using a questionnaire and receive their responses

enforce a legal hold by preserving relevant data in a forensically sound manner without employee assistance.

The EnCase Legal Hold, First Pass Review and the Analysis components of eDiscovery run on a proprietary server known as the EnCase Command Centre Web Server (ECC Web Server). Access to the First Pass Review and Analysis components are provided through a web browser (the ECC Web Client).

The ECC Web uses role-based security to determine what a particular user is permitted to do and which folders they are authorised to access. This type of security focuses on role identity not user identity. This approach allows the one-time creation of a set of permissions and the assignment of those permissions to an entire group. This also means that the system administrator need not individually configure each user's permissions.

Roles are set up within ECC Web and permissions are determined by enabling or disabling security descriptors for a given role. Roles are associated with active directory groups set up by the network administrator. Members of an active directory group have permission to do anything that the role(s) associated with that group enables them to do. If more than one role is associated with an active directory group, the members of that group have all the permissions from all the roles associated with the group.

Table 4 sets out the new eDiscovery functionality provided by the ECC Web interface.

Functionality	Description
Assess data early in a case to determine scope and strategy	- Immediately start examining data as soon as the first target is completed - Browse collected files and e-mails - Summarize collection results in report format - Categorize items with tags
Analyse indexed data to quickly find relevant, responsive items	- Search through indexed data using keywords and phrases - Calculate search term statistics - View overall search statistics - Refine searches by custodian or tag category
Review collected data for more in depth research	- View e-mail or file contents in a variety of ways - View e-mail or file properties and metadata - View e-mail conversation threads

Table 4: New functionality for the ECC Web with eDiscovery

At any time during the assessment and analysis process, an investigator can review the contents of the collected data. Print preview provides the ability to see a document or e-mail as it would be printed. Text view provides just the textual content without any formatting. You can also examine the forensic properties of a document or e-mail, as well as the metadata of the file or e-mail message. A collection and review history is also kept for each item.

3.6 User interface

ECC Desktop is a Windows-based user interface that manages the entire ECC system. Administrators can use ECC Desktop to:

configure databases and data stores

schedule collections

process jobs

analyse jobs

generate reports.

Item	Description
Case	A case is represented by data stored in its own SQL database.
Source	A source contains the data to be collected. This may include e-mail on a server (or archived PST or NSF files), document repository files, files on network shares, or files from the EnCase SAFE (representing live remote computers).
Custodian and its target	Custodians are users that are associated with target data on a particular source. Custodians can be imported from Active Directory or entered manually.
Case-specific report	A report describing the characteristics of the target data.

Table 5: ECC Desktop interface features

Jobs scheduled in ECC Desktop are processed by the EnCase eDiscovery software. Any scheduled jobs are automatically picked up and processed. These jobs may include uncompressing and indexing collected data, running keyword searches, and connecting directly to mail servers, document servers and workstations for data collection.

3.6.1 EnCase eDiscovery

The EnCase eDiscovery software is a judicially accepted solution for internal, legal and regulatory investigations. With EnCase eDiscovery, users take control and perform everything they want to do in-house.

The key objectives of an e-discovery package are to:

establish a consistent and scalable process to manage the identification, collection, preservation, processing, review and production of electronic data in a systemised and repeatable manner

use a defendable process that enables effective compliance, including the timely and systemised execution of litigation holds

reduce the size of preserved and collected data to only the potentially relevant material.

As Figure 10 shows, EnCase eDiscovery supports all the core stages of the e-discovery process: identification, collection, preservation, processing, analysis and review.

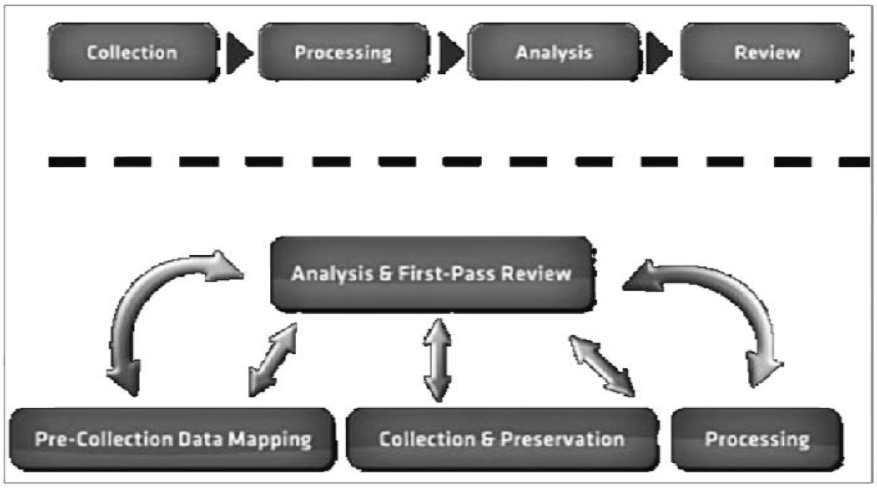


Figure 10: Core stages of eDiscovery process

EnCase eDiscovery provides the ability to perform case assessments before and during collection. The pre-collection scan, search analytics and first-pass review features enable investigators to conduct analysis throughout the process.

The methodology facilitates collaboration and communication between legal and IT teams. The early case assessment process can be used in an iterative fashion, allowing legal and IT teams to go through cycles of testing and sampling of various keywords, performing queries and analysis, and reviewing the results before and after collection.

Table 6 outlines some of the different features and functionality available within the eDiscovery platform.

Feature	Description
Legal Hold	Notifies at the earliest outset of litigation and helps organisations to address their duty to preserve relevant information through custodian identification, hold notification, questionnaires, document preservation, and hold releases.
Pre-Collection Analytics	Scan collects metadata only and analyses it, scoping and assessing the case and the size of the collection so that both legal and IT teams can plan accordingly. The need to perform pre-collection testing & analytics consists of the ability to do assessment work prior to actual collection as well as obtain statistics about the types of ESI that exist in a given environment, and report the locations of files.
Collection and Preservation	Technology allows users to keep working during the collection process with minimal to no interruption of productivity. Distributed technology culls at the point of collection, preserving only potentially relevant ESI
Processing	EnCase customers cull their dataset down by 90% compared to their existing methods. Greatly reduce the overall data set collected by setting aside irrelevant files based on keywords, hash values or any file system metadata property such as file type, date, path, or custodian. Reduce data further by removing duplicates on custodian or case level.
Analysis and First-pass Review	Browse and view documents and e-mails prior to indexing or perform linear review with hit highlighting, relevance rankings, e-mail thread and conversation viewing to identify responsive ESI, and tagging with comments to classify, categorize, and manage relevant content. Plus, EnCase can identify unique e-mails and documents per search expression, suggest search terms and provide corresponding hit counts.

Table 6: EnCase eDiscovery features and functionality

4.0 The testbed

4.1 Network topology

The investigators were based in the Records Department of the Knowledge and Information Management Division (KIMD) of the Welsh Government (WG). They used thin clients to access the dedicated testbed servers over the normal networking infrastructure. There was no dedicated networking infrastructure. The server room was offsite.

The thin clients and the servers were on a dedicated subnet. The subnet was not isolated. The network performance experiments were conducted in order to make projections for the scale of the data manipulation operations if performed over the whole estate. These projections are discussed in Section 5.8.

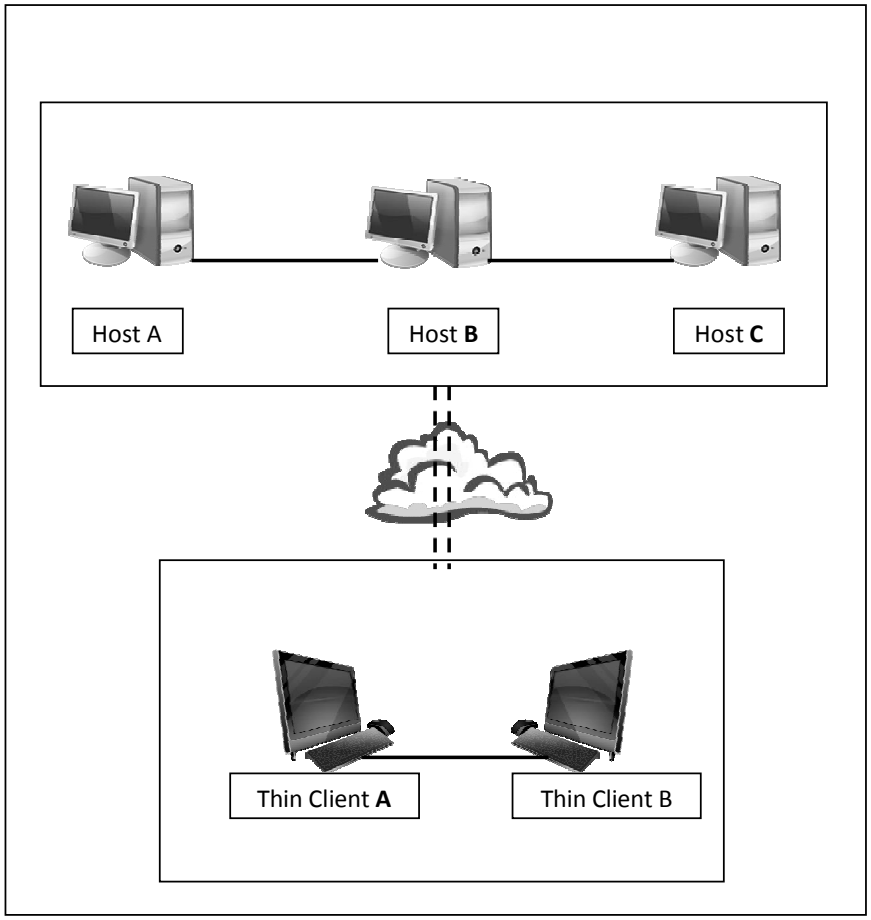


Figure 11: Testbed topology

4.2 Servers

The servers were members of a dedicated Windows 2003 domain. There were three physical machines (Hosts A, B, and C) running a number of VMs as required by the eDiscovery suite architecture.

Host A was responsible for managing the testbed domain and running the active directory.

Host B was responsible for running the server part of the eDiscovery suite. One VM was running the SAFE, one VM was running the ECC and one VM was running the Web Server and the database server.

Host C was responsible for running the examiner part of the eDiscovery suite. It was running several ECC Examiners.

Some other software was installed on the servers for the experiment, including:

MS Office

Windows Explorer

SQL Server Express.

4.3 Thin clients

The thin clients used by the investigators were standard desktop computers that WG employees use for their daily activities. They were locked-down thin clients that could not access the internet. The normal WG policies were enforced on the user accounts used by the investigators.

4.4 The experiment

4.4.1 ECC Web

Much of the EnCase eDiscovery process takes place in the ECC Web environment. This simple web interface can receive, compile and analyse custodian acknowledgments and create case-related interview style questions. The investigators used the ECC Web for analysing and examining the metadata information in the files on the testbed.

Cases must be reserved first from the ECC Desktop before they can be created in ECC Web. After a legal hold is started, custodians and holds can be viewed from ECC Desktop in the case view holds tab. Reports can be generated from either ECC Web or ECC Desktop. All legal hold editing is done in ECC Web.

ECC Desktop supports the ECC Web interface by:

- assigning the active directory keymaster group
- setting up reserved cases
- viewing the legal hold information
- running reports to obtain custodian and hold statistics
- creating and maintaining processing jobs for the assessments data set –these data sets are prepared for document and transcript viewing by ECC Web browsers
- running jobs using ECC Examiner
- creating and maintaining delivery jobs that include or exclude tags created and applied by ECC Web users.

Initially a case had to be reserved for ECC Web by the investigators. This was necessary to enable ECC Web. Before reserving a case for ECC Web, the investigators needed to create a specific SQL case database. This is the same as a database that would be used as a framework for an ECC Desktop case. Once the SQL database is created, a case can be reserved in ECC Desktop.

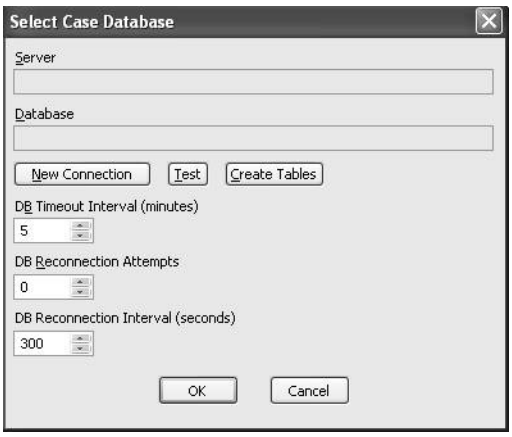


Figure 12: The select database dialog

After reserving the case, it can be created in ECC Web by first opening the 'cases' tab in ECC Web (see Figure 13). Then create the new case by clicking the 'new' button and choose a name for the case in the 'case' dialog display

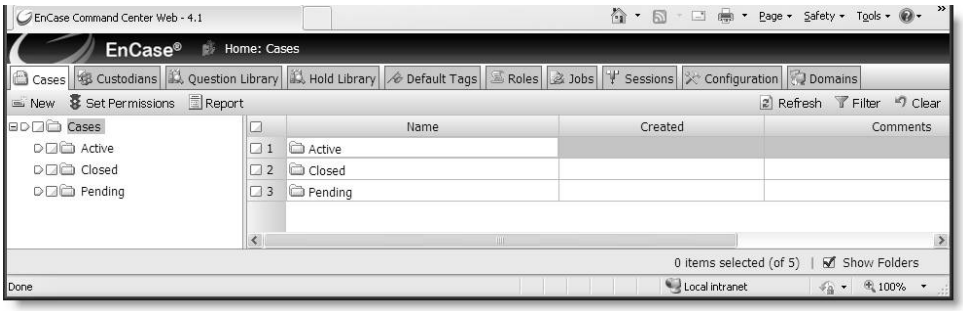


Figure 13: Cases screen on ECC Web

4.4.2 Summarising collection results

An investigator can view how much data has been collected, processed and indexed by looking at the summary numbers for each data set. These numbers also indicate the size of the collected document and e-mail stores.

1. Open the current case.

   From the ECC Web home page, click the ‘cases’ tab.

   Double click on the case name you want to work with.

2. Open the ‘data tab’. The data for the current collections is then displayed.

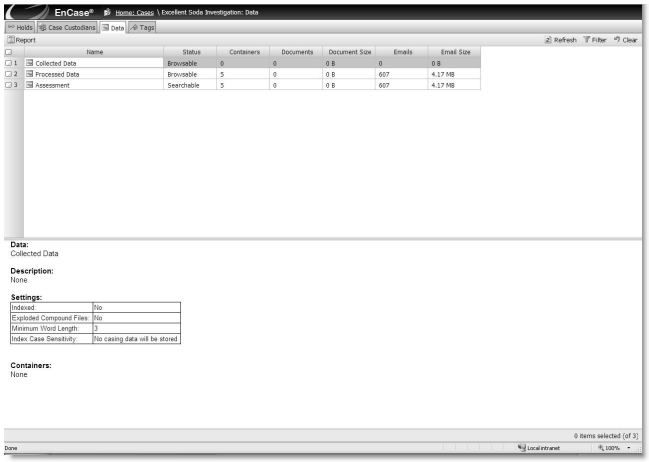


Figure 14: The data tab

The data tab displays various statistics for the data that has been collected or processed using ECC. There are three default data sets in ECC Web.

The ECC Desktop administrator determines exactly what can be seen in which data set.

Collected Data contains data gathered using an ECC Desktop collection job.

Processed Data contains data output from a processing job.

Data in the Assessment data set is indexed and can be analysed using the search query language available in ECC Web.

4.4.3 Conditions and criteria

The eDiscovery software collects and processes electronically stored information (ESI) to produce potentially relevant data. This is achieved by setting various criteria to determine what data is included in the responsive data set.

Four types of conditions can be set.

1. Metadata comparisons. The investigators set date ranges and file extensions to determine which system files were included. This facility can also be used to filter the metadata.
2. Keyword searches. Conditions are set based on keywords. It is possible to search for keywords in several ways, including by proximity, Boolean logic and index queries.
3. Matching files. Conditions are created to include (or exclude) file sets by size and hash value.
4. Compound file mounting. This can be used to expand compressed files to include their content in your culling.

Criteria sets are examination tools that can be used to search for and collect data. These tools include conditions, keywords, queries and sets of matching file hash values. When a collection, processing or delivery job is created, a criteria set has to be specified, which could include any or all of these components. Each job can have only one criteria set.

The ‘criteria’ tab in the Case View dialog screen is used to organise criteria sets for the jobs that need to be run (see Figure 15).

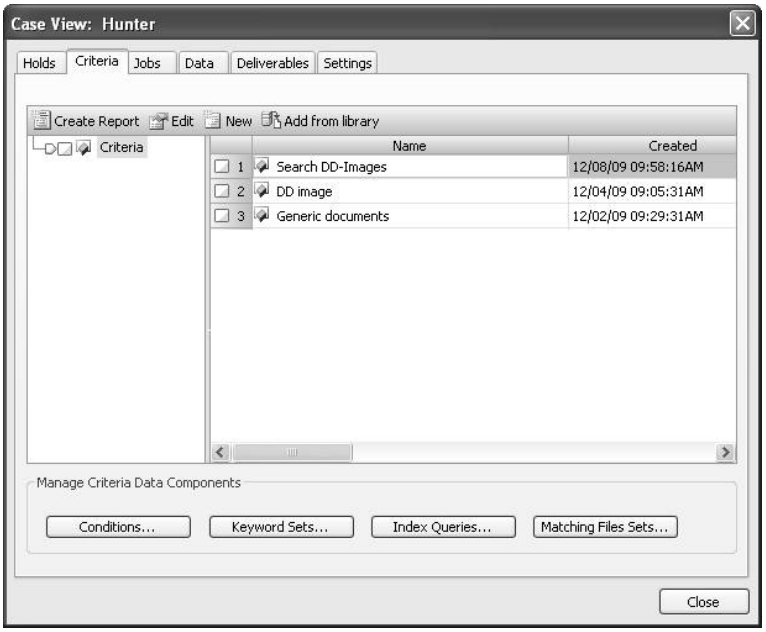


Figure 15: The criteria tab

When the data type is selected, the 'conditions' window displays. This can be used to add the conditions. Two additional components can be included in criteria sets under the 'responsive item conditions' group.

1. Index Query Settings is used to select the desired indexed case sensitivity option.
2. Compound file conditions can be set by clicking 'conditions' and writing a condition for files with an internal structure, such as OLE files or zip-compressed files.

The conditions in a criteria set are used to determine which data is collected or culled from different types of data sources. Each condition in a criteria set can contain keyword sets, index queries and matching file sets, so that an investigator can use a combination to meet the needs of a specific type of job. Conditions are saved in the global or case database to use when setting up the criteria for jobs.

The condition term options vary according to the source data type. The default is 'any' source type, with 'entry' as the data type.

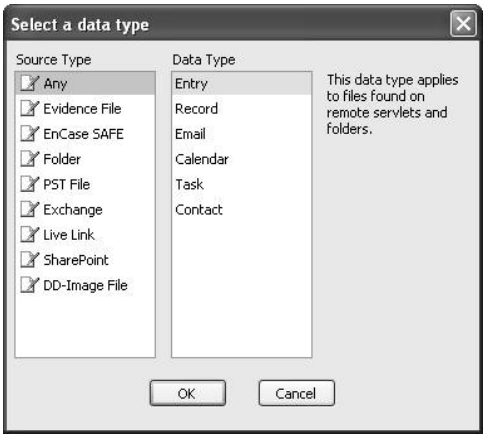


Figure 16: Data type dialog screen

There are two starting points for entering conditions.

To add to the global database, open the ECC Desktop window, then click the 'criteria library' tab.

To add to the case database, open the Case View window and select the 'criteria' tab.

These steps were used to create a condition.

1. Under Manage Criteria Data Components, click Conditions (see Figure 17).
2. In the Conditions window, click New.

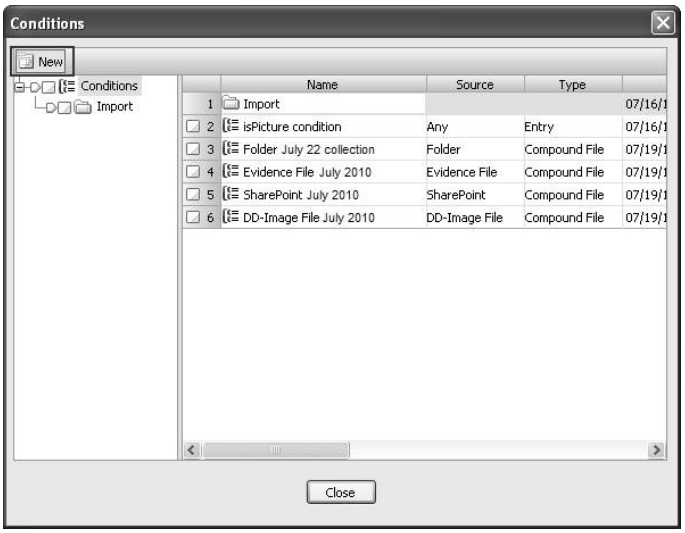


Figure 17: Conditions screen

3. In the Select a data type dialog, under Source Type, select the desired data store and the desired data type.
4. Click OK after selecting the data type. This opens the condition editing window for this type of data.

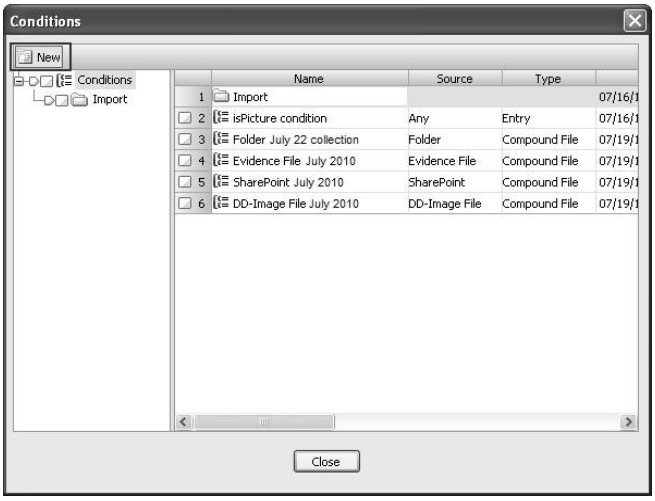


Figure 18: The condition editing window

5. In the Conditions window, give the new condition a name.

In some instances, it was necessary to construct search terms with conditions. Terms are composed of properties, operators and values. The condition term tool is under the 'terms' tab and displays properties, and then expands to display operators, a value text box and other options, depending on which of the properties are selected. Properties allow the investigator to specify what information to search for, and the operators specify how to filter this information.

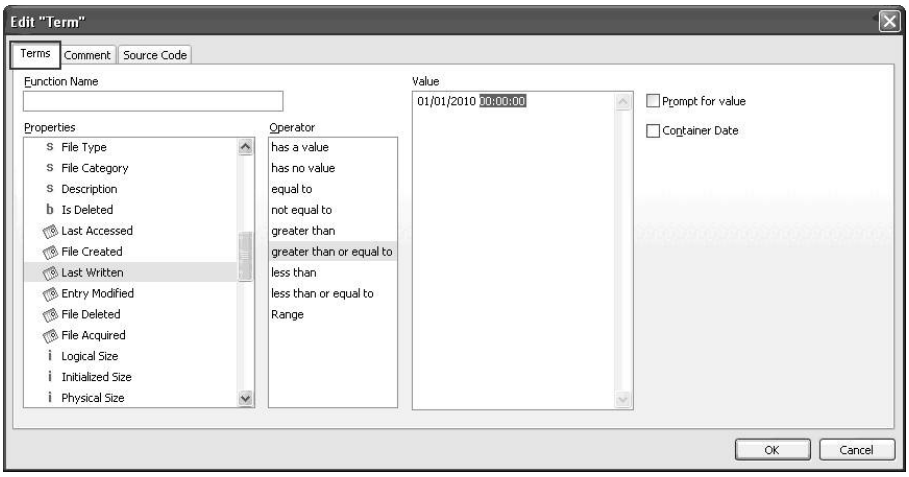


Figure 19: The condition term tab

For example, a search that was conducted on the testbed was to search by file extension. This specify the file types to be collected, as well as file types to be excluded from the search.

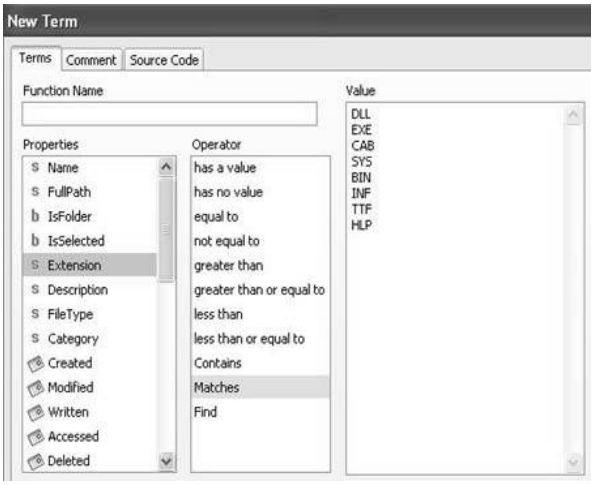


Figure 20: Searching by file types

Another search conducted on the testbed was to filter by date. This was used to apply that part of the retention strategy which excluded files created prior to a certain date.

Care must be taken when dealing with compound files. Within the New Term dialog, there is a feature for working with the contents of compound files (for example, the files within a zip file). This option, Container Date, applies the date of the compound file (parent) to any files nested within it (children) that do not have the date preserved. This option looks at files whose source type is EnCase SAFE, Folder and Evidence File, and whose data type is Entry. When searching the EnCase SAFE nodes, network directory folders and evidence files, the investigator must select the Container Date option to ensure that all nested files are collected if they match other criteria.

Dates are stored in GMT in the database. Under Value, the hours:minutes:seconds is always assumed 0 regardless of the operator chosen. It is best practice to specify them to avoid any possible confusion when others review the criteria.

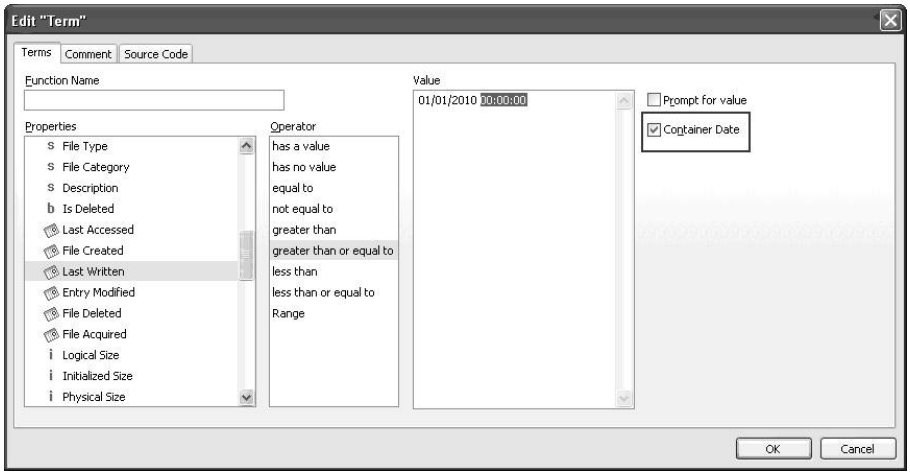


Figure 21: Searching by last written date

4.4.4 Keyword sets

Another feature of the ECC Web that proved useful in this study was the keyword search sets. The keyword sets were used within conditions as part of the search criteria. They can be set from the 'criteria' tab and the 'keyword sets' button. To create a new keyword set, follow these steps.

1. From ECC Desktop, select the 'criteria library' tab, or from the Case View, select the 'criteria' tab.
2. Under Manage Criteria Data Components, click Keyword Sets.
3. In the Keyword Sets window, click New.

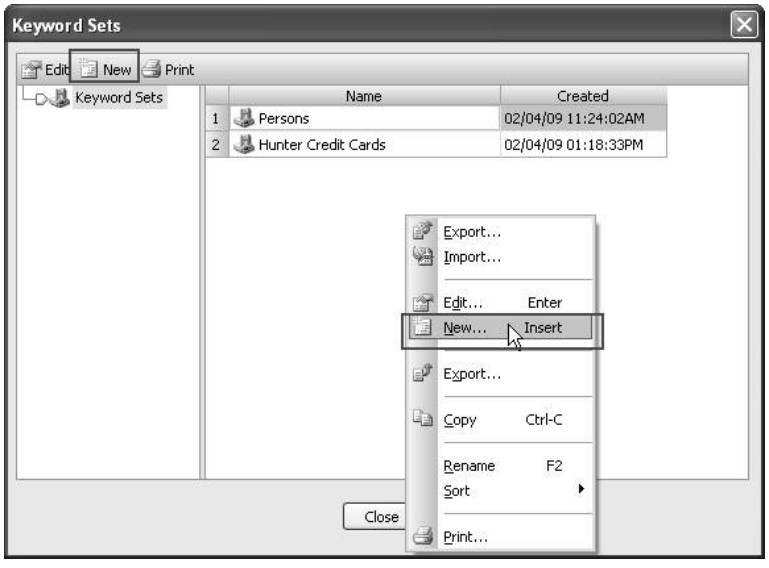


Figure 22: New keyword set screen

4. In the Keywords dialog, enter a keyword set name, then right click in the details pane and select Add Keyword List.

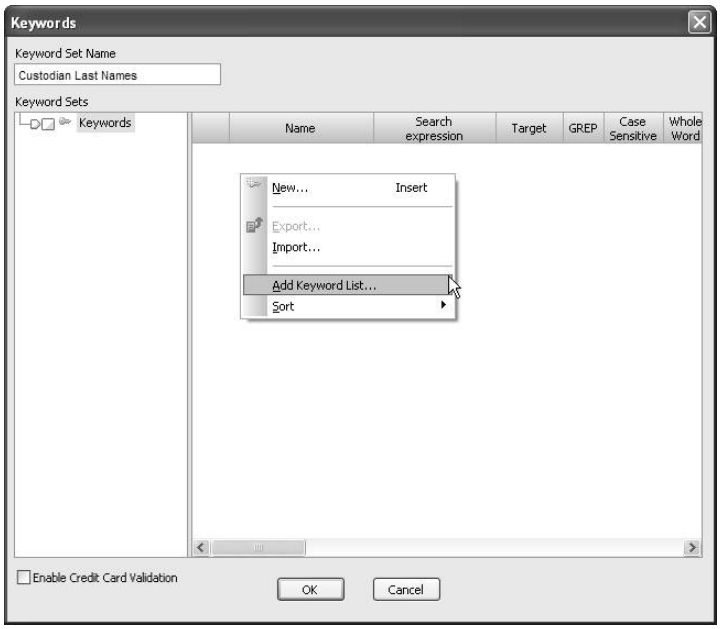


Figure 23: Keywords dialog

5. In the Add Keyword List dialog, enter or paste a keyword, one for each line. Select options to apply to the keywords, such as GREP, Case Sensitive, or Whole Word. These selections apply to all terms.

6. Click OK to complete the add keyword list process. The Keywords dialog displays the search expressions and any selected attributes.

Some search expressions were created in some cases in order to analyse the testbed data. A search expression is created by clicking New in the Keywords dialog, as shown in Figure 23.

Figure 24 shows a keywords list as displayed in ECC Web.

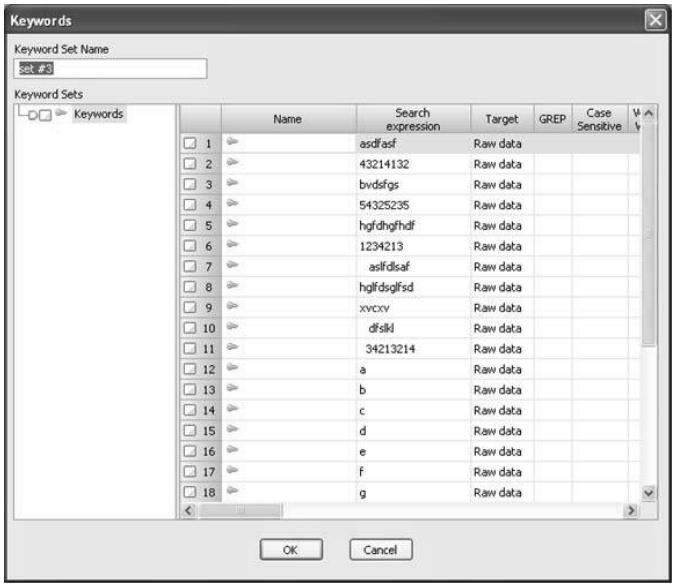


Figure 24: Keywords list display

4.4.5 Summary reports

There are two reporting options that can be used for looking at the gathered data. They were both used during the project. A current view report of selected data containers or a case screening report can be generated. Either report can be output in HTML, RTF or PDF format.

The case screening report shows a breakdown of the selection of data containers, and provides a summary of the information, custodian statistics and the types of file extensions found. It can be viewed by clicking on the Report button. The reports can be exported as a Microsoft Excel file for further analysis.

4.4.6 Browsing collected files and e-mails

Any data container generated from a job run in an ECC version, such as a Logical Evidence File (or LEF), can be viewed at any time.

The forensic properties, such as size and collection date of a collected item, a flat list of collected files or e-mails, a print preview of a collected file or e-mail can also be viewed at any time.

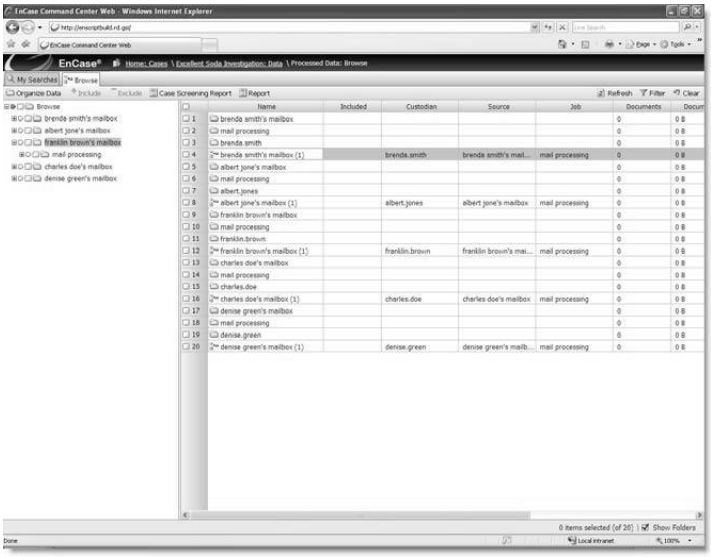


Figure 25: The browse tab

Any collected file can be downloaded in its native format, or any collected e-mail as a file in MSG format. Another option is to export any collected file or e-mail in PDF format for easy e-mailing or viewing outside the system. The 'items' tab (see Figure 26) shows the contents of the data container.

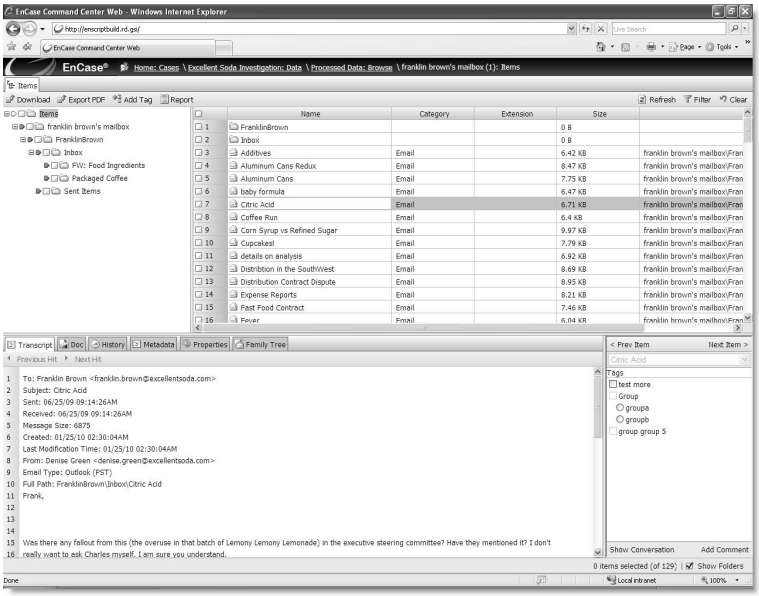


Figure 26: The items tab

ECC Web's transcript view displays the text contents of a file or e-mail, and includes all the information in the item that is stored as text. In this view, the text is reformatted for easy reading. The transcript of an e-mail includes the subject line and basic

message headers (to, from, cc, bcc and time stamps). This does not include text stored as images in the file or e-mail. The print preview of the file or e-mail is displayed from the ‘doc’ tab.

During pre-processing, any internal metadata available in the file (such as author or organisation details) is extracted from files and e-mails. This metadata can be accessed either while browsing or searching. It is only available for searchable data that has already been indexed.

The internal metadata of the selected item displays as a text block in the ‘metadata’ tab.

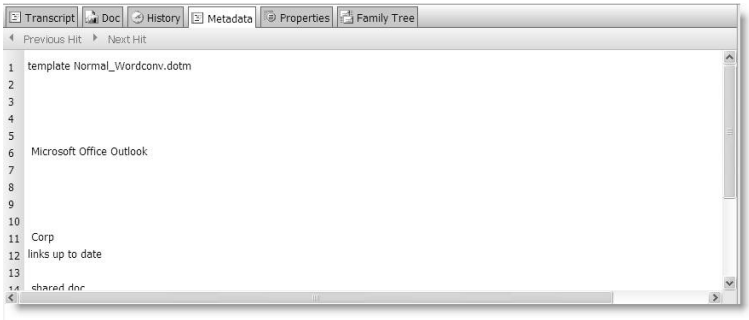


Figure 27: The metadata tab

During the project, the metadata from a sample of files in the testbed were previewed in this tab in order to identify the fields that contained records that could be used for further metadata analysis.

4.4.7 Categorising items with tags

The results of searches needed to be tagged in order to be exported. Tags are used to separate data items into different categories.

Tags are managed from within the case in the ‘tags’ tab. Tag groups are displayed in the tree view on the left, and the tags within each group are shown on the right. Tags can be used to categorise or organise files and e-mails quickly and easily. These tags are stored in the item’s history and may be exported to other review platforms.

The ‘add tag’ button is used for tagging an item. Then the Add Tag dialog appears (see Figure 28).

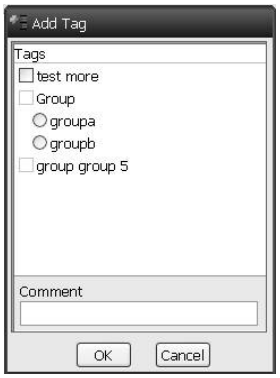


Figure 28: The add tag dialog

An item's history is updated every time a tag is added or removed, or a comment is added to that item. This history includes logging the date and time of the change, a description of the operation, details of who initiated the event, and a record of any comments that were added.

4.4.8 Metadata analysis

The collected data was indexed for further analysis and afterwards imported into the ECC Web. The indexing is a separate process and, once completed, enables an investigator to search through the text in the data set to find precisely what is required. Keywords and phrases can be found within any indexed data set. Indexed data sets are marked ‘searchable’ in the status column of the ECC Web Data tab.

The metadata analysis in this project required complex searches to be constructed. ECC Web can be used to search for more than simple lists of keywords. By using the ECC query language, an investigator can create a complex set of search criteria that can expand or contract the search results in very specific ways.

These are the steps required to create a new search.

1. Open the current case.

   From the ECC Web home page, click the ‘cases’ tab.

   Double click on the case name you want to work with.

2. Open the ‘data’ tab. The data for the current collections in the case displays.

   In the ‘Status’ column, ‘browsable’ means the investigator can browse through the data in the set. ‘Searchable’ means that the set has been indexed and you can run searches against the data in the set.

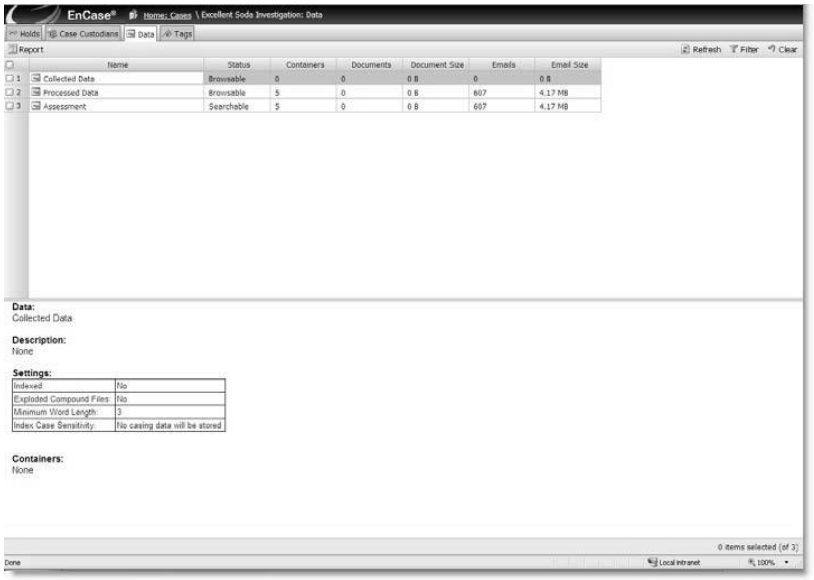


Figure 29: The data tab

3. Double click on a searchable data set. The ‘my searches’ tab opens, showing any searches that may have been saved.

4. Click ‘New’. The New Search dialog box appears.

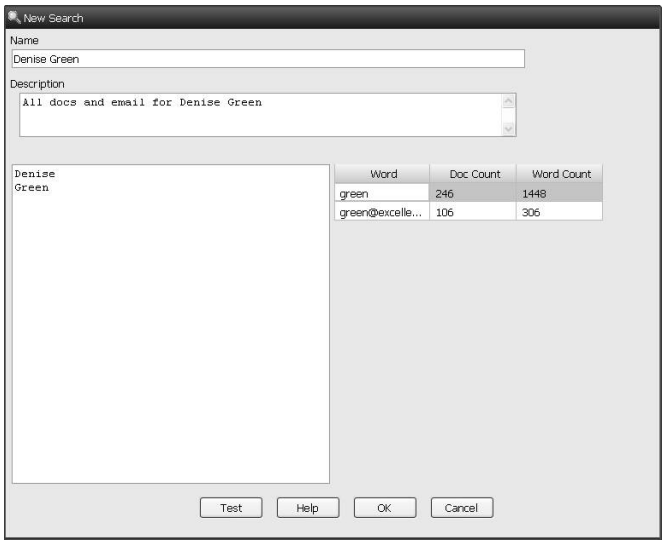


Figure 30: New search dialog

5. Enter a name for the search and a term into the Search Term text box.

   For complex searches, special operators and wildcards can be used. For example, the fields that are of interest can be restricted by using the bracket ([ ]) field specifier. [MetadataTranscript] searches the internal document metadata,

such as the author's name and time stamp of last revision. This was used for the metadata analysis.

Running a search may take a few minutes, depending on the number of independent terms in the search and the size of the data being searched. From the ‘my searches’ tab double click on the search you want to run. The ‘terms’ tab displays. From either the terms, tags, or custodians tabs, click Run.

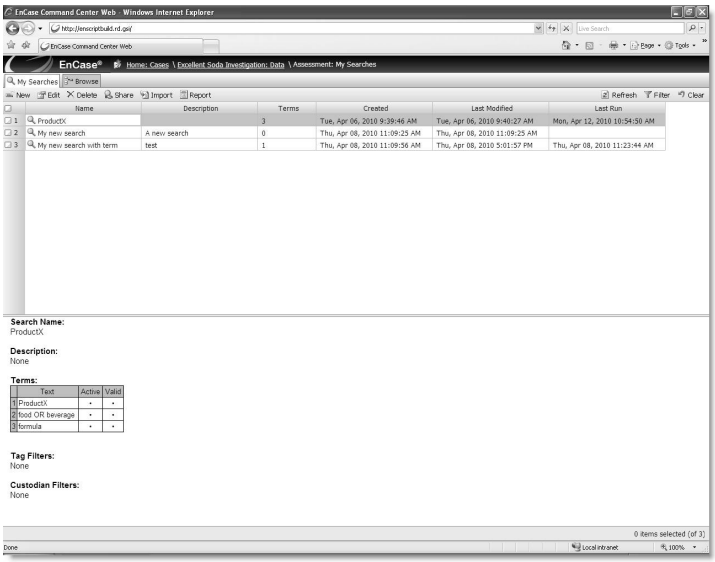


Figure 31: My searches tab

As the search runs on the server, the Status column for all active terms changes to ‘Out-of-date’.

When the search is complete, the Status column for all active terms changes to ‘Complete’ and the statistics are updated.

The ‘search hits’ tab is populated with items responsive to the search. By default, the maximum number of hits returned is 100,000.

Deactivating a term removes it from the search without deleting it. This allows an investigator to determine the effect of removing a particular term or keyword set without having to delete and recreate the search over and over. Inactive terms are ignored by the search engine.

If a search term that has been previously deactivated needs to be used again, it can be activated easily. Click ‘Activate’ on the term that needs to be turned on. The Status column of the selected terms switches to ‘Out-of-date’ until the search is rerun.

Each term in a search is run independently against the index and statistics are calculated on a per-term basis. On the ‘terms’ tab, ECC Web provides a set of statistics that enable an investigator to assess which terms are the most effective.

‘Total Items’ is the total number of items that were responsive to the search term.

‘Unique E-mails’ is the total number of e-mails that were responsive only to the search term and not responsive to any of the other terms in the search.

‘Unique Docs’ is the total number of documents that were responsive only to the search term and not responsive to any of the other terms in the search.

‘# E-mails’ is the total number of e-mails that were responsive to the search term.

‘# Docs’ is the total number of documents that were responsive to the search term.

‘E-mail Size’ is the total size of all the e-mails responsive to the search term.

‘Doc Size’ is the total size of all the documents responsive to the search term.

The total and unique numbers can determine the effectiveness of individual terms.

5.0 Findings

5.1 De-duplication

Before the de-duplication process can be undertaken, it is necessary to acquire the relevant data. Data acquisition within digital forensics refers to the process of imaging a computer or computer-related device in a manner that does not alter or modify the contents in anyway.

Imaging is best described as creating a bitstream copy of the original hard drive. It provides a panoramic view of all the data present within the system, like a photographic snapshot of the highest resolution that records even the most miniscule detail, which in forensic terms can mean the deleted bits and fragments of files that have been deleted but not fully replaced by other files.

The acquisition process is validated by an automated hash value that is calculated before the data is imaged. The importance and validity of the compare-by-hash procedure is described in section 2.2.2.2.

The process of data acquisition is nearly always a lengthy procedure and hugely processor intensive. In this case, the acquisition of all the system data took just over one day, specifically 24 hours and 3 minutes.

Once the acquisition phase was complete, an initial analysis and filtration was conducted on the imaged data. Figure 32 provides an overview of the original sample of data and comparison with the data collated after Phase 1 of the project was completed.

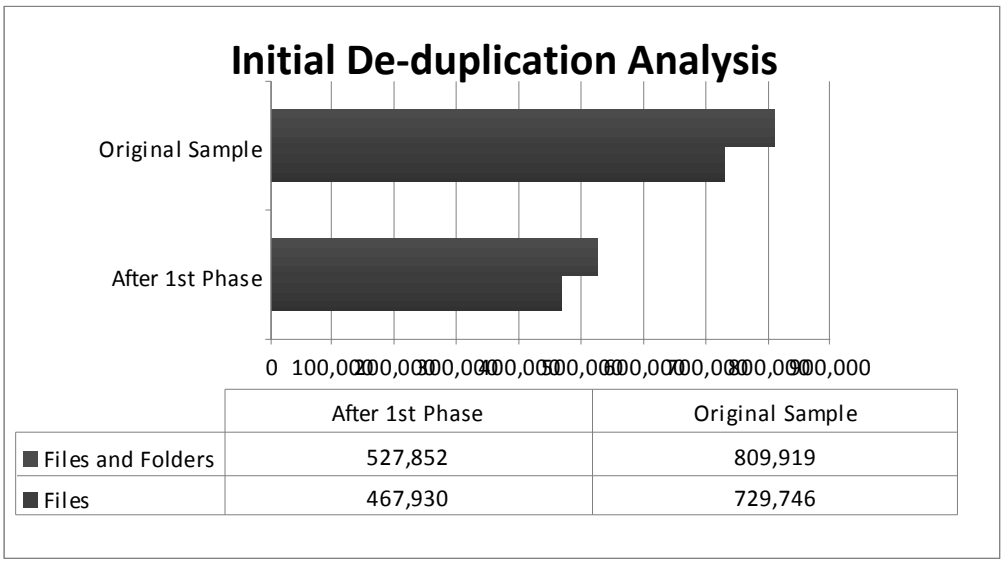


Figure 32: Initial de-duplication analysis

The original sample contained:

809,919 files and folders

729,746 files

809,919 – 729,746 = 80,173 folders

Once the first phase of the project was completed, there were:

527,852 files and folders

467,930 unique files

This first phase demonstrated how many of the files were duplicates within the original sample. Its findings are that **35.88%** of files in the system were duplicates. Duplication has occurred through:

multiple persons working on the same document

multiple copies of the same file being saved

multiple persons saving identical files to the system.

De-duplication obviously produces savings in data storage. The original volume of data was 211.9 gigabytes (GB); after de-duplication there remained 149.4GB of data. This generates:

62.5GB of free space on the system

The de-duplication process therefore liberates **29.49%** of the previously occupied storage space. This not only increases efficiency, but provides more room for data storage and is a more cost effective use of IT resources.

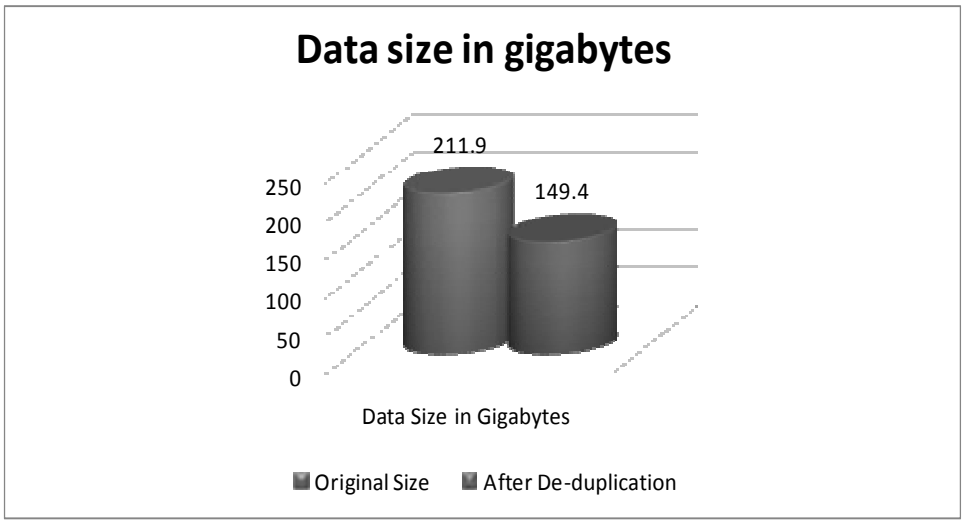


Figure 33: Data size in gigabytes before and after the de-duplication

5.1.1 Types of data before and after the de-duplication

Once the data was analysed and de-duplicated, filtration of the files enabled specific file types to be classified. It was then possible to calculate the percentage of storage space occupied by each file type.

The majority of files present within the system are MS Office files, most notably in the .doc file format (that is, MS Word files). Some 53% of the files on the system are .doc files. There are also at least 11 other types of files located within the data set. This illustrates the varied nature of work undertaken by those utilising the file system.

Table 7 shows the breakdown of each type of file that was found within the data image of the system. The comparison between the original data set with the de-duplicated data set shows which type of files occupy the data space and highlights the type of file that is duplicated most often.

Data Type	Original data		De-duplicated data	
	File size	File percentage	File size	File percentage
doc	383676	53%	247411	53%
xls	48698	7%	34219	7%
ppt	11979	1%	7054	1%
pdf	38956	5%	20338	4%
msg	75811	10%	60759	13%
rtf	27478	4%	18429	4%
txt	11220	1%	7633	2%
xsl	71	Less than 1%	71	Less than 1%
image files	58522	8%	24179	5%
compressed files	88	Less than 1%	38	Less than 1%
htm, html	37466	5%	25952	6%
other files	35781	5%	21847	5%
Total Files	729746	100%	467930	100%

Table 7: Different types of data before and after the de-duplication

It is worth noting that there are the same percentage (53%) of MS Word documents in both the original data set and the de-duplicated set. The figures in Table 7 show that there were:

136,265 duplicate MS Word documents present on the system

5.1.2 Charts

This data can be presented graphically. Figure 34 shows the percentages of different file types found in the original data.

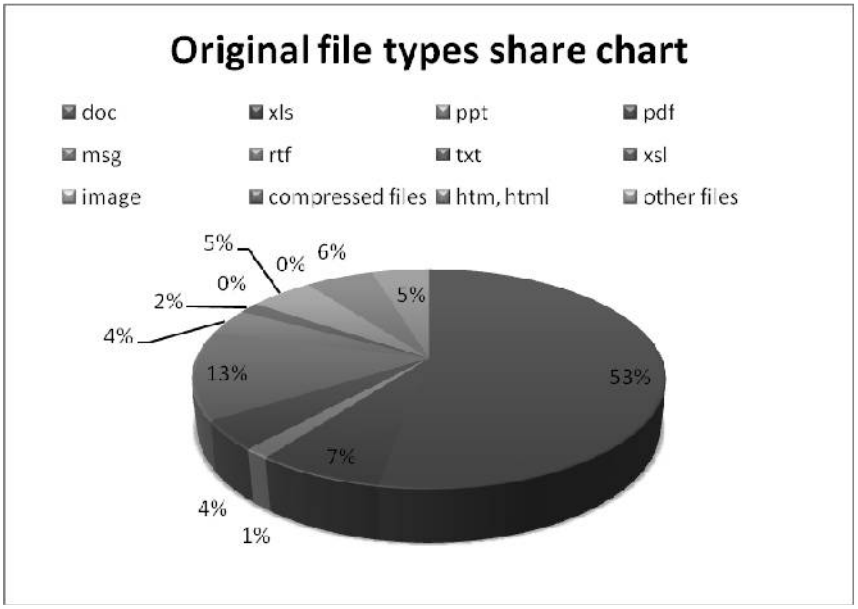


Figure 34: File types in original data

Figure 35 presents a similar analysis for the de-duplicated set. It provides an overview of what type of file is present and how often it is uniquely occuring.

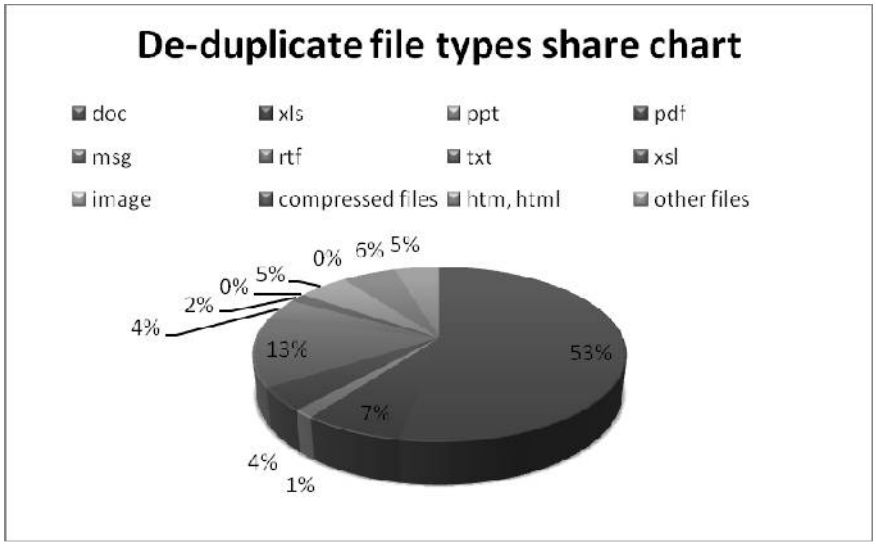


Figure 35: File types in de-duplicated data set

5.2 Data to migrate

A script that identifies the data that are to be excluded under the general retention strategy rule (see section 2.2.2.2) was applied immediately after the de-duplication. This is Phase 2, Process 2, Activity 1 of the methodology.

A query that identified all the files that were created or modified 10 years ago was written and executed in EnCase. The query returned the following results:

Modified before 25 March 2001	:	826
Created before 25 March 2001	:	608

We would normally expect this query to show more created than modified files, but some files didn't have a creation date. This usually occurs when a file is closed without being saved and its date reverts to the last-modified date. A further examination of the files that showed up in EnCase without ‘created’ dates indicate that they were all created on 1 January 1970. This is obviously misleading system information.

5.3 Metadata identification

Metadata, simply defined as data about the file attributes, provides information that helps classify and understand files in terms of their modified, last accessed and created (MAC) times. All files have metadata and by studying the attributes associated with each file, regardless of their file type, basic fundamental details can be established, including:

the time the file was created

the author of the file

the last time the file was accessed

the last time the file was modified

any other attributes related to file creation and maintenance as well as any organisational information embedded within the system setup that automatically adds itself to a file's metadata.

Out of the 467,930 unique files contained in the test bed, there were 354,233 files with metadata. Therefore:

75.7% of the files had metadata

Consequently, **113,697** out of the 467,930 unique files within the testbed data set did not contain information that allowed a metadata-based classification. These are considered as residual files at this point. However, further analysis could assist with classification.

5.4 Classification of data

An analysis of the metadata from files within the data set showed that they did not contain rich metadata. It was only possible to identify the creator of every file, along with MAC times and dates. Through further analysis, it was possible to ascertain these attributes within the metadata.

1. File name
2. Company name, which corresponds to the WG department name in most cases
3. Author's name, which corresponds to the member of the staff that created the file
4. Creation date

The file name metadata is not particularly useful. Many file names are similar if not identical. Reasons for this, other than duplication, are that several users worked on

the same file and saved their own versions, and that many file names are generic so that different files get saved with identical names.

Other constraints that have hindered the process using this metadata to classify the data.

1. The author name is not a unique identifier. There are many WG staff with similar or identical first names and/or surnames
2. The author name is does not uniquely identify the department in which the document was created. An author might have worked for several departments over time.
3. The staff list required to identify the department of the member of staff that authored a particular file is only available from 2006–2010.

5.4.1 Classification by department

An assessment of the metadata in a sample of files from the testbed focused the classification towards a departmental categorisation. The information located within the 'company' field of the metadata can be used for an initial classification of the files by department. This provides a hierarchical system that enables the files to be classified according to the department from which they originate.

The WG organisation departmental structure from 2010 was used for this categorisation. Table 8 illustrates the categories that have been created based on this structure. It shows the total number of files allocated to each department category.

Note that further categories had to be created to classify files that could have originated from more than one department, or which originated from outside of the WG and could be linked with several departments. In some cases, the actual content of the files will need to be individually examined. Other files have been categorised as consultant, council or personal.

Consultant is used to classify files originating from consultancy firms that the WG is using for specific operations. The consultant's name has been extracted from the metadata.

Council is used to classify files originating from a council. The council's name is retrieved from the metadata. The relationship between a council and the WG department could involve communication with several departments. Therefore, it was decided that the councils should form an individual category.

Personal is used to classify those files that were identified as being personal communications. Users tend to use their work computers for some personal communications. It is therefore unsurprising that there was some personal data in the files.

Department	Total number of files
Constitutional Affairs	26431
Corporate Services	99
Economy and Transport	2232
Education	116857
Environment, sustainability and housing	27
Finance	414
Health	3
I.T.	5357
Consultant	2968
Council	2771
Personal	535
Need to check the authors/ general information	88724
Total	**246418**

Table 8: Classification by department

Out of the total **354,233** files that contained information in the 'company' metadata field, **157,694** files provided sufficient metadata to make a classification by department. However, many files that could not be directly classified. There were **88,724** files which did not specify a department in the company name metadata field. Instead, they often had more generic information, such as WAG or National Assembly for Wales. These files were further analysed in the next activity to see if a departmental classification could be made on the basis of the author's name. The remaining **107,815** files contained invalid or unsearchable names in the company name field, such as 'Home user'. This provided insufficient information that could assist the data classification.

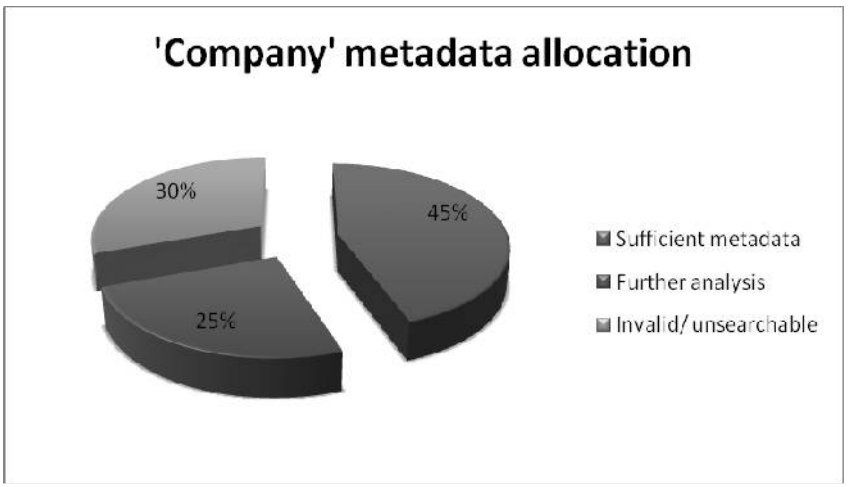


Figure 36: 'Company' metadata allocation

As Figure 36 shows, 45% of the files can be directly classified by department based on the extracted 'company name', while another 25% requires further analysis. The remaining 30% of the files had invalid or unsearchable metadata. Some of these files could also be classified following analysis of the secondary classification source (author's name).

```
Name Of Document: \AT S\1. CONTACT DATABASES ALL BRANCHES\2008-09\PI  Branch South
Wales AT Contact List Mail Merge Table Main.xls
Company: NCETW
Title:
Subject:
Author: linda.wilkes
Keywords:
Comments:
Last Saved By: williamsd16
Template:
Version: Microsoft Excel
Revision:
Create Date: 07/Jun/2006 08:59:35AM
Last Revision Date: 18/May/2009 04:10:05PM
Last Print Date: 11/Feb/2009 12:25:34PM
Number of Pages: 0
Number of Characters: 0
Number of Paragraphs: 0
Number of Words: 0
Hash: F81CF779D9A146A2ECF1A100728AE9CE
```

Figure 37: Metadata extraction, example 1

Figure 37 is an example of the metadata extracted from a single file. It has been illustrate the approach used in categorising the information. As stated in section 2.2.2.2, the title field usually contains insufficient information for classification purposes. In this example, the title field is empty. The 'company' field is much more useful for making departmental classification. The information extracted from these fields can be matched against the WG's departments and division. In this example, the company name is given as NCETW. This acronym stands for National Council Education and Training Wales. It was established that this is part of the Department of Education, Lifelong Learning and Skills. Therefore the file can be extracted and placed in the appropriate folder.

The Education department as expected had the argest number of files: 116,857 unique files can be categorised within the Education department, and this is more than 50% of the classifiable content on the system.

Figure 38 provides a further overview of the total number of files allocated by department. It compares the overall allocation and shows the departments from which the largest amount of unique data originated. This categorisation provides an overview of the hierarchical structure of the files on the system. It can be further analysed to show the type of file and the originating division within each department.

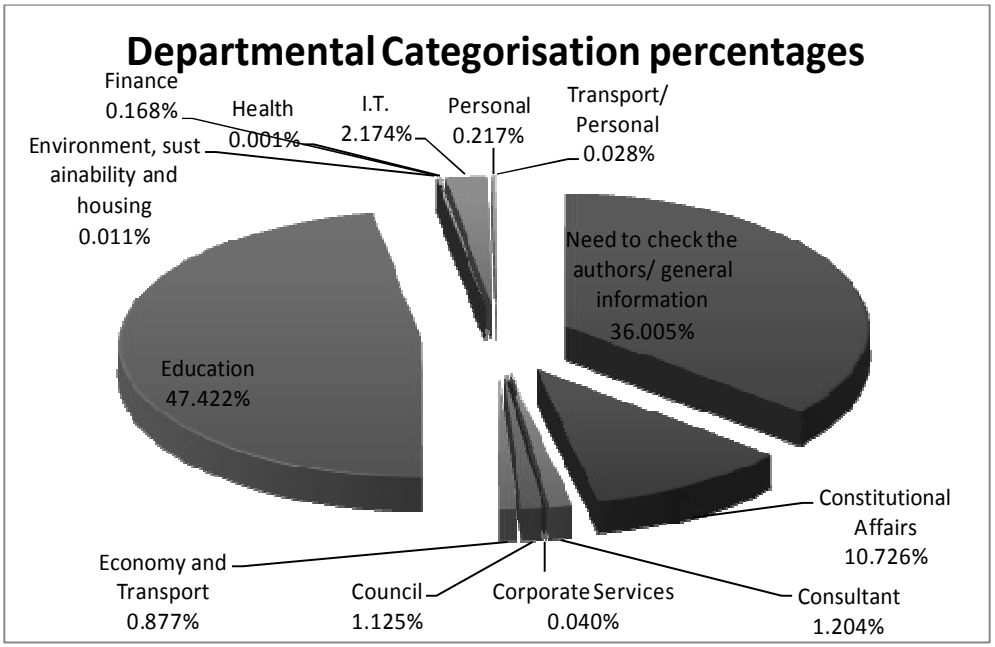


Figure 38: Departmental categorisation: overview

5.4.2 Classification by 'author'

The classification by author is the secondary classification source. It is used when there is insufficient evidence in the company name metadata field to provide a classification by department. The extraction and analysis of the 'author' field is a more complex process.

The same EnScript routine that extracted the 'company' information was also used to extract the 'author' metadata. The next stage is to match the author field with a specific department. This process is best illustrated by an example. Figure 40 contains the metadata from a single file. There is no information in the company field. Therefore, an attempt is made to identify the author of the file, and then to infer the originating department from the author. The author field in Figure 40 contains 'sargentd'. This appears to be a user name rather than someone's full name. A

search in the WG staff list database shows that there has been an employee named Sargent D. (the initial is used instead of the full name for privacy reasons). During 2006-2010, the period for which the staff list is available, this employee has worked for the Department for the Economy and Transport and for two different divisions within the Education department.

```
Name Of Document: \AT S\1. CONTACT DATABASES ALL BRANCHES\2008-09\080611 Directors
of Education All Wales.xls
Company:
Title:
Subject:
Author: sargentd
Keywords:
Comments:
Last Saved By: sargentd
Template:
Version:
Revision:
Create Date: 03/Oct/2007 11:27:02AM
Last Revision Date: 07/Nov/2008 11:06:38AM
Last Print Date: 30/Jun/2008 10:08:43AM
Number of Pages: 0
Number of Characters: 0
Number of Paragraphs: 0
Number of Words: 0
Hash: 644230E6CDBF6E1A003137B604E389D0
```

Figure 39: Metadata extraction, example 2

The next step is to use the tertiary classification source – the creation date of the file – to find out which department the author was assigned to when the file was created. Figure 39 shows that the creation date was 3 October 2007 and at this time the author worked for the Department for the Economy and Transport. This file could therefore be classified as originating from the Department for the Economy and Transport

Figure 40 shows the number of files that were referred for further analysis after the classification based on the company name metadata that contained the author's name in the metadata. From the 88,724 files referred for further analysis, there were 16,380 files that contained the author's name within the metadata. Some names appeared very infrequently. There were 1351 files linked to authors whose names appeared three times or less in the metadata. A check on a sample of these names found that most could not be found on the staff list. Due to the time constraints of the project, it was decided to focus on those authors that were linked to several (that is, more than three) files. Therefore, these 1351 files were added to the set of residual files. That left **15,029** unique files where an attempt was made to link author's name within the metadata to a department.

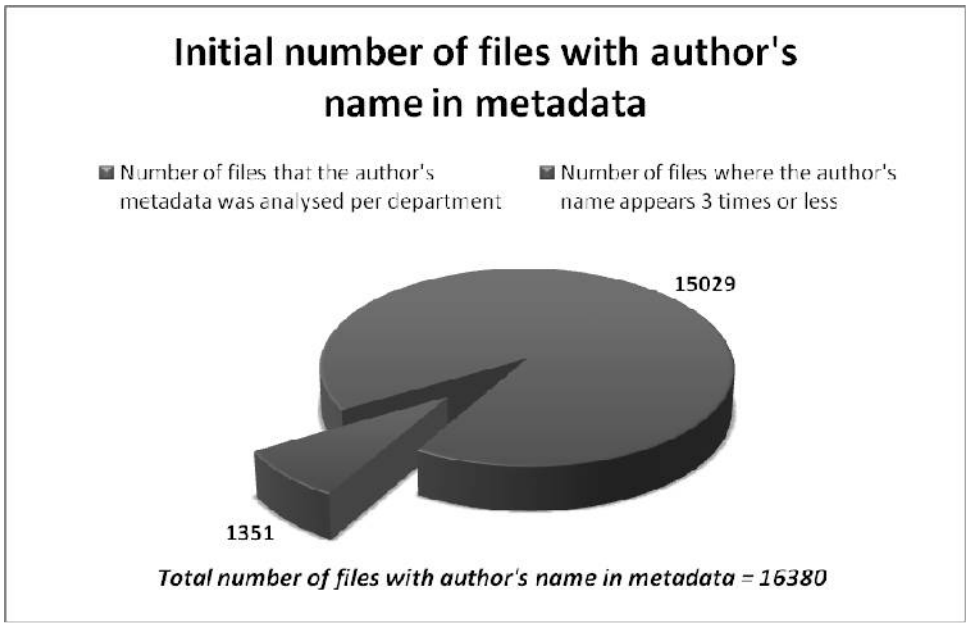


Figure 40: Number of files with author's name in metadata

Table 9 shows the number of files allocated to each department based on matching the author's name with the WG's staff list. It shows that there are 15,029 unique files that were analysed have been sorted into 15 categories.

Department	Total Number of files per Department
Multiple Departments	916
Ungrouped	83
Education	3094
Economy & Transport	444
Environment, Sustainability & Housing	6
Finance	214
First Minister	375
Health & Social Services	92
Public Services and Local Government	122
Sustainable Futures	43
I.T.	392
Legal Services Department	10
Invalid/ Unknown	4384
Multiple Names	2287
Not Found 2006-2010	2567
Total	**15029**

Table 9: Using author name to classify by department

Of these 15 categories in Table 9, five have been set up to for files that could be linked for various reasons to a specific department. These five categories are defined below. The first two categories can be subject to further analysis, but the remaining three categories cannot be investigated further to obtain an originating department.

1. Multiple departments – according to the staff lists, the author worked in more than one department. This need further analysis. The creation date must be used to categorise these files by department.

2. Ungrouped – files that do not originate within WG departments but from an outside organisation, such as a council or consultancy firm.

3. Invalid/unknown – invalid or unknown names have been recorded in the metadata, including 3871 files with no 'author' information.

4. Multiple names – files that cannot be attributed to a specific author, because the author does not have a unique name, or files that are attributed to several authors working in more than one department

5. Not found 2006–2010 – files that were created before 2006. Since the project did not have access to staff lists covering this period, it is not possible to know in which department the author was working when the file was created.

Figure 41 shows how the unique author files have been categorised by department.

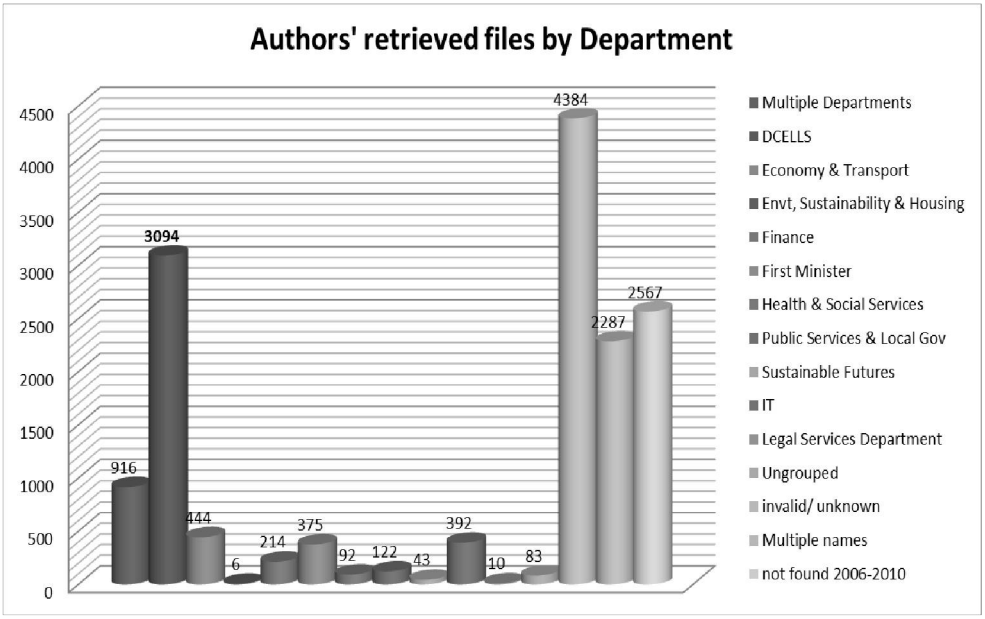


Figure 41: Classification of files by department based on author's name

In many cases the information in author's name metadata was insufficient to classify a file by department. As Figure 42 shows, of the 15,029 files analysed only 39% (5791 files) contained enough information for them to be classified into the hierarchical file storage structure.

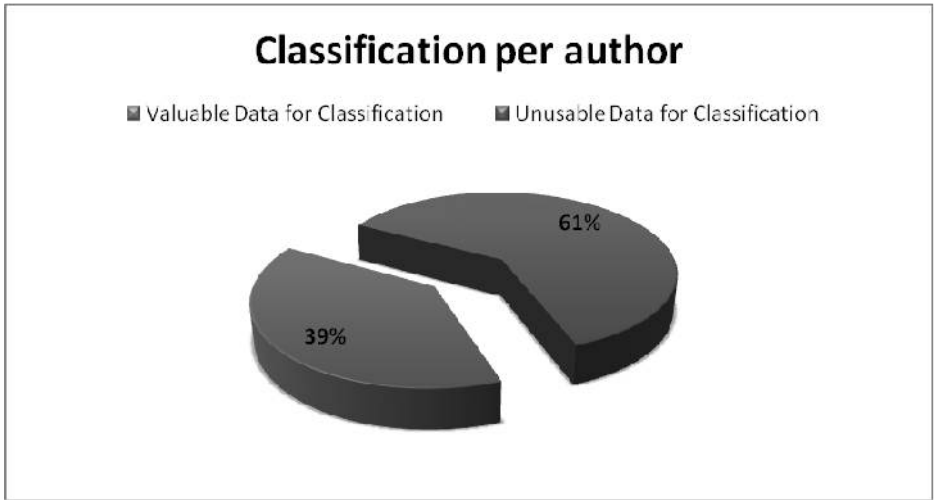


Figure 42: Success of attempting to classify using author's name

Table 10 shows how these 5791 files were categorised. The majority of the files belong to the Education department (DCELLS). These 3094 files were distributed between its different divisions, including Business Development, Children and Schools, Lifelong Learning and Skills, Qualifications and Curriculum, Higher Learning Group.

Authors By Department	Valuable data for Classification
Multiple Departments	916
DCELLS - Education	3094
Economy & Transport	444
Environment, Sustainability & Housing	6
Finance	214
First Minister	375
Health & Social Services	92
Public Services & Local Government	122
Sustainable Futures	43
IT	392
Legal Services Department	10
Ungrouped	83
Total number of files	**5791**

Table 10: Classification of files by department by using author's name

Figure 43 provides an overview of the data that has been classified through an examination of the 'company' metadata and then further classified through the author metadata combined with the department metadata and the provided staff list.

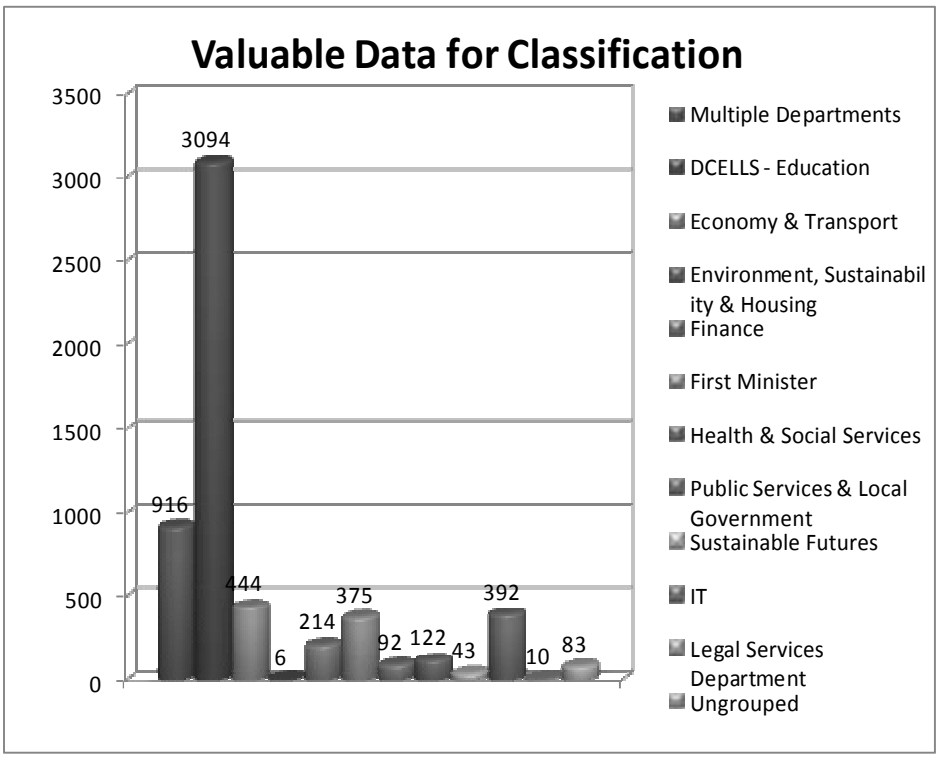


Figure 43: Overview of data that has been classified

Some 61% of the data extracted from the author name metadata field is unusable for classification. Table 11 groups the reasons why this data could not be used into three categories.

Reason why data is unusable	Number of files
Invalid/unknown	4384
Multiple names	2287
Not found 2006-2010	2567
Total number of files	**9238**

Table 11: Unusable author name data

This figure of 61% seems high but there are some extenuating circumstances which prevent the categorisation process being more comprehensive. One of these circumstances is the lack of valid staff lists and associated information for the years

before 2006. There were 2567 files that could not be associated with a department because they were created by their authors prior to 2006.

The invalid and unknown category contains files that could not be classified for one of these reasons.

1. Unknown author
2. Unknown department
3. Invalid or missing file signatures
4. Unrecognised file formats
5. A combination of the reasons 1–4

Some author names appear in many documents. However if there is no matching record in the staff list, then there is no basis for matching the name (and, therefore, the file) to a department. It might be possible to identify the department from the contents of the document. However this is necessarily a subjective process, and could risk the accuracy of the classification. In any case, it is impossible to go through all the files due to time constraints.

To conclude, the extraction of author name metadata yielded valuable data for 5791 unique files. This figure would be much higher if:

staff lists were available dating back to 1998 – the oldest file creation date is 1998 and there are 2567 files which were created between 1998 and 2006

authors used a formal metadata standard – there are 4384 files with an invalid or unknown author's name

There was a method of identify each staff member uniquely – there are 2287 names that could not be distinguished from staff with similar names and initials.

5.5 File types – residual data

The residual files includes files that have been identified as having an unknown file format and or unusable extensions. These files do not necessarily belong to the 113,697 files that do not contain sufficient metadata. A check of a sample of these file types showed that some of them could be classified if time permitted. These files could broadly be categorises into eight groups.

Compound files – these include compressed files and files associated with UNIX file systems.

Database files – these contain a combination of .mdb, .db and .csv files.

Executable files – these .exe files are executable program files that have either been created or downloaded from the internet. Their content is unknown and can be malicious in nature.

Registry files – again these files are unknown unless the content is analysed. They can be associated with software installed on the system by an individual, created by an individual for unknown reasons or downloaded. Modification of registry files may cause the data and or the system to become unstable.

Image files – these contain images and have various file extensions.

Videos – these files contain .mp4 and related file signature formats. Further searches should isolate these particular file types, either for classification or removal.

.HTM and HTML webpages – these files contain various web page information and messages.

Sound files – these files contain .mp3 and .wav file format signatures. Again further searches will isolate these particular files, either for classification or removal.

Some of these files have been further analysed and classified by type, depending upon their file signature and/or content. Table 11 illustrates the breakdown of these residual files.

Identified residual File Types	Total number of files
Compound files	743
Audio files	706
Database files	1447
Executable files	1624
Movie files	27
Image files	24,347
.HTM & HTML files	25,951
Total	54,845

Table 12: Breakdown of identified residual files by type

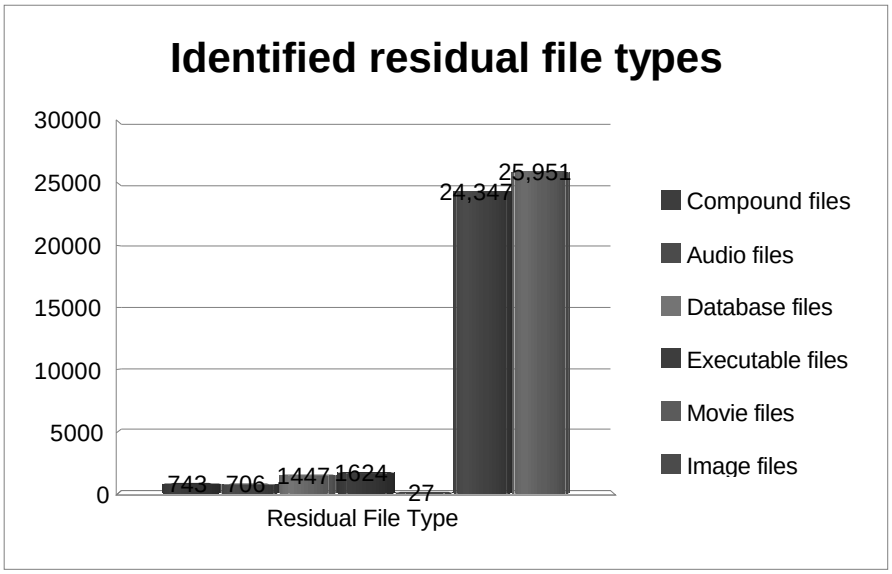


Figure 44: Identified residual file types – overview chart

Some types of residual file can be further analysed and classified for further action. There are 743 compound files and they have been bookmarked as 'archive files'. They can be further catalogued as follows:

Number	Type
3	.cab MS compressed
5	.gz UNIX GZIP
1	.rar
3	.tar UNIX tape archived
4	.uue UUEncode
727	.zip

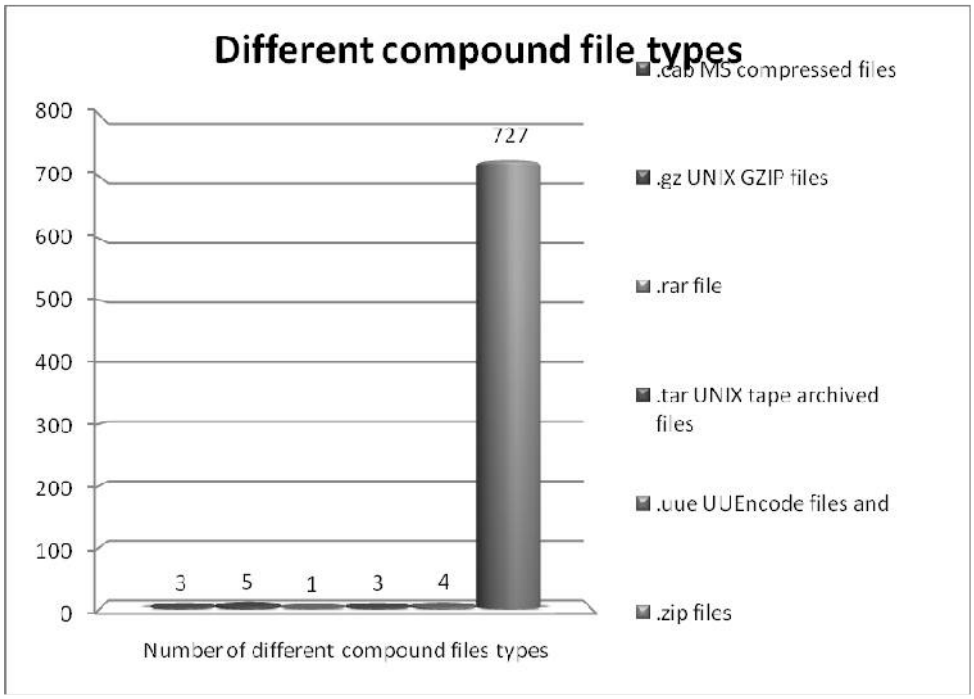


Figure 45: Compound file types

There are 24,923 picture files identified in the testbed. It is impossible to identify which may be personal and which are corporate without further investigation. For this reason, picture files are classified by file type and grouped together under the hierarchy structure as an individual group. Further manual processes could identify the purpose of these files.

There are a total of 706 audio files, which have been bookmarked as 'audio files'. These files can be further classified as:

Number	Type
91	.wav Waveform audio files
615	.mp3 MPEG audio files

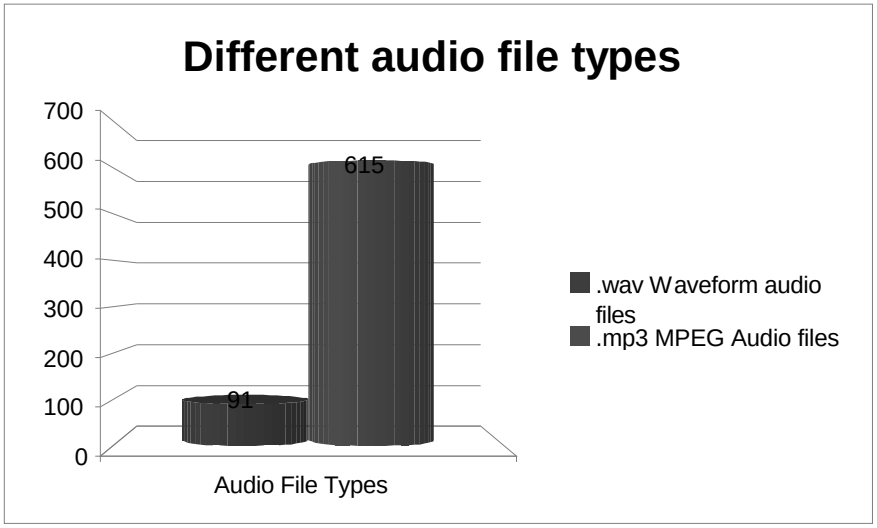


Figure 46: Audio file types

There are 1447 databases files have been bookmarked as 'database'. These files include a combination of .mdb, .db and .csv files. It was not possible to get a further breakdown of the number of files by each file extension due to time constraints and issues with the server and software installed within the WG.

There are a total of 1624 executable files, which have been bookmarked as 'executable files'. There are 447 .exe files and 1177.dll files. It is not known why these dynamic link library (.dll) files, which relate to the registry, are bookmarked within this folder. This requires further analysis.

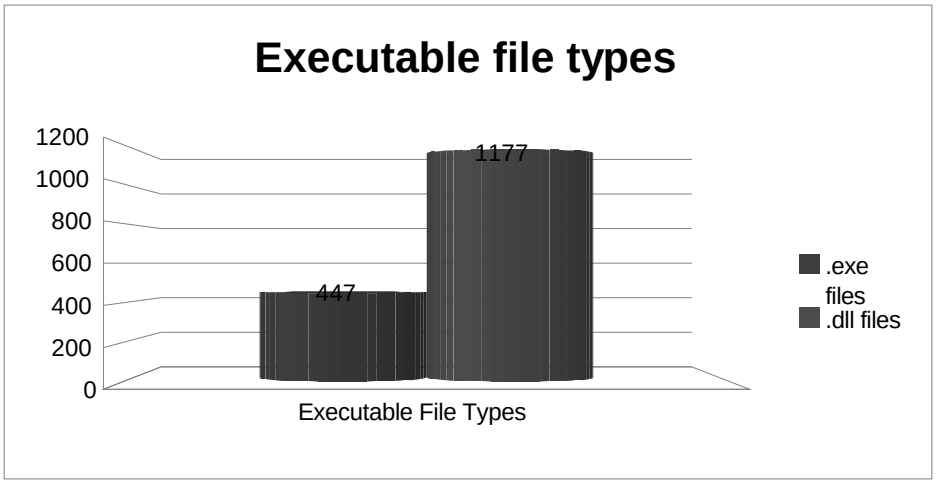


Figure 47: Executable file types

There are a total of 27 movie or video files, bookmarked within 'Movie'. They contain .asf files, .avi files, .mmm files, .mov files and .mpg files. Figure 48 contains a breakdown of the number of each type of movie file.

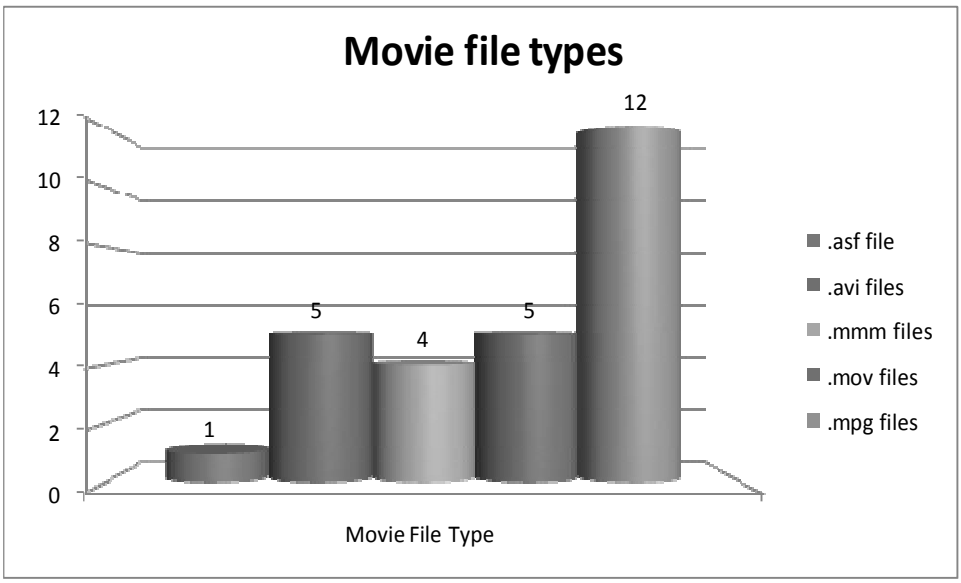


Figure 48: Movie file types

The final residual file type that has been categorised is .htm and .html web pages. There are 25,951 of these files bookmarked. A visual search of these files has identified that some of these files relate to e-mail messages that were not analysed due to their lack of metadata. These files have not been searched for classification because they don't contain metadata. However, further specific keyword combination searches could be conducted to attempt to classify these files.

In total, this means there are **54,845 files** that were identified by file type but not classified based on their metadata.

5.6 Final classification results

The earlier part of this section describes how files have been analysed based on their metadata and classified according to the WG's departmental structure. The creation of duplicates is inevitable in this process due to the lack of structured and complete metadata. After the metadata examination therefore, it is essential to ensure that remove any duplicate files that may have been created in the analysis of two sources of metadata: the 'company name' and the 'author name'. This was achieved by running a hash check on the created logical files.

The tables presented below show the results before and after the de-duplication of the logical files.

	Number of files	Number of unique files
DCELLS	25848	24791
Qualifications and Curriculum	53576	52518
Lifelong learning and skills	33595	32687
Higher learning group	5607	5083
Children and schools	3828	3605
Total	122454	**118684**

Table 13: Files in the Education department classified by division

Table 13 shows the classification of files originating from the Education department. The column 'Number of files' includes the duplicate files that were created during the classification. Through de-duplication it is possible to reduce this sets so that there are only unique files. This provides an accurate and efficient overview of the system. As this calculation shows, there were 3770 duplicates files tagged to the Education department created during the classification process.

Total number of files – Total number of unique files = created duplicates

122,454 – 118,684 = 3770

Figure 49 charts this data. The files listed under the DCELLS heading are education files, but the metadata contained in these files did not allow an attribution to a particular division within the department.

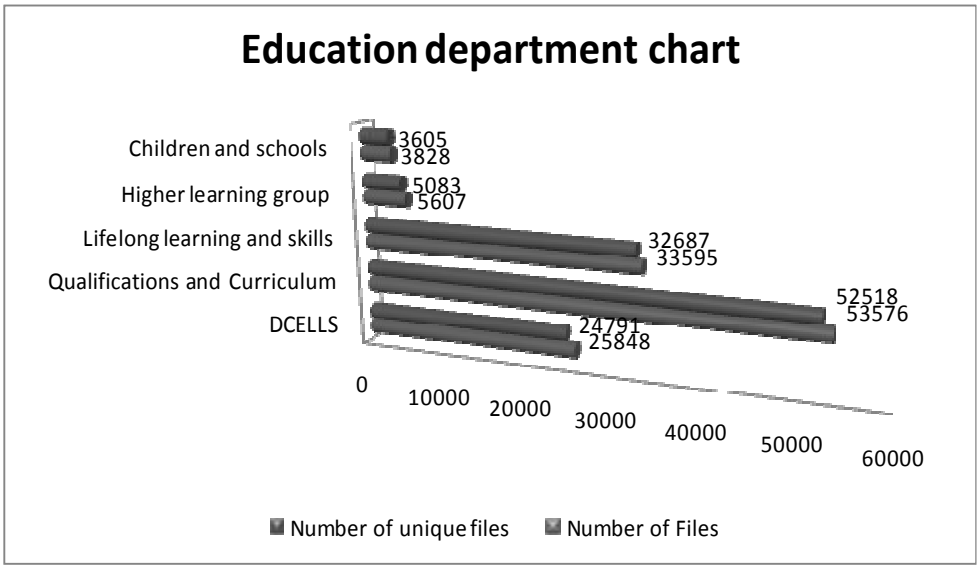


Figure 49: Education department files

Table 14 shows the effect of de-duplication after the initial classification for all files in the system.

	Number of files	Number of unique files
Multiple Departments	1475	673
Education	1475	118684
Economy & Transport	853	819
Envt, Sustainability & Housing	0	0
Finance	9363	8494
Health & Social Services	404	102
Public Services & Local Gov	591	395
IT	1530	1346
Legal Services Department	16	12
Author No Details Available	88614	47881
Constitutional Affairs	19821	19820
Consultant	2100	2099
Corporate Services	66	66
First Minister	4047	1138
Personal	13	9
Council	1726	1724
Total	132094	**203262**

Table 14: Files classified by department

Table 15 presents the classification of unique files in the hierarchy folder structure presented in 2.2.2.3. This is presented graphically in Figure 50.

Hierarchy Folder Structure		
	Departments	**Number of files**
1.	First Minister (including IT and Corporate Services)	2550
2.	Economy and transport	819
3.	Counsrl General and Leader of the house (Constitutional Affairs)	19820
4.	Social justice and Local government	12
5.	Finance Public Services and delivery	8889
6.	Health and Social Services	102
7.	Environment, sustainability and Housing	0
8.	Children Education And Lifelong Learning and Skills	118684
9.	Heritage	0
10.	Rural Affairs	0
11.	AUTHORS	47881
12.	COUNCIL	1724
13.	CONSULTANT	2099
14.	PERSONAL	9
15.	FILE TYPES (including Multiple Departments)	55518
	Total	**258107**

Table 15: Number of files in the hierarchy folder structure

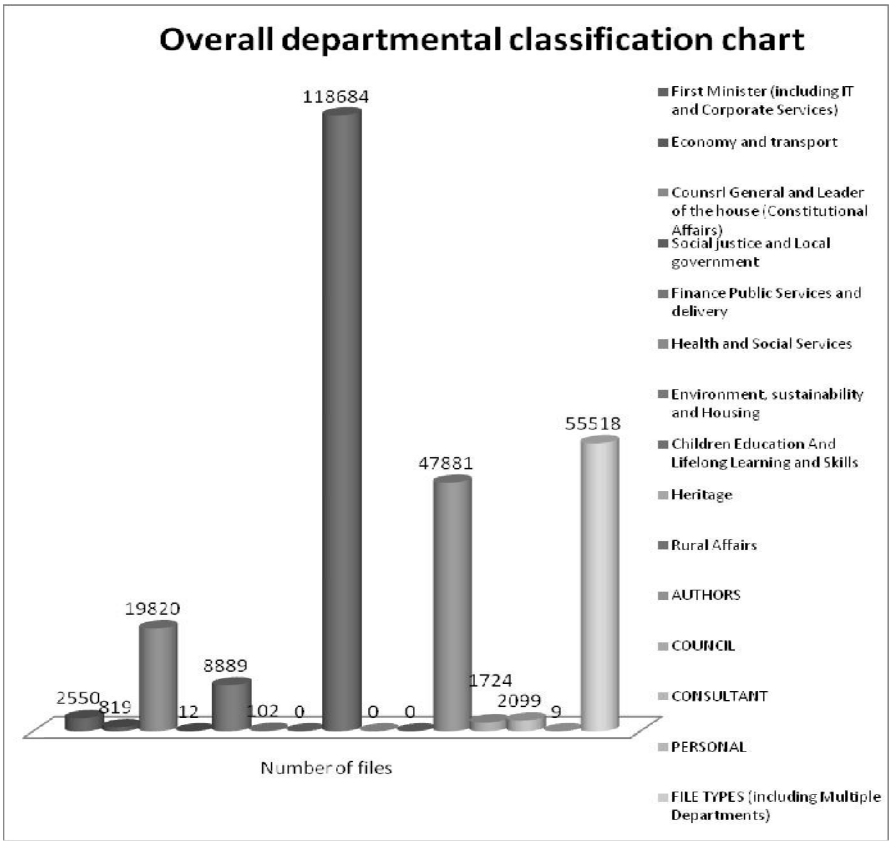


Figure 50: Overall departmental classification

In summary then, it proved possible to extract and classify !The Formula Not In Table files from the **354,233** unique files that contained metadata available for classification. In percentage terms, **65%** of the files with metadata could be classified on the basis of that metadata.

In total, there were **467,930** unique files in the testbed after the de-duplication, including files with no metadata. Again in percentage terms, it was possible to extract and classify **55%** of these testbed files. This leaves **210,496** residual files that will need further treatment or manual intervention based on their content. These residual files could be Microsoft Office files, e-mails as well as other files types. They could be further examined by using methods such as keyword lists. This might yield information that could be used to make an attribution to a department. This exercise could not be performed in the project due to time constraints.

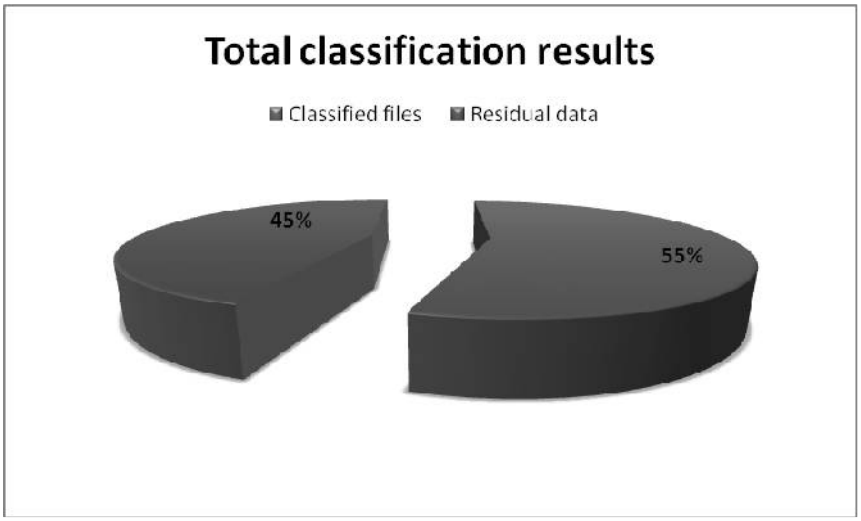


Figure 51: Summary of classification results

5.7 Extraction into iShare

A methodology has been developed in order to characterise each iShare category. This categorisation should enable the classified files to be integrated into the iShare active filing system used by the WG for classifying current data. The process will eradicate any duplication. It will allow the storage of files within the system to be organised efficiently and within a methodical structure.

Some files will need to be classified through keyword searches. This is because it is not possible to use metadata to classify files that have not been originally created for this type of system. However, a classification could be made based on the content of the files. This requirement can be achieved by performing keyword searches on the files. A successful data extraction of this type should ensure that there is significant evidence of the content or that the keywords are precise enough to allow classification.

Files need to be first extracted by categories based on departmental classification. This will enable us to:

- compare and verify the results obtained from the two different classification processes
- exclude files that have no usage.

This should allow the files to be categorised faster, and it should be a more accurate way of incorporating them into the iShare system. It should also avoid the duplication

of files that may contain more than one keyword. It is vital that duplicate files are not created for the successful integration of the files into the iShare system. This will also enable a calculation of the percentages of the departmental classification that will fit into the specific iShare categories.

The files to be extracted for the iShare file system will be identified using the different departmental categories that have been established following the classification based on the 'company' metadata. This means that when a particular file is stored in the hierarchical structure, it will be stored according to the department to which it is related. For example, if a file is created by a member of the Finance department, then the file when it is saved it will be classified within the Finance department. If the file is created by the Education department, then when it is saved it will be classified in the Education department, and so on.

5.7.1 iShare results

Searches have been conducted based on the keyword lists provided by the WG. They have been conducted against the 'DCELLS' and 'Author no details' data sets. An examination of these results suggests when keywords had more specific terminology (such as the finance keywords), the results are more accurate. This is to be expected, since the more specific a search term the more relevant the results. Other more general keywords produced results that were not particularly accurate or relevant. However, the results for the finance keyword searches suggest it may be possible to apply the retention policies that relate to finance documents.

In general, keyword searches will allow the WG to identify further documents to integrate into the iShare system, but further work is needed to make sure that keywords are not too generic and will generate accurate search results. Recommendations for compiling specific keyword search lists will be discussed within the conclusion of this report.

In order to illustrate how the keyword searches operate, a keyword search was conducted specifically for financial management files present within the DCELLS category. The results are presented in Table 16 and Figure 52. This shows that **7.7%** of files that were identified as financial management files, and they could potentially be included within the iShare system with the specific retention policies in place.

Keyword Search parameters	Total number of files
DCELLS total number of files	118684
Financial Management files identified	9179

Table 16: Results of keyword search

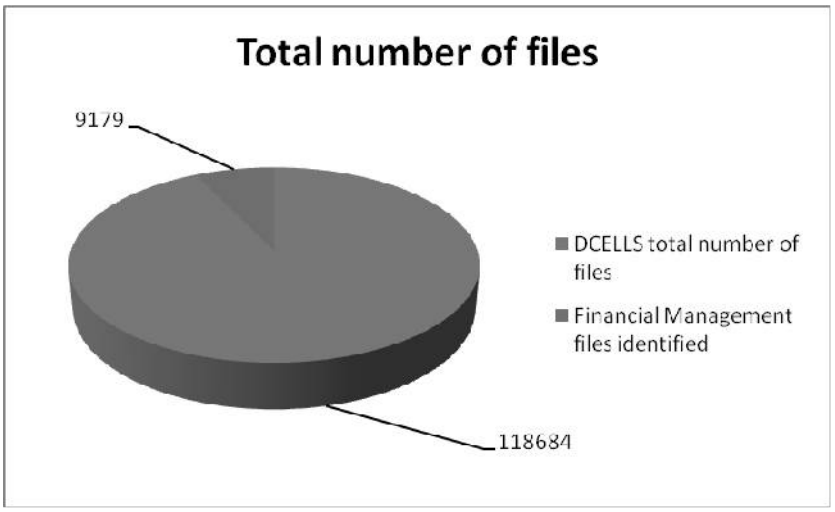


Figure 52: Keyword search within DCELLS category

Keyword combination searches could potentially be utilised to classify additional files, but more work would need to be carried out to ensure that specific rather than generic keywords are used when creating the lists.

5.8 Timescales and projections

The project allows some projections to be made for the application of this methodology to the whole estate and, in particular, the implications for:

network performance

data acquisition

hashing & de-duplication

indexing.

It is based on these parameters supplied by WG employees: there are 61 servers in the WG infrastructure and approximately 60TB of data. The network performance experiments were conducted in the UWN's computer laboratories and in a fully switched 100mbps environment using standard networking kit.

Networks are collections of independent computers that can communicate with one another over a shared medium using network protocols. Network protocols are standards that allow computers to communicate. A typical protocol defines how computers should identify one another on a network, the form that the data should take in transit, and how this information should be processed once it reaches its final destination. The TCP protocol was selected for examination as the eDiscovery Suite is using it for the communications of its different components.

One of the important aspects of a network protocol is its throughput performance. Often for the shake of simplicity, the interoperability and security aspects of performance are ignored. In our experiments we measured the throughput against block size of data. According to the first rule of network performance, throughput is largely dependent on CPU performance. Hence, the processing power of the computers running the eDiscovery components is important. The second rule says that block size is proportionate to the throughput, hence the larger the block size the better the throughput. However, the MTU for a standard TCP/IP network is only 1500 bytes, so the protocol has to fragment the data that the user wishes to send over the network.

The initial idea was to make peak and off peak tests between two clients and between a client and a server. In order to measure ‘efficiency’, we used a large sample of data. There were three different packet sizes in each test: 10, 40 and 80MB. These sizes are much larger than the requests of an average corporate user but well within the scope of requests generated by an investigator using the Examiner component of the eDiscovery Suite.

Figure 53 shows the TCP client-server performance. The network performance data is given in Appendix G.

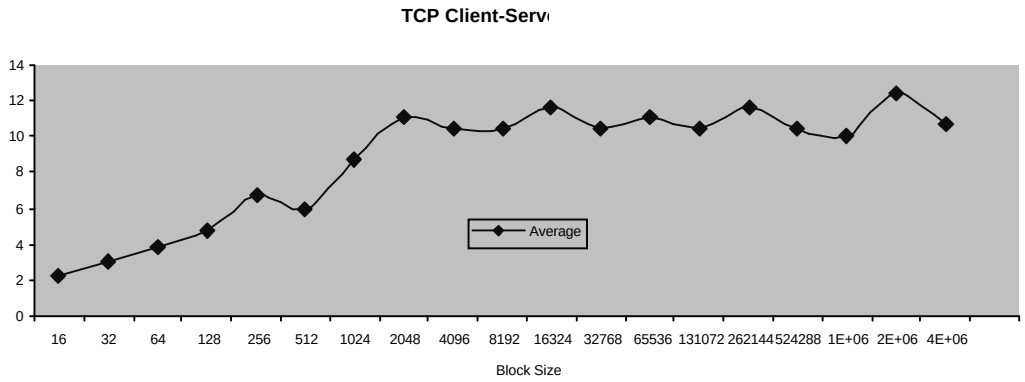


Figure 53: TCP client-server performance

TCP is a heavyweight network protocol. It can be seen that TCP performance is heavily dependent on the packet size. Once the packet size is greater than 1MB, is the performance improves drastically. The factors that reduce throughput for TCP are the time it takes to:

establish a connection and also close that connection once transmission has completed

resend packets that have not reached their destination

handle flow control

ensure packets are in sequence and unduplicated

check that packets actually reach their destination (the system waits for an acknowledgement).

The smaller the packet size, the greater the number of packets that have to be sent, hence the more time TCP has to spend in error checking.

Figure 54 shows the TCP client-client performance. The two ‘valleys’ seen in the graph can either be due to CPU load or network traffic generated by other users and applications.

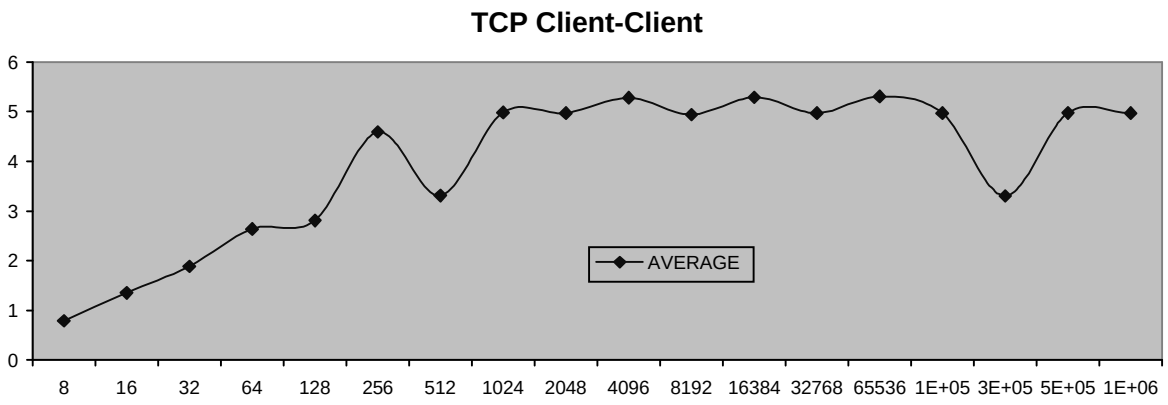


Figure 54: TCP client-client performance

The data acquisition took 24 hours and 3 minutes for 211.9GB. This is 150.371MB/min, which is within the lower range of the results from the network performance experiments. Projecting to the whole infrastructure of WG, it is estimated that a straightforward data acquisition through eDiscovery would take **290.5** days. However, if the maximum performance in the network experiments could be achieved – in other words, if we could to get maximum performance from TCP – then this time would fall to **60.6** days. Of course, even this second value is not feasible or practical. It is therefore recommended that the data set is fragmented and the operation is parallelised.

The de-duplication process took 5 hours for 211.9GB. Projecting to the whole of WG's infrastructure, it is estimated that a full de-duplication would take **60.4** days. Again this is not feasible, and it is recommended that we follow the acquisition fragmentation and further parallelise this operation.

The indexing process took 5 days for 149.4GB. Using the 35.88% duplication figure, this suggest that some 39395GB would have to be indexed. This would take **1318.451** days. Of course with the suggested fragmentation of the data set and the parallelisation of the operation, this estimate would also come down.

6.0 Methodology assessment

This section evaluates the specific methodology used for the classification of the WG's ERMS for the Digital Continuity project. In general terms, the evaluation procedure should prove or disprove the initial project objective.

6.1 Software

For the successful application of the methodology, the software must be able to perform the necessary functions. Several forensic tools and software packages were evaluated before choosing the most appropriate one.

The software tool needs to have specific attributes that would allow the classification a large data set in a remote network environment. The chosen software, the Guidance EnCase eDiscovery package, is not designed for these specific purposes. However, despite some difficulties (see Section 7.1), it proved to be a viable tool with the capabilities to assist the purpose of the project.

6.2 Methodology

The methodology is outlined in Section 2 of this report. The weaknesses of the provided system had to be studied before an appropriate method could be applied to classify and archive the data. The first weakness of the system was the existence of duplicate files. The second was the lack of any structured archiving of the files. The methodology is designed to tackle these two issues by providing first a procedure for de-duplication and then for classification.

6.2.1 De-duplication

Section 5.1.1 set out the reasons why there are duplicate files in the shared drives. Digital forensics can overcome this issue through the application of cryptographic hash functions. The compare-by-hash technique (see Section 2.2.2.2) can identify duplicate files. EnCase eDiscovery identifies duplicate data and performs dynamic de-duplication of data when it is queried. It does not eliminate duplicate files, and the duplicates are preserved in the system until their manual deletion. This allows WG staff to assess the results of the de-duplication process.

7.0 Conclusions

7.1 Problems encountered

The project encountered several problems, from the design to the experiment execution phase. These caused unforeseeable issues in meeting the proposed targets within the specified timeframe.

The issues that have affected the project encompass three specific areas of execution. They can be defined as follows:

- issues with the software required for executing the project successfully
- problems with the set-up and implementation of the testbed
- unforeseeable disruptions and limitations caused by the network performance.

These issues are described and analysed below.

7.1.1 Software

The software used for this project severely hindered the smooth progress of the project from inception to delivery. There were two major issues:

- the SAFE dongle failed twice
- the software was appropriate but it was not designed specifically for the requirements of this project.

Guidance Software employs a dongle authorisation system for using its software. If the dongle does not authenticate the licence correctly, access to the required software is restricted or denied completely. On two separate occasions, the SAFE dongle failed to authenticate. On the first occasion, a software malfunction on the server resulted in a halt of the dongle. The reason for this disruption had to be investigated. It cost the project six working days. On the second occasion, the dongle licence expired. It was initially planned for the dongle to expire in the middle of February. Because of the several challenges for the project, an extension to the license was required. The renewal procedure took two working days. In total, the failure of the dongle resulted in the project losing eight working days before the issues could be rectified.

In regards to the capabilities of the software, EnCase eDiscovery has the ability to do what was required for the project. However, it is not specifically designed for this

purpose. As a result, it was necessary to use several different components of eDiscovery. Proceeding in this manner demonstrated that the procedure cannot be completely automated in the future (as stated in the initial proposal) without any specialist interaction among the different nodes of eDiscovery. This limitation is a concern as the process cannot be conducted by any person periodically running an automated script, but rather it requires a specialist to conduct the data searching through manual interaction and supervision. Issues of automation and transparency could be addressed in a future version of eDiscovery.

7.1.2 Testbed

The testbed that was initially proposed for the project did not coincide with the layout of the testbed that was actually used. The WG employs a different system than the one described within the project proposal. This meant the methodology had to be adjusted so that project requirements were satisfied whilst adhering to the layout of the WG system set-up (which, of course, could not be adjusted).

Further disruptions occurred as a result of the state of the data. Several days were spent studying the nature of the data, the amount of metadata that was held in the files and deciding on the appropriate measures for each of the different data types.

The data provided by the WG about the file structure and the metadata that the files contained was not sufficiently precise. At the start of the project it was thought that the metadata would be rich enough to assist the classification of the records according to The National Archive's metadata standard. However, this prerequisite was not met in the testbed. Most files were created in FAT32 file systems that do not support the storage of rich metadata. Furthermore, descriptive metadata was not automatically or manually entered into the majority of the files by WG employees when the files were created.

Since the file structure data was not accurate or reliable, the data and the metadata had be manually scrutinised. This was a hugely time-consuming exercise, taking into consideration the sheer volume of data that needed analysis and categorisation.

Towards the end of the project there was not sufficient memory to load the case and initiate the keyword searches for further analysing the residual data. The low specification of the testbed had an impact on the human resource utilisation and resulted in unnecessary delays.

7.1.3 Network performance

The performance of the network and the physical layout of the network structure caused serious disruptions and issues that impeded the project's progress. These were the key issues:

winter weather disruption

physical network set-up

The severe weather disruption caused by the continual heavy snowfall and subsequent icy conditions caused not only issues with getting to the WG premises but, more importantly, caused the WG's local and remote servers to fail. This was further exacerbated by the fact that the network set-up consisted of remote desktops (thin clients) linked to the physical servers in a separate location.

At the start of the project, there were some issues concerning the architecture of the closed and independent network that was required for the implementation. Due to the lack of the required network set-up, the virtual LAN and the network domain had to be created at the same time with the eDiscovery software and after the researchers had accessed the WG facilities for conducting the first phase of the project.

Furthermore, the WG places user access limitations on physical access to the servers. This meant that technical assistance was required to tackle any problems, from fault finding to simple file copies and rectifications. Even though the network engineers were fully co-operative, this caused significant delays throughout the project's lifecycle.

The effect of these delays resulted in interruptions to other normally less time-critical tasks, and together with the remote desktop layout, this resulted in further delays. For example, the complete data indexing took almost three weeks to complete, which meant that the analysis phase was hugely belated. The indexing of data also required additional hard disk space, and resources needed to be found to increase the hard disk capacity. This cost the project another working day. The extended hard disk space was required for indexing the testbed data.

The other major issue within the project was the limitations imposed by using approved WG software. The web browser needed for successful metadata analysis was Microsoft Internet Explorer 7 (IE7). However, the WG security policy allows only Internet Explorer 6 to be installed on the systems. Native access to IE7 was required for ECC Web Server to allow data tagging and since the WG had not conducted full

testing on IE7, this web browser could not be installed at the beginning of the project. This imposed serious time constraints that increased the amount of resource needed to complete the project successfully within the specified timeframe. Eventually, IE7 was installed on the server as the technical department agreed that it was essential for the successful outcome of the project.

7.2 Lessons learned

Several issues had to be resolved during the project. As above, there have been some technical (and non-technical) issues that affected the servers and caused disruption.

The virtualisation of the e-Discovery components was problematic as virtualising within a virtual environment caused instabilities to the majority of the eDiscovery components.

Legacy data types created in FAT32 systems do not hold rich metadata. This means that a simple e-discovery process could not produce metadata that meets The National Archives standards. The retrieved metadata was not sufficient to answer all classification queries. Interviews had to be performed in order to collect additional primary data about the current practice of classifying records in WG.

Having an isolated network and dedicated hardware resources is important. Without these resources, the performance of the eDiscovery Suite components suffers, particularly affecting the acquisition, hashing and indexing operations.

Towards the end of the project there was not sufficient memory to load the case and initiate the keyword searches for further analysis of the residual data. It is imperative that state-of-the-art computers with adequate processing power and memory capacity are used to host all -the different eDiscovery Suite components.

7.3 Recommendations and conclusions

Recommendations for a full scale roll-out of the operations described in this report cover:

- general issues
- physical and logical architecture of the analysis infrastructure
- software applications for the classification

classification methodology.

The WG ICT infrastructure is governed by clear policies addressing relevant legislation regarding information and data. Given the problems that were encountered during the project, we recommend that the future classification operations be given ‘root’ access. The very nature of the classification operations goes against the existing ICT policies, and the need to work around these polices caused many instabilities and delays. An appropriate environment (including infrastructure, policies and users) will have to be created. It is recommended that the WG’s ICT supplier should not participate on a support basis but as a full stakeholder with dedicated access to the classification operations resources. This will negate the issue of external parties (the investigators) lacking appropriate permission levels and/or appropriate security clearance. Furthermore, it will allow for the transparent fragmentation of the data hosted in WG’s ICT infrastructure and their seamless acquisition.

Having dedicated hardware resources (including networking resources) is imperative. To overcome the problems encountered during the project, we recommend the use of a secure ‘war-room type’ environment (call it the classification environment) with root access to the whole of WG’s ICT infrastructure and access to the private cloud running over high-performance computing (HPC) resources. All the human resources participating in the classification operations (network engineers, IT support personnel and investigators) should be based in this secure environment. A means of easily isolating this classification environment from the rest of the infrastructure should be considered.

For efficiency, we recommend the fragmentation of the data during the acquisition and hashing operations. The de-duplication operations will not be affected by this fragmentation. Several servers (call them eDiscovery servers) with appropriate computing power and memory capacity should be based in the classification environment. These servers will be used for running the software applications required for the analysis and classification of data as well as for temporarily storing the data under examination. After the successful classification of the data, the records will be exported to predefined data repositories in the normal WG ICT infrastructure and their logical evidence files will be deleted from the classification environment.

The eDiscovery servers should be connected to a number of computers running the Examiner modules. The Examiners can be virtualised so the host computers can run

a number of virtual Examiners according to the requirements of the classification operations.

In order to minimise the turnaround time of the operations, we recommend the use of HPC resources. Fujitsu is set to bring high-performance computing to Wales. It will provide a distributed grid in a five-year project costing up to £40 million. The grid will include over 1400 nodes that will be spread across more than eight sites, linked using Fujitsu's middleware technology SynfiniWay, which will deliver an aggregated performance of more than 190 petaflops.

Grid computing is a technology that enables people and machines to effectively capture, publish, share and manage resources. There are several types of grids but the main types are data grids, computational grids and knowledge grids. Data and computational grids are quite similar in that they are used to manage and analyse data. With technology increasing and developing at such a dramatic rate, average computers cannot cope with the amount of data or the calculations they are being asked to perform. To analyse a complicated set of data could take a standard computer a few days or even weeks. If a grid is used to perform the same analysis, it could take considerably less time because it would harness the computational power available on the grid, parallelise the load and allow the calculations to be performed with a small turnaround time.

Regarding the applications used for the analysis of the data, we recommend the eDiscovery Suite from Guidance Software and the Pingar API from Pingar. One issue we had with the residual files after the metadata extraction was our inability to properly classify them due to the lack of appropriate metadata. With the Pingar API package we could have managed the residual data including documents, webpages, e-mails or any kind of text for performing these operations.

Entity extraction – Pingar API has a suite of cutting-edge tools that turn documents into useful lists of entities including people's names, telephone numbers, organisations and department/division names. This feature can be used for automatically generating metadata about the residual files after the eDiscovery process.

Content analysis – Pingar API provides precision keyword extraction and document summarisation. This feature can be used for extracting knowledge about each file that will be used to complement the results of the eDiscovery process. This should simplify the classification of the records.

Appendix A: Initial project plan

The project has a 24 week timeframe and 54 man-weeks are required according to the proposed project plan. The suggested start date is 4 October 2010 and the suggested end date is 1 April. The following table contains the details of the project work-packages, suggested start and end dates and project deliverables. The work-packages are ordered according to their number. Please note that certain work-packages overlap and that WP1 and WP5 start in week 0, and that there is a two week gap between 25 December 2010 and 10 January 2011.

Work-package No[1]	Work-package title	Lead partner No[2]	Person-weeks[3]	Start week[4]	End week[5]	Deliverable No[6]
WP1	Classification Methodology	P1	18	0	24	D1.1, D1.2
WP2	Pilot Application Set-up	P2	2	0	8	D2.1, D2.2, D2.3
WP3	Operational Experiment	P1	32	5	19	D3.1, D3.2
WP4	Exploitation/Dissemination and Management	P1	2	0	19	D4.1, D4.2
	TOTAL		54			

Work-package List

[1]Workpackage number: WP 1 – WP n.
[2]Number of the contractor leading the work in this work-package.
[3] The total number of person-weeks allocated to each work-package.
[4] Relative start date for the work in the specific workpackages, week 0 marking the start of the project, and all other start dates being relative to this start date.
[5]Relative end date, week 0 marking the start of the project, and all ends dates being relative to this start date.
[6] Deliverable number: Number for the deliverable(s)/result(s) mentioned in the work-package: D1 - Dn.

The following table contains details on the project deliverables such as their nature and their suggested dissemination level, as well as their suggested delivery dates. The deliverables are ordered according to their delivery date.

Del No[7]	Deliverable title	Delivery date[8]	Nature [9]	Dissemination level[10]
D1.1	Draft Classification Methodology	4	R	CO
D2.1	Pilot application Set-up	4	D	CO
D2.2	Integration & Testing Report	4	R	CO
D2.3	APIs for Test-bed	8	P	PU
D3.1	Draft En-Scripts	12	P	CO
D3.2	Operational Experiment Results	19	D	CO
D4.1	Exploitation/Dissemination & Use Plan (draft)	19	R	PU
D4.2	Technology Implementation Plan (draft)	19	R	PU
D1.2	Final Report	24	R	CO

Deliverables List

[7] Deliverable numbers in order of delivery dates: D1 – Dn

[8] Month in which the deliverables will be available.Month 0 marking the start of the project, and all delivery dates being relative to this start date.

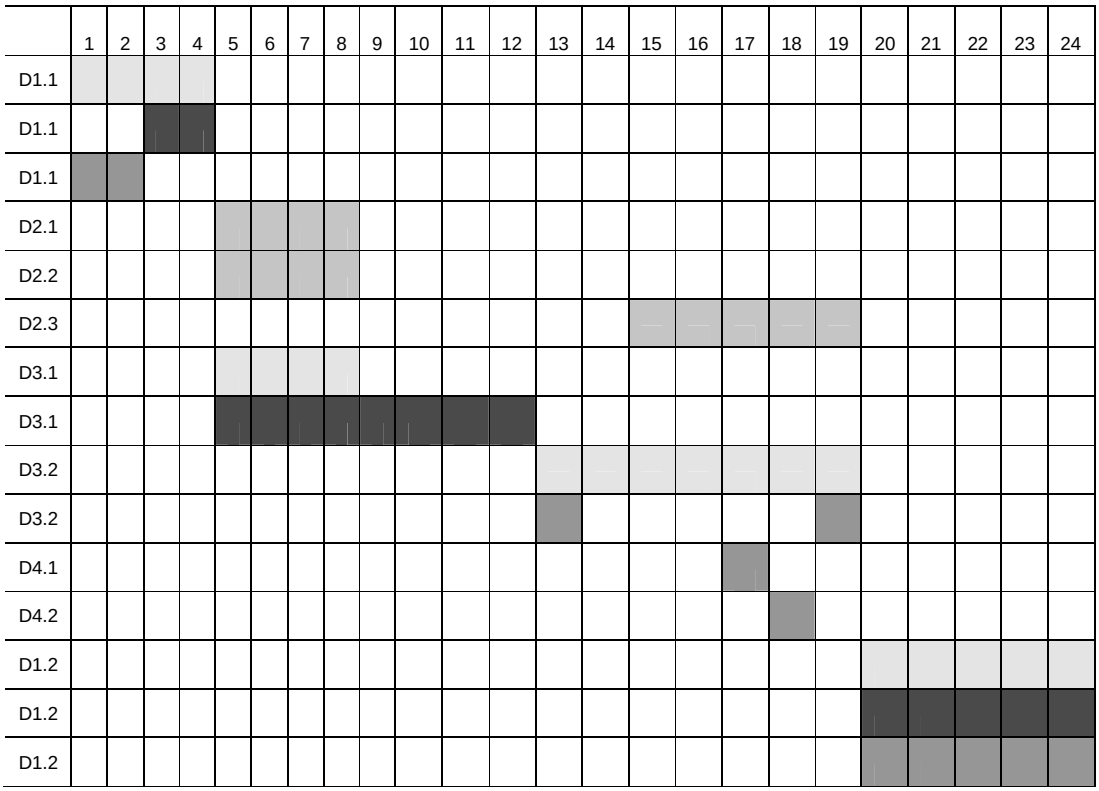
[9] The nature of the deliverable is indicated with one of the following codes:

R = Report
P = Prototype
D = Demonstrator
O = Other

[10] The dissemination level is indicated with one of the following codes:

PU = Public
RE = Restricted to a group specified by the consortium.
CO = Confidential, only for members of the consortium.

Appendix B: Resource utilisation



Appendix C: Work progress

Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
					1	2
3	4 Project Initialisation – Project plan meeting	5 Familiarisation with the project requirements	6	7 1st project progress meeting	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19 Security vetting received	20
21	22	23	24 1st day at the WG	25	26	27
28	29	30				

October 2010

November 2010

Working on WP1 Design draft report: create the Digital Continuity classification framework

Working on finalising the Digital Continuity classification framework

Setting up the servers and eDiscovery

Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
			1 Familiarisation with the test bed data Weaknesses identified (see end of document)	2	3 Data acquisition	4
5	6 Attempt to de-duplicate	7 Test bed De-duplication	8	9	10	11
12	13 Adaptation of the framework to the new requirements (see end of document)	14	15	16 EDiscovery dongle failed Decision to index files for full access on transcripts ○ Need for additional hard	17 Weather disruption	18
19	20 Weather disruption	21	22	23 Index running	24	25
26	27	28	29	30	31	1
2	3	4	5	6 Welsh/ English files separation script running	7 EDiscovery dongle failed	8
9	10	11	12	13	14	15
16	17 Metadata search to attempt classification on	18 Metadata search to attempt classification on	19 Treatment of results	20 Re-application of indexing	21	22
23	24	25 Treatment of the results	26 Attempt to run specific script that extracts all files	27 Treatment of the results Improvement of the script	28 Re-application of the script Gathering of the	29

December 2010

January 2011

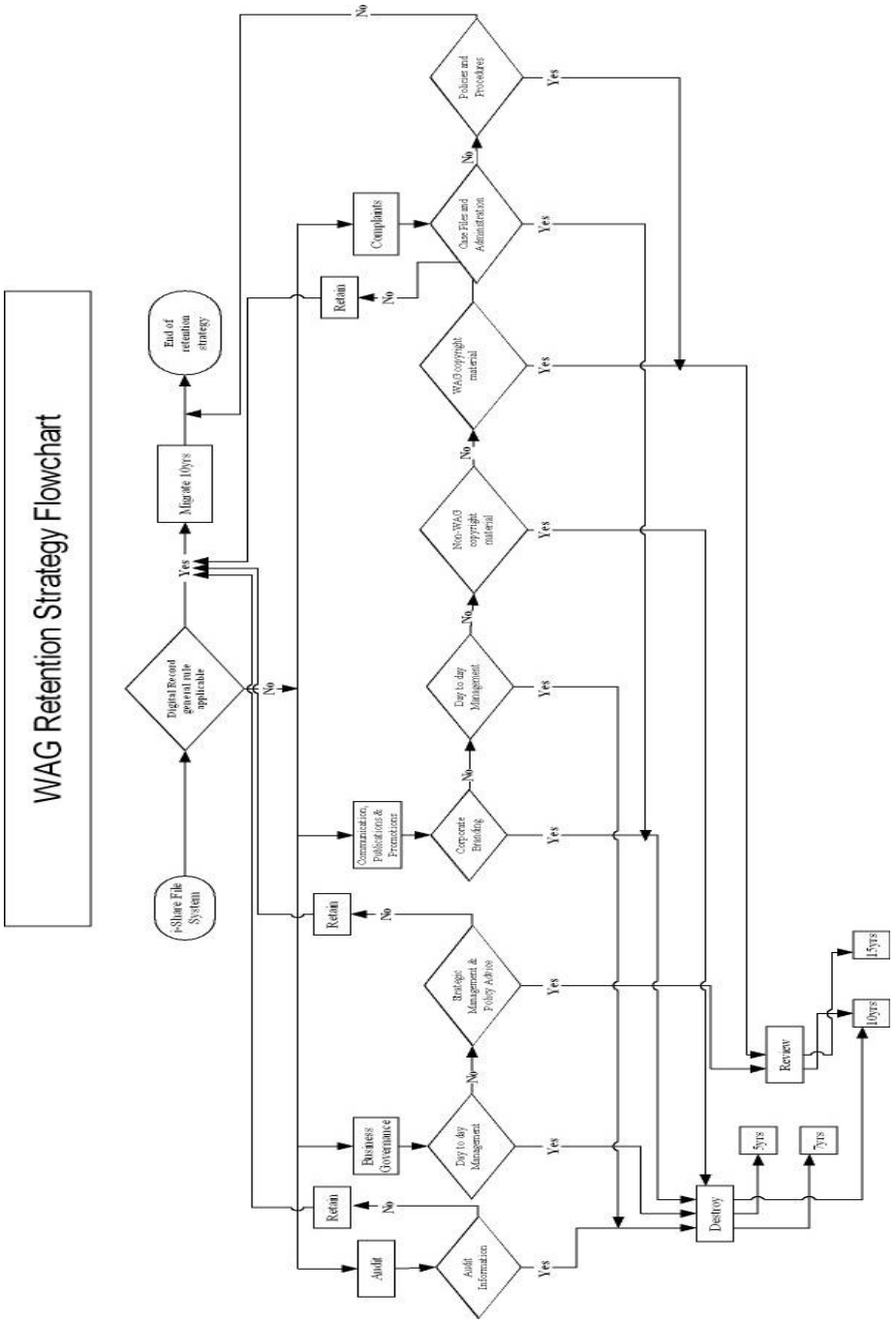
Attempts to recover – Communication with Guidance Software

30	31 File disposal script applied	1 Specific searches on metadata containing 'company' (department) information (see end of document) and specific author names	2	3	4	5
6	7	8	9	10	11	12
13	14	15 Initialisation of results' processing	16	17	18	19
20	21	22	23	24 Managing dongle and extraction issues	25	26
27	28	1 Categorising files by file type and examining the results	2	3 Preparing keywords for the iShare classification	4 Report writing preparation	5
6	7 Report writing meeting	8 Examining and finalising the classification results	9	10	11	12
13	14 Input to the final project report	15	16	17	18	19
20	21	22	23	24	25	26

February 2011

March 2011

Appendix D: Sample retention policy flowchart



Appendix E: Duplicates sample report

Name	File Type	Hash Value	Is Duplicate
E-mail - Ex RR (Accommodation Projects - Bedwas&Llanishen) 25.11.05.doc	Word Document	0000825ac5905e83afafcc358e0f66e6	No
fundraising-e.ppt	MS Powerpoint Template	0000aa88da34ad52454068c018a3f614	No
Pynciauargyfer 2010-11.doc	Word Document	0000bc852b2e5d94ba8008cd4f168264	No
BC003094.doc	Word Document	0000e8b199760bfe0512163355e4ea3e	No
Reply Slip (3).doc	Word Document	0000fdd3e72e5474379b86856b8e2ae0	Yes
Reply Slip.doc	Word Document	0000fdd3e72e5474379b86856b8e2ae0	Yes
Reply Slip (2).doc	Word Document	0000fdd3e72e5474379b86856b8e2ae0	No
Reply Slip.doc	Word Document	0000fdd3e72e5474379b86856b8e2ae0	Yes
20100326WBAC.XLS	MS Excel Spreadsheet	00013bcf0aea841bac1deb8d551102a3	No
020304arpWBC.doc	Word Document	00018321cb09282be7fc72d8457686e7	No
Kyrgystan.doc	Word Document	00019a00d9849b7058aee1f1e7cb1713	No
Kyrgystan.doc	Word Document	00019a00d9849b7058aee1f1e7cb1713	Yes
AMANDA RYAN.msg	MS Outlook Item	0001a7efda8f5f4e6010a11db47e82b9	No
Protocol - Local Social Partnerships Working v1 for SPs.doc	Word Document	0001b025e0b4b7c70326150a77763d7a	No
FW Probation Form Kate Allen (CP) 8.03.10.msg	MS Outlook Item	0001bbdc6000876608d7c0aca0809598	No
image9.gif	GIF	0001c62a68062473564256a0f8ec7f31	No

JVW conference intro_ English.ppt	MS Powerpoint Template	00020d7ca00d1874228cad39caedd87e	No
SFJD 0381-04 19th Nov.doc	Word Document	000216a7a19314e39e62c63aa74128de	No
SFJD 0381-04 19th Nov.doc	Word Document	000216a7a19314e39e62c63aa74128de	Yes
Cyfarfod i drafodgwerthusiadgwasanaethyrAthrawon Bro Cynnwys y fanyleb.rtf	Rich Text Format	0002520ce951407f5bc60daaa17d8b73	No
JC001296.doc	Word Document	00028ee828a75bfd04a87cb06e5b0f39	No
Mr Earle Letter - 3 August 2010.doc	Word Document	0002c69bfd1d006a8a75ad3ba609d502	No
091201rcg - Support Grant Letter w.doc	Word Document	00032254c2c4ae52d3d3deece1f0d293	No
Phil Rogers from GB, re grant offer letter, 27 March 08.doc	Word Document	00034b589603c9c05838a8edd0e5199d	Yes
Phil Rogers from GB, re grant offer letter, 27 March 08.doc	Word Document	00034b589603c9c05838a8edd0e5199d	No
BDD 48-06 Departmental Welsh Action Plan.msg	MS Outlook Item	000353336d8c7b4c5da279febec62d34	No
JISC(09)16 Annex A JISC Related Bodies.docm		000363a82653797fc93bda4cfa12cfa0	Yes
JISC(09)16 Annex A JISC Related Bodies.docm		000363a82653797fc93bda4cfa12cfa0	Yes
JISC(09)16 Annex A JISC Related Bodies.docm		000363a82653797fc93bda4cfa12cfa0	No
E-mail Attainment Units for new Welsh for Adults qualification.rtf	Rich Text Format	0003cca776537a711db29b069136bed2	No
Bilingual Budget plan round 03-07 M Jenkins 23sept03.doc	Word Document	0003cf1c38fb0d943c6ee05141af1a1c	Yes
Bilingual Budget plan round 03-07 M Jenkins 23sept03.doc	Word Document	0003cf1c38fb0d943c6ee05141af1a1c	No
ALG FE Regs 2007 Minute to CR 01 11 06doc.doc	Word Document	000404c6f8cdb5c742546b58f9e6811f	No
Skills for Health SQS feedback.doc	Word Document	00040ab9f5a87a7951da7cbad685fde2	No
Dash Training Annex A doc.doc	Word Document	00041b9734af02649120b7dd222395e8	No

YML Ystadegau staff - ColegGlannau Dyfrdwy.msg	MS Outlook Item	00041d606bacf7db4ef34e5c344c2804	No
D081114 CSkillsCrossBorder info for Bethan Milton.doc	Word Document	00044c18f6a4dbe8c2cbea4f8dca3a7c	No
Joanne Phillips.doc	Word Document	00048919be0fb5cce513cb897a315f15	Yes
Joanne Phillips.doc	Word Document	00048919be0fb5cce513cb897a315f15	No
Master W.doc	Word Document	0004b3f12b77fa124639241906d32d3f	No
W Exp PerfArts Inner.pdf	Adobe PDF	0004c89f55558af083988e08f2a683ba	No
W Exp PerfArts Inner.pdf	Adobe PDF	0004c89f55558af083988e08f2a683ba	Yes
Memo - Finance (T&S Claim 10.06.02).doc	Word Document	0004f516ed21b02e3059c0498a8690c8	No
Network Training Services Ltd - english.doc	Word Document	0004f741ae0a32eb6c546ac3321e5256	No
15 Jun 06 MD RE Submission on Functional Skills.htm	Web Page	00051dbd4f3670fbfcc8008c6de7e51e	No
ashley_contract06.doc	Word Document	000525b7233c5cc10bfc40386f9e0e5d	No
BC002761.doc	Word Document	00058edc82cde5c940bc99b7e0989132	No
NEW HEADS CONFERENCE 2003 - Programme W.doc	Word Document	0005b16b819637d7a9803cfdf47e9d86	Yes
NEW HEADS CONFERENCE 2003 - Programme W.doc	Word Document	0005b16b819637d7a9803cfdf47e9d86	No
path.gif	GIF	000636261f67c3e89c457131c2836e91	No
path.gif	GIF	000636261f67c3e89c457131c2836e91	Yes
path.gif	GIF	000636261f67c3e89c457131c2836e91	Yes
E-mail - John Jones (Serviced Office Accommodation).htm	Web Page	000642a519a8a1180ec024733ecaa81d	No
Y9strikingandfielding (2).doc	Word Document	0006776e41271ce9781132c3c63fd74a	Yes
Y9strikingandfielding (2).doc	Word Document	0006776e41271ce9781132c3c63fd74a	No

E-mail - Confirmation Authorisation to Award.rtf	Rich Text Format	00068a5db7260a8ed252bf21d0f5a69a	Yes
E-mail - Confirmation Authorisation to Award.rtf	Rich Text Format	00068a5db7260a8ed252bf21d0f5a69a	No
050711 Estyn performance frameworks - Martin Rolph response re Cabinet Sub Committee - Local Government and Public Services Monday 4th July 2005.msg	MS Outlook Item	000691e9ef9d9aec64c92854ec3301ad	No
0005 KS2 Ph4 DRAFT June27 for Sept05.doc	Word Document	0006a6042c4473642f7e88895985971b	No
0005 KS2 Ph4 DRAFT June27 for Sept05.doc	Word Document	0006a6042c4473642f7e88895985971b	Yes
3 KS2 DRAFT for Sept05.doc	Word Document	0006a6042c4473642f7e88895985971b	Yes
3 KS2 DRAFT for Sept05.doc	Word Document	0006a6042c4473642f7e88895985971b	Yes
PL1.min to AlunHuws - ELL Policy Board - 3 November 03.doc	Word Document	0006aa03e77c830724a7c7b46eaa3075	Yes
PL1.min to AlunHuws - ELL Policy Board - 3 November 03.doc	Word Document	0006aa03e77c830724a7c7b46eaa3075	No
ltr.re.learnPlanReview_RodAshley.LD.25.7.06.doc	Word Document	000751bcc3450f11b4154d46585a847e	No
06-09-14 TL2020 Workforce CPD paper TPO'S & JA 050906.doc	Word Document	000780d70437b178056cecc9b2acd6dd	No
06-09-14 TL2020 Workforce CPD paper TPO'S & JA 050906.doc	Word Document	000780d70437b178056cecc9b2acd6dd	Yes
Annex C - KS3 Group Cardiff.doc	Word Document	00079ca2968133c7ec7ea5dbd9c38fdc	No
The Education (National Curriculum) (Foundation Stage) (Wales) Order 2008 (W).doc	Word Document	0007ad33327c45c01a559197e89968fa	No
PatsPPtoKSPDCo-ords_14Jan08_V2.ppt	MS Powerpoint Template	0007e2624bcf526d3f10488cf575dab6	No
07-10-08 FW Confidential - consolidated version of evidence plus versions in mark-up of responsibilities and leadership sections.msg	MS Outlook Item	000807b194bb894b69a103aa0e07ca22	Yes

07-10-08 FW Confidential - consolidated version of evidence plus versions in mark-up of responsibilities and leadership sections.msg	MS Outlook Item	000807b194bb894b69a103aa0e07ca22	No
FINAL AGENDA 2 june 2006.doc	Word Document	00081199cc9d022e7a8114eddeabb570	Yes
FINAL AGENDA 2 june 2006.doc	Word Document	00081199cc9d022e7a8114eddeabb570	No
ST000516.doc	Word Document	0008144c7066bdaa5d4e50e1e81e4bb7	No
Caerphilly Activity 5B.doc	Word Document	0008527851edd6674fdaef207392ddb4	Yes
Caerphilly Activity 5B.doc	Word Document	0008527851edd6674fdaef207392ddb4	No
Caerphilly Activity 5B.doc	Word Document	0008527851edd6674fdaef207392ddb4	Yes
Caerphilly Activity 5B.doc	Word Document	0008527851edd6674fdaef207392ddb4	Yes
RE Learning Development Questionnaire.msg	MS Outlook Item	0008638d1f1cb373887120b9f2fdc0da	Yes
RE Learning Development Questionnaire.msg	MS Outlook Item	0008638d1f1cb373887120b9f2fdc0da	No
Workshop 22 - Ewloe [081001].doc	Word Document	0008689ca3ce0f83d4ec1c6d343182e6	No
Workshop 22 - Ewloe [081001].doc	Word Document	0008689ca3ce0f83d4ec1c6d343182e6	Yes
FW SCHOOL BASED INSET DAYS.htm	Web Page	000891155c8c30601c456104b3573c58	No
RE Welsh Synthetic Voice Meeting.msg	MS Outlook Item	0008af4c1307d7fcb669be1be72cdefd	No
RE Welsh Synthetic Voice Meeting.msg	MS Outlook Item	0008af4c1307d7fcb669be1be72cdefd	Yes
Final CILT UK grant letter Mar06.doc	Word Document	0008d5cd7d49b7bf4cefbda72ce16ece	No
020314arpInstructions on Accommodation Centres1.doc	Word Document	0008db09c2a81f61bce039f886c1c1fd	No
les1_B.doc	Word Document	0009126f1794b231bba94bee91663ce2	Yes
les1_B.doc	Word Document	0009126f1794b231bba94bee91663ce2	Yes
les1_B.doc	Word Document	0009126f1794b231bba94bee91663ce2	Yes

les1_B.doc	Word Document	0009126f1794b231bba94bee91663ce2	No
Gifts and Hospitality proforma.February 03.doc	Word Document	00094209cb7c0dea5b6a41e7cdfdcf4d	No
Gifts and Hospitality proforma.February 03.doc	Word Document	00094209cb7c0dea5b6a41e7cdfdcf4d	Yes
2010-03-23 RAE - year end information response (8)(BSCU).msg	MS Outlook Item	00095c578f695b9d337c7e8e2302bd18	No
081113 Gwernyfed High School.msg	MS Outlook Item	000985f025fc66bb65aae189e8f06606	No
JW001965.doc	Word Document	00099522bab8a4d3a5597772b1b084b7	No
100303 MFL Profile (To Jan '10).xls	MS Excel Spreadsheet	0009c1b268ded3f8c3314ddfdaeba1c8	No
Environment leaflet 23.9.10 (E).pdf	Adobe PDF	0009d2c949deee467b6fa1c7be02456f	No
05pubschedule01.xls	MS Excel Spreadsheet	0009d972138934440bb4f2658d752739	No
RE Post reference number required.rtf	Rich Text Format	0009f272f5d7927de10b31bd3a95133a	No
RE Post reference number required.rtf	Rich Text Format	0009f272f5d7927de10b31bd3a95133a	Yes
KEF operplan.doc	Word Document	000a0d452daa0f2ab178f961326ca217	No
minutes from 6th July (2).doc	Word Document	000a163e5c632b9467a9f46ad7582d33	No
BAA113 Contribute to innovation in a business environment.docm		000a3cb9dbba53749b89b9ddf41a5898	No
Lighting a stage.doc	Word Document	000a3e1124f10c1986d0811a326498ff	No
Tredegar_comp.pdf	Adobe PDF	000a4ef18d523653da0010ef597e3f33	No
RE Leitch 9 meeting in March 2007 - GJ availability.htm	Web Page	000a6e2e97358293baf0b48b27be357c	No
8 jAN Essential Skills Wales Entry Level Credit Rating Proposals.msg	MS Outlook Item	000ac9c0dddb6883e4d4f149c371e519	No
go-top.png	Portable Networks Graphic	000ad1f4103fb59d33a5e8b90e82342b	No
YSGOLYWURN0608.JPG	JPEG	000adf4ac74e618c75ac019659b60a0c	No

B3 Deall siapiau.jpg	JPEG	000b1886dc8b0b8809101115fe3c2ca7	No
General feedback from Minister-FW MB-JH-0994-08 - Address at the Bassaleg School Awards Evening .msg	MS Outlook Item	000b3027444b453ab5ed41fab95dd616	Yes
General feedback from Minister-FW MB-JH-0994-08 - Address at the Bassaleg School Awards Evening .msg	MS Outlook Item	000b3027444b453ab5ed41fab95dd616	Yes
General feedback from Minister-FW MB-JH-0994-08 - Address at the Bassaleg School Awards Evening .msg	MS Outlook Item	000b3027444b453ab5ed41fab95dd616	No
13.ico	Windows Icon	000b35a8d564bcdd84f7b8386870362d	No
133.ico	Windows Icon	000b35a8d564bcdd84f7b8386870362d	Yes
73.ico	Windows Icon	000b35a8d564bcdd84f7b8386870362d	Yes
VMDAH40.DOC	Word Document	000b58de33dd52016cbd60c4f642fa00	No
VMDAH40.DOC	Word Document	000b58de33dd52016cbd60c4f642fa00	Yes
Gan MM re cyfarfod efo DWJ - Rhag 07.msg	MS Outlook Item	000b9f597126b15acc6100bfd3223ce0	No
Template Perfromancemanagament form march 07.doc	Word Document	000bc2b3830c294e079ba49cf32258ea	No
carms notes.doc	Word Document	000bcb0c992ab2006186b21121eecbae	No
060419rcg - FP-PB-06-07 - MINUTES of Project Board - APR 06.doc	Word Document	000bf9e16a7a40d9cd2fc338e8e9ea5d	No
PS001557.doc	Word Document	000c0b17adc596872e7a61e7170433de	No
080117 SfB AAG Paper 6.doc	Word Document	000cb329defba16dc2bdb049df04c9cc	No
Dylunio a Thechnoleg - cynradd.doc	Word Document	000cdd123759c28db93483feecd7736f	No
Doc 1 Summary of Engagement.doc	Word Document	000cf3dabe1548b2569cb8ff1886afce	No
VT000630.doc	Word Document	000cf9ff9972a37ded58a3bbe7790d86	No

Merged letter to candidates Eng.doc	Word Document	000d0df1a22a5f7ec52caba296f70d09	No
FW Education UK - Wales Brochures.msg	MS Outlook Item	000d1719c411047b1d5cc1d615c3a30d	No
Headship programme.rtf	Rich Text Format	000d1f145241b7fe0fea303306356daf	Yes
Headship programme.rtf	Rich Text Format	000d1f145241b7fe0fea303306356daf	No
Headship programme.rtf	Rich Text Format	000d1f145241b7fe0fea303306356daf	Yes
Headship programme.rtf	Rich Text Format	000d1f145241b7fe0fea303306356daf	Yes
T&D 02 Using and Interpreting Engineering Drawings and Documentation.pdf	Adobe PDF	000d25f083e6836fd7d9171e965039ae	No
071127 MFL at KS2 WG mtg 071203 Agenda.doc	Word Document	000d2827baf9e52b7264e5fb5c7d9779	No
Final Version Report 12 May 2009.doc	Word Document	000d5722983d1f3b6ceaafd473721969	No
08 Apr 08 RE Welsh Baccalaureate - minor spec amendments2.txt	Text	000d58f4559952ed9b3a02b0c6f591f2	No
04 Risk registerJan 08.doc	Word Document	000dc8c3a144a17a2462527c63a85694	No
Websites 3.doc	Word Document	000dd16069eb32012b7ac3f9a1226ca2	No
David Morgan - English.doc	Word Document	000ddb8884718e7fc65a86928412b3f3	No
David Morgan - English.doc	Word Document	000ddb8884718e7fc65a86928412b3f3	Yes
Definitive list of Rec Bodies.doc	Word Document	000e2debae6e07427343d6db74fc8e39	No
Salary and Payee List.xls	MS Excel Spreadsheet	000e3ebfed898082e45b3aaa7d4bab7a	No
e-safety plan sept 09 - march 10.doc	Word Document	000e5ae6a6422edf125c075f3a4b1fdd	No
e-safety plan sept 09 - march 10.doc	Word Document	000e5ae6a6422edf125c075f3a4b1fdd	Yes
08SFS060.msg	MS Outlook Item	000e68680d7993fd15ac0ca768a35a16	No
Journal Voucher 021006 (2).xls	MS Excel Spreadsheet	000e89177a9dff270a9ff57d62f2689e	No
Journal Voucher 021006.xls	MS Excel Spreadsheet	000e89177a9dff270a9ff57d62f2689e	Yes
Journal Voucher 021006.xls	MS Excel Spreadsheet	000e89177a9dff270a9ff57d62f2689e	Yes

Journal Voucher 021006 (3).xls	MS Excel Spreadsheet	000e89177a9dff270a9ff57d62f2689e	Yes
040630 Handling of IT Area Inspection REport.msg	MS Outlook Item	000e938196248442345471782324 35fc	No
12_Open_Writer.mp3	MPEG-1 Audio Layer 3	000eafe2c420c81c9f24233b937af39b	No
Gwernyfed.Form H Final Report.doc	Word Document	000ec277501d7cf738dfbe3187146957	No
Blank Filing Tabs Template.xls	MS Excel Spreadsheet	000edddf6b2704f9b05ffd67431f85f6	No
Curriculum 7-19 Branch Operational Plan and Targets 2010-2011 V3.doc	Word Document	000f0dc24ed911f6fd9012bf0acd8bb2	No
Note of the Learning Skills and Qualifications Sub group 26 September.doc	Word Document	000f5f9d476434946508846e993f7bb	No
standard letter to DCSF.doc	Word Document	000f8e59b0018efc85604378d8a9b7ca	No
Deborah Davies (English).doc	Word Document	000f93a77790ee6055c3d835fa54dd25	No
Deborah Davies (English).doc	Word Document	000f93a77790ee6055c3d835fa54dd25	Yes
07.06.27 2008-09 PID - draft v6.doc	Word Document	000f95b0d3c4338c89739dc6b262f5a8	No
PT000355.doc	Word Document	000fad0af2cf416892d8acf5072b09ec	No
RE Business Cards.msg	MS Outlook Item	000fceeaed0c2a7e06a5e3e290a3f179	No
fu21509.png	Portable Networks Graphic	000fd245e52190400b5142ea2aeb1ba5	No
BE025229.jpg	JPEG	000fd8136ffd17604feebbe06995cac7	No
SY - deliver Key Skills training.doc	Word Document	000fe7fdf95e45b40801b9eeaf854d4a	No
org.eclipse.core.runtime.nl_zh_3.4.0.v20081130043401.jar	Compressed Java Archive	00101318a3a37a05a37d055b7a45b17c	No
Assessemnet.doc	Word Document	00102ca70170067760bc96f3988b70ae	No
Assessemnet.doc	Word Document	00102ca70170067760bc96f3988b70ae	Yes
FW Further to David's list.txt	Text	0010c087cf71986c39ec8f5da3bf69fc	No
Contractor Passes for the Coffee Shop.rtf	Rich Text Format	0010dc1791f4bef6cea49b8a25157ed1	No

LIPDD may 08 .xls	MS Excel Spreadsheet	00110129c52abafe2a4795d6777d38e0	No
Actual Spend 2009 2010.msg	MS Outlook Item	00113e565ed5b0f3c8cc14e0f247e197	No
Enquiry 080206 Terms of Appointment.rtf	Rich Text Format	001203352f72195ee9fd1dd1bc076d4e	Yes
Enquiry 080206 Terms of Appointment.rtf	Rich Text Format	001203352f72195ee9fd1dd1bc076d4e	No
CQFW brief words.doc	Word Document	00122e1a440ae4cba88d03748fe6dfc8	No
17 Feb 06 FS consultation 8 March.htm	Web Page	001234d278bd395525a1112accec785b	No
Response DH letter Dr Sibani Roy July 2010 V2.doc	Word Document	0012b66f28cabcdd245b3e6728e8a2dd	No
Contract covering letter Jan 08.doc	Word Document	0012bd2caa67d4eac37de61a4745d329	No
Covering letter 19 June 09.doc	Word Document	0012e7a77e016d2908a22e22cd0b8605	No
AT-JH-05541-09 Response.xml	XML Document	00131d2a0b7028b14fd346af261eef9f	No
Sandy Mills.doc	Word Document	0013252aa5eca87b9e8efd848021f816	No
ALL CHANGE TO FINANCE SYSTEMS-April 1st .rtf	Rich Text Format	0013a7feda98f9698208d11805745874	No
ALL CHANGE TO FINANCE SYSTEMS-April 1st .rtf	Rich Text Format	0013a7feda98f9698208d11805745874	Yes
ALL CHANGE TO FINANCE SYSTEMS-April 1st .rtf	Rich Text Format	0013a7feda98f9698208d11805745874	Yes
ALL CHANGE TO FINANCE SYSTEMS-April 1st .rtf	Rich Text Format	0013a7feda98f9698208d11805745874	Yes
E1 Invitation.pdf	Adobe PDF	0013abd42c227a03c1f80df7c4dc2ed6	Yes
E1 Invitation.pdf	Adobe PDF	0013abd42c227a03c1f80df7c4dc2ed6	No
minutes of team meeting 12.02.09.doc	Word Document	0013b2c0e362c80f8924f1f8bcfcd8c6	No
tasganllenyddol 3 (3).doc	Word Document	0013bd26513af7d194c836adddf9d4cd	No
070619 Briefing on inspection outcomes tables in Estyn Annual Report publication of inspection outcomes.doc	Word Document	0013ec74dc6cff364b2b4cf2e6e654ad	No

FW JD0105106 Dcleared reply.msg	MS Outlook Item	001404b7f2e8a36fc2db97d5665504b6	No
Tesco Computers for Schools Presentation Speech 29 Sep 03.doc	Word Document	0014126e4a1bf0d13eef2200b62f91a9	No
further contribution re contentious issues.rtf	Rich Text Format	001422e6dbd9ecbe87357eae21611efd	No
Edexcel Partnership - Project Board minutes		001444ae673ef64ee1e72422ff11776c	No
Cover Letter Llanidloes.doc	Word Document	001445dfc140a1600aac1154a69639eb	No
W_X_Bar2.png	Portable Networks Graphic	001474f4fafdb4910e2f8aa6a587074c	Yes
W_X_Bar2.png	Portable Networks Graphic	001474f4fafdb4910e2f8aa6a587074c	No
W_X_Bar2.png	Portable Networks Graphic	001474f4fafdb4910e2f8aa6a587074c	Yes
W_X_Bar2.png	Portable Networks Graphic	001474f4fafdb4910e2f8aa6a587074c	Yes
Successful Letter - Tenby Infants school - Julie Hurlow - Dwyieithog - (23-11-09).doc	Word Document	00149947b6258bfb787355b86eeab711	No
BC002879.doc	Word Document	00149a2d69b3caa0eb128ce078e7fc25	No
Aug 07 Student Support.doc	Word Document	00149a463c6b7b35f470c9040cef47b1	No
Cit skills landscape.doc	Word Document	0014b5d8f3ba0c51fd59518a2cd839cc	No
PPR Evaluation Form - BTS 09.06.doc	Word Document	0014cd0b6358df1c06ca9bdcaefc54ea	No
DK000103.doc	Word Document	0014ec386fc9091f01b5f5e6ffc37d57	No
Bus3_B.doc	Word Document	00150994e5061b600ca0700054b66456	No
E-mailt o AJ xchangewales - Request for additional desks.htm	Web Page	001524dac4dfd28d041a02460de27061	No
Scheme v2 not tracked.doc	Word Document	00153e9ad77bd29977601c8980106b08	Yes

Appendix F: Metadata sample report

Name Of Document: Case_AfterDeduplication\WGdata (3)\172.33.5.196\TestSource\AT S\1. CONTACT DATABASES ALL BRANCHES\2008-09\080611 Directors of Education All Wales.xls
Company:
Title:
Subject:
Author:
Keywords:
Comments:
Last Saved By: sargentd
Template:
Version:
Revision:
Create Date: 03/Oct/2007 11:27:02AM
Last Revision Date: 07/Nov/2008 11:06:38AM
Last Print Date: 30/Jun/2008 10:08:43AM
Number of Pages: 0
Number of Characters: 0
Number of Paragraphs: 0
Number of Words: 0
Hash: 644230E6CDBF6E1A003137B604E389D0

Name Of Document: Case_AfterDeduplication\WGdata (3)\172.33.5.196\TestSource\AT S\1. CONTACT DATABASES ALL BRANCHES\2008-09\PI Branch South Wales AT Contact List Mail Merge Table Main.xls
Company: NCETW
Title:
Subject:
Author: linda.wilkes
Keywords:
Comments:
Last Saved By: williamsd16
Template:
Version: Microsoft Excel
Revision:
Create Date: 07/Jun/2006 08:59:35AM
Last Revision Date: 18/May/2009 04:10:05PM
Last Print Date: 11/Feb/2009 12:25:34PM
Number of Pages: 0
Number of Characters: 0
Number of Paragraphs: 0
Number of Words: 0
Hash: F81CF779D9A146A2ECF1A100728AE9CE

Name Of Document: Case_AfterDeduplication\WGdata (3)\172.33.5.196\TestSource\AT S\1. CONTACT DATABASES ALL BRANCHES\2009-10\080611 Directors of Education All Wales.xls
Company:
Title:
Subject:
Author:
Keywords:
Comments:
Last Saved By: kinga1
Template:
Version:
Revision:
Create Date: 03/Oct/2007 11:27:02AM
Last Revision Date: 17/Aug/2009 12:53:21PM
Last Print Date: 30/Jun/2008 10:08:43AM
Number of Pages: 0
Number of Characters: 0
Number of Paragraphs: 0
Number of Words: 0
Hash: 310CA4E2C795454DDE1E7BEA69880F35

Name Of Document: Case_AfterDeduplication\WGdata (3)\172.33.5.196\TestSource\AT S\1. CONTACT DATABASES ALL BRANCHES\2009-10\Network and Implementation contacts for Stakeholder Questionnaire.xls
Company: Welsh Assembly Government
Title:
Subject:
Author: BrowningL1
Keywords:
Comments:
Last Saved By: BrowningL1
Template:
Version: Microsoft Excel
Revision:
Create Date: 23/Jul/2009 09:31:08AM
Last Revision Date: 23/Jul/2009 03:19:32PM
Last Print Date:
Number of Pages: 0
Number of Characters: 0
Number of Paragraphs: 0
Number of Words: 0
Hash: DA55D23559CB1D1A65938D21C6683487

Name Of Document: Case_AfterDeduplication\WGdata (3)\172.33.5.196\TestSource\AT S\1. CONTACT DATABASES ALL BRANCHES\2009-10\PI Branch South Wales AT Contact List Mail Merge Table Main.xls
Company: NCETW
Title:
Subject:
Author: linda.wilkes
Keywords:
Comments:
Last Saved By: williamsd16
Template:
Version: Microsoft Excel
Revision:
Create Date: 07/Jun/2006 08:59:35AM
Last Revision Date: 18/May/2009 04:10:05PM
Last Print Date: 11/Feb/2009 12:25:34PM
Number of Pages: 0
Number of Characters: 0
Number of Paragraphs: 0
Number of Words: 0
Hash: F0D3EB08A54DD6BA987EFF95358949FF

Name Of Document: Case_AfterDeduplication\WGdata (3)\172.33.5.196\TestSource\AT S\10. HR\2008-09\081120 ATSW U Access upload update from T&D 20 Nov 08.xls
Company:
Title:
Subject:
Author: Crystal Decisions
Keywords:
Comments: Powered by Crystal
Last Saved By: jane.leek
Template:
Version:
Revision:
Create Date: 20/Nov/2008 09:53:06AM
Last Revision Date: 21/Nov/2008 03:50:29PM
Last Print Date: 21/Nov/2008 03:48:59PM
Number of Pages: 0
Number of Characters: 0
Number of Paragraphs: 0
Number of Words: 0
Hash: CF49F2B0D520A0D9C70E0B183133BAEA

Name Of Document: Case_AfterDeduplication\WGdata (3)\172.33.5.196\TestSource\AT S\10. HR\2008-09\Annual Leave form.doc
Company: National Assembly for Wales
Title: Application for Leave of Absence
Subject:
Author: Debra Wong
Keywords:
Comments:
Last Saved By: leekj
Template: Normal
Version: Microsoft Office Word
Revision: 2

Create Date: 20/Feb/2009 04:37:00PM
Last Revision Date: 20/Feb/2009 04:37:00PM
Last Print Date: 22/Jul/2008 12:11:00PM
Number of Pages: 2
Number of Characters: 1359
Number of Paragraphs: 3
Number of Words: 238
Hash: 16187B5D82F8FC800E57649354D3F46B

Name Of Document: Case_AfterDeduplication\WGdata (3)\172.33.5.196\TestSource\AT S\10. HR\2008-09\AT South 08-09 PMP as at 17 Sept 2008.xls
Company: National Assembly for Wales
Title:
Subject:
Author: thomasc12
Keywords:
Comments:
Last Saved By: jane.leek
Template:
Version: Microsoft Excel
Revision:
Create Date: 18/Sep/2008 01:30:21PM
Last Revision Date: 23/Sep/2008 01:35:34PM
Last Print Date:
Number of Pages: 0
Number of Characters: 0
Number of Paragraphs: 0
Number of Words: 0
Hash: A70FA43A229347E75CC4AFC9F5345815

Name Of Document: Case_AfterDeduplication\WGdata (3)\172.33.5.196\TestSource\AT S\10. HR\2008-09\Minute to HoDs re trigger point reports Feb 09.doc
Company: Welsh Assembly Government
Title: To:
Subject:
Author: olderg
Keywords:
Comments:
Last Saved By: rossh
Template: Normal
Version: Microsoft Office Word
Revision: 2
Create Date: 17/Feb/2009 10:03:00AM
Last Revision Date: 17/Feb/2009 10:03:00AM
Last Print Date:
Number of Pages: 1
Number of Characters: 2269
Number of Paragraphs: 5
Number of Words: 397
Hash: 94818058E33BD708D2C9B372525752A4

Name Of Document: Case_AfterDeduplication\WGdata (3)\172.33.5.196\TestSource\AT S\10. HR\2008-09\U-Access -Good Practice (HR).doc
Company: NCETW
Title: Good Practice ?updating U-access for the Business Directory
Subject:
Author: alex.king
Keywords:
Comments:
Last Saved By: alex.king
Template: Normal.dot
Version: Microsoft Office Word
Revision: 1
Create Date: 14/Nov/2008 10:27:00AM
Last Revision Date: 14/Nov/2008 10:34:00AM
Last Print Date:
Number of Pages: 1
Number of Characters: 4918
Number of Paragraphs: 11
Number of Words: 862
Hash: B59F3FD7C241C2A5E674665B28AC2FE8

Name Of Document: Case_AfterDeduplication\WGdata (3)\172.33.5.196\TestSource\AT S\10. HR\2009-10\AT South 08-09 PMP as at 17 Sept 2008.xls
Company: National Assembly for Wales

Title:
Subject:
Author: thomasc12
Keywords:
Comments:
Last Saved By: jane.leek
Template:
Version: Microsoft Excel
Revision:
Create Date: 18/Sep/2008 01:30:21PM
Last Revision Date: 23/Sep/2008 01:35:34PM
Last Print Date:
Number of Pages: 0
Number of Characters: 0
Number of Paragraphs: 0
Number of Words: 0
Hash: F182FCA15A46493733E5E4DC65EAE763

Name Of Document: Case_AfterDeduplication\WGdata (3)\172.33.5.196\TestSource\AT S\2. AREA INFORMATION\2008-09\1 BRIDGEND\Area Statistics and Data\LA DATA\080806 LA Data Briefing.doc
Company: NCETW
Title: POPULATION
Subject:
Author: Robert.Joyce
Keywords:
Comments:
Last Saved By: Robert.Joyce
Template: Normal
Version: Microsoft Office Word
Revision: 2
Create Date: 06/Aug/2008 04:02:00PM
Last Revision Date: 06/Aug/2008 04:09:00PM
Last Print Date:
Number of Pages: 1
Number of Characters: 25884
Number of Paragraphs: 60
Number of Words: 4540
Hash: 0566555979EDF230F1537FF31715B692

Name Of Document: Case_AfterDeduplication\WGdata (3)\172.33.5.196\TestSource\AT S\2. AREA INFORMATION\2008-09\1 BRIDGEND\Area Statistics and Data\LA DATA\Bridgend Briefing.doc
Company: NCETW
Title: According to the latest available Mid Year Estimates, the total population of Bridgend has steadily increased in recent years and stood at 132,584 in 2006
Subject:
Author: Robert.Joyce
Keywords:
Comments:
Last Saved By: Robert.Joyce
Template: Normal
Version: Microsoft Office Word
Revision: 13
Create Date: 20/Aug/2008 11:00:00AM
Last Revision Date: 20/Aug/2008 05:18:00PM
Last Print Date:
Number of Pages: 1
Number of Characters: 3109
Number of Paragraphs: 7
Number of Words: 545
Hash: 70B7CD0A20CC09AC66F43722F6C775B9

Name Of Document: Case_AfterDeduplication\WGdata (3)\172.33.5.196\TestSource\AT S\2. AREA INFORMATION\2008-09\1 BRIDGEND\Area Statistics and Data\LA DATA\LA Data Briefing.doc
Company: NCETW
Title: Population - current and projected, Welsh speakers;
Subject:
Author: Robert.Joyce
Keywords:
Comments:
Last Saved By: Robert.Joyce
Template: Normal
Version: Microsoft Office Word
Revision: 11
Create Date: 06/Aug/2008 11:19:00AM

Last Revision Date: 06/Aug/2008 02:52:00PM
Last Print Date: 06/Aug/2008 02:11:00PM
Number of Pages: 1
Number of Characters: 25883
Number of Paragraphs: 60
Number of Words: 4540
Hash: 58A42434D246C82C81B7E81778DE4247

Name Of Document: Case_AfterDeduplication\WGdata (3)\172.33.5.196\TestSource\AT S\2. AREA INFORMATION\2008-09\1 BRIDGEND\Area Statistics and Data\LA DATA\WBL 0506.xls
Company: NCETW
Title:
Subject:
Author: Robert.Joyce
Keywords:
Comments:
Last Saved By: Robert.Joyce
Template:
Version: Microsoft Excel
Revision:
Create Date: 06/Aug/2008 02:30:16PM
Last Revision Date: 06/Aug/2008 02:53:50PM
Last Print Date:
Number of Pages: 0
Number of Characters: 0
Number of Paragraphs: 0
Number of Words: 0
Hash: 3C7EE231BA2BC8B0748F24F776E71C41

Name Of Document: Case_AfterDeduplication\WGdata (3)\172.33.5.196\TestSource\AT S\2. AREA INFORMATION\2008-09\1 BRIDGEND\Area Statistics and Data\MAPS\09 08 07 Word Version.doc
Company: Welsh Assembly Government
Title:
Subject:
Author: jenkinsp2
Keywords:
Comments:
Last Saved By: jenkinsp2
Template: Normal
Version: Microsoft Office Word
Revision: 1
Create Date: 07/Aug/2009 09:11:00AM
Last Revision Date: 07/Aug/2009 09:15:00AM
Last Print Date:
Number of Pages: 1
Number of Characters: 1
Number of Paragraphs: 1
Number of Words: 0
Hash: D8B66086794D117F02624BB09A04D0D7

Name Of Document: Case_AfterDeduplication\WGdata (3)\172.33.5.196\TestSource\AT S\2. AREA INFORMATION\2008-09\1 BRIDGEND\Bridgend\Schools\List of Schools -Bridgend.xls
Company: Welsh Assembly Government
Title:
Subject:
Author: kinga1
Keywords:
Comments:
Last Saved By: jenkinsp2
Template:
Version: Microsoft Excel
Revision:
Create Date: 30/Apr/2009 01:39:06PM
Last Revision Date: 07/Sep/2009 09:25:11AM
Last Print Date: 07/Sep/2009 09:25:06AM
Number of Pages: 0
Number of Characters: 0
Number of Paragraphs: 0
Number of Words: 0
Hash: 549C730382F03761741B2AA65CD95FD9

Name Of Document: Case_AfterDeduplication\WGdata (3)\172.33.5.196\TestSource\AT S\2. AREA INFORMATION\2008-09\1 BRIDGEND\Capital Project Funding\Assessment of Capital Funding Applications OBC Sept 2008.doc
Company:

Title: Institution Name:
Subject:
Author:
Keywords:
Comments:
Last Saved By: Brian.Foster
Template: Normal
Version: Microsoft Office Word
Revision: 2
Create Date: 30/Sep/2008 12:20:00PM
Last Revision Date: 30/Sep/2008 12:20:00PM
Last Print Date: 25/Apr/2008 09:26:00AM
Number of Pages: 1
Number of Characters: 1575
Number of Paragraphs: 3
Number of Words: 276
Hash: 47AA62C7FE355F591D0A831C0321111D

Name Of Document: Case_AfterDeduplication\WGdata (3)\172.33.5.196\TestSource\AT S\2. AREA INFORMATION\2008-09\1 BRIDGEND\Capital Project Funding\capital project funding apllications-final.xls
Company: NCETW
Title:
Subject:
Author: Nicola.Owen
Keywords:
Comments:
Last Saved By: Nicola.Owen
Template:
Version: Microsoft Excel
Revision:
Create Date: 09/Sep/2008 02:25:31PM
Last Revision Date: 22/Sep/2008 09:26:04AM
Last Print Date: 15/Sep/2008 09:31:09AM
Number of Pages: 0
Number of Characters: 0
Number of Paragraphs: 0
Number of Words: 0
Hash: 0F297987A8A130B4D2F1309A211BF914

Name Of Document: Case_AfterDeduplication\WGdata (3)\172.33.5.196\TestSource\AT S\2. AREA INFORMATION\2008-09\1 BRIDGEND\Capital Project Funding\CM Bridgend Bid.doc
Company:
Title: Institution Name:
Subject:
Author:
Keywords:
Comments:
Last Saved By: Brian.Foster
Template: Normal
Version: Microsoft Office Word
Revision: 2
Create Date: 30/Sep/2008 02:53:00PM
Last Revision Date: 30/Sep/2008 02:53:00PM
Last Print Date: 25/Apr/2008 09:26:00AM
Number of Pages: 1
Number of Characters: 1926
Number of Paragraphs: 4
Number of Words: 337
Hash: 8FFD972FA69B69510481FB97BC9CEE02

Name Of Document: Case_AfterDeduplication\WGdata (3)\172.33.5.196\TestSource\AT S\2. AREA INFORMATION\2008-09\1 BRIDGEND\Capital Project Funding\Rhymney College Bid.doc
Company:
Title: Institution Name:
Subject:
Author:
Keywords:
Comments:
Last Saved By: Brian.Foster
Template: Normal
Version: Microsoft Office Word
Revision: 2
Create Date: 30/Sep/2008 01:42:00PM
Last Revision Date: 30/Sep/2008 01:42:00PM
Last Print Date: 25/Apr/2008 09:26:00AM

Number of Pages: 1
Number of Characters: 2012
Number of Paragraphs: 4
Number of Words: 352
Hash: ABA4A3B45FCF179219516DC45247C519

Name Of Document: Case_AfterDeduplication\WGdata (3)\172.33.5.196\TestSource\AT S\2. AREA INFORMATION\2008-09\1 BRIDGEND\Capital Project Funding\SCIF Draft letter to FEIs 14-19 network reps v1.doc
Company: COI Communications
Title: Eichcyf
Subject:
Author: Margaret.Lomer
Keywords:
Comments:
Last Saved By: Brian.Foster
Template: Letter
Version: Microsoft Office Word
Revision: 2
Create Date: 07/Oct/2008 12:41:00PM
Last Revision Date: 07/Oct/2008 12:41:00PM
Last Print Date: 26/Feb/2008 11:13:00AM
Number of Pages: 1
Number of Characters: 4191
Number of Paragraphs: 9
Number of Words: 735
Hash: 4E257D2FB2C3A6A7234102F7EEFFD338

Name Of Document: Case_AfterDeduplication\WGdata (3)\172.33.5.196\TestSource\AT S\2. AREA INFORMATION\2008-09\1 BRIDGEND\Dysg\Capital Project Funding\capital project funding apllications-final.xls
Company: NCETW
Title:
Subject:
Author: Nicola.Owen
Keywords:
Comments:
Last Saved By: Nicola.Owen
Template:
Version: Microsoft Excel
Revision:
Create Date: 09/Sep/2008 02:25:31PM
Last Revision Date: 22/Sep/2008 09:26:04AM
Last Print Date: 15/Sep/2008 09:31:09AM
Number of Pages: 0
Number of Characters: 0
Number of Paragraphs: 0
Number of Words: 0
Hash: 73085EC8B8C4050FE959369A7AFFCAC8

Name Of Document: Case_AfterDeduplication\WGdata (3)\172.33.5.196\TestSource\AT S\2. AREA INFORMATION\2008-09\2 CAERPHILLY\Area Statistics and Data\FEI data\YstradMynach College.doc
Company: NCETW
Title: YstradMynach College is one of two colleges located within the County Borough of Caerphilly, the other being Coleg Gwent, Crosskeys Campus
Subject:
Author: Paul.Jenkins
Keywords:
Comments:
Last Saved By: jenkinsp2
Template: Normal
Version: Microsoft Office Word
Revision: 3
Create Date: 24/Nov/2008 01:21:00PM
Last Revision Date: 30/Apr/2009 12:22:00PM
Last Print Date: 28/Aug/2008 02:45:00PM
Number of Pages: 1
Number of Characters: 35127
Number of Paragraphs: 82
Number of Words: 6162
Hash: 32AFDB21BDBFFC6E8D84452663BDE5C5

Name Of Document: Case_AfterDeduplication\WGdata (3)\172.33.5.196\TestSource\AT S\2. AREA INFORMATION\2008-09\2 CAERPHILLY\Area Statistics and Data\General Information\09 04 27 CAERPHILLY.doc
Company: Welsh Assembly Government
Title: CAERPHILLY
Subject:

Author: jenkinsp2
Keywords:
Comments:
Last Saved By: jenkinsp2
Template: Normal
Version: Microsoft Office Word
Revision: 2
Create Date: 27/Apr/2009 08:29:00AM
Last Revision Date: 27/Apr/2009 08:29:00AM
Last Print Date:
Number of Pages: 1
Number of Characters: 18555
Number of Paragraphs: 43
Number of Words: 3255
Hash: 44A52071D71D4AD998C1F258DC475656

Name Of Document: Case_AfterDeduplication\WGdata (3)\172.33.5.196\TestSource\AT S\2. AREA INFORMATION\2008-09\2 CAERPHILLY\Area Statistics and Data\General statisitics\Base Data.xls
Company: Welsh Assembly Government
Title:
Subject:
Author: kinga1
Keywords:
Comments:
Last Saved By: kinga1
Template:
Version: Microsoft Excel
Revision:
Create Date: 10/Mar/2009 10:28:44AM
Last Revision Date: 10/Mar/2009 12:44:26PM
Last Print Date:
Number of Pages: 0
Number of Characters: 0
Number of Paragraphs: 0
Number of Words: 0
Hash: 72275D8EA5D5C5637C0E7732C3D0039C

Name Of Document: Case_AfterDeduplication\WGdata (3)\172.33.5.196\TestSource\AT S\2. AREA INFORMATION\2008-09\2 CAERPHILLY\Area Statistics and Data\LA DATA\080806 LA Data Briefing.doc
Company: NCETW
Title: POPULATION
Subject:
Author: Robert.Joyce
Keywords:
Comments:
Last Saved By: jenkinsp2
Template: Normal
Version: Microsoft Office Word
Revision: 3
Create Date: 06/Aug/2008 04:02:00PM
Last Revision Date: 04/Aug/2009 07:49:00AM
Last Print Date:
Number of Pages: 1
Number of Characters: 25884
Number of Paragraphs: 60
Number of Words: 4540
Hash: 3725185E8D28CD15DC44453C9552B6D4

Name Of Document: Case_AfterDeduplication\WGdata (3)\172.33.5.196\TestSource\AT S\2. AREA INFORMATION\2008-09\2 CAERPHILLY\Area Statistics and Data\LA DATA\LA Data Briefing.doc
Company: NCETW
Title: Population - current and projected, Welsh speakers;
Subject:
Author: Robert.Joyce
Keywords:
Comments:
Last Saved By: jenkinsp2
Template: Normal
Version: Microsoft Office Word
Revision: 12
Create Date: 06/Aug/2008 11:19:00AM
Last Revision Date: 04/Aug/2009 07:54:00AM
Last Print Date: 06/Aug/2008 02:11:00PM
Number of Pages: 1
Number of Characters: 25883

Number of Paragraphs: 60
Number of Words: 4540
Hash: 2525DA73FCE8D0945F9DB88A2F47857C

Name Of Document: Case_AfterDeduplication\WGdata (3)\172.33.5.196\TestSource\AT S\2. AREA INFORMATION\2008-09\2 CAERPHILLY\Area Statistics and Data\LA DATA\WBL 0506.xls
Company: NCETW
Title:
Subject:
Author: Robert.Joyce
Keywords:
Comments:
Last Saved By: Robert.Joyce
Template:
Version: Microsoft Excel
Revision:
Create Date: 06/Aug/2008 02:30:16PM
Last Revision Date: 06/Aug/2008 02:53:50PM
Last Print Date:
Number of Pages: 0
Number of Characters: 0
Number of Paragraphs: 0
Number of Words: 0
Hash: 58FFB964A67BBB4BC0F72C96518967BA

Name Of Document: Case_AfterDeduplication\WGdata (3)\172.33.5.196\TestSource\AT S\2. AREA INFORMATION\2008-09\2 CAERPHILLY\Area Statistics and Data\MAPS\Map of Caerphilly 108e 08-09 Education (2).doc
Company: NCETW
Title:
Subject:
Author: Paul.Jenkins
Keywords:
Comments:
Last Saved By: jenkinsp2
Template: Normal
Version: Microsoft Office Word
Revision: 3
Create Date: 24/Nov/2008 01:06:00PM
Last Revision Date: 04/Aug/2009 07:39:00AM
Last Print Date: 04/Aug/2009 07:38:00AM
Number of Pages: 1
Number of Characters: 1
Number of Paragraphs: 1
Number of Words: 0
Hash: 6EE425A32C374D96F9209098495D5E19

Name Of Document: Case_AfterDeduplication\WGdata (3)\172.33.5.196\TestSource\AT S\2. AREA INFORMATION\2008-09\2 CAERPHILLY\Caerphilly\Schools\List of Schools -Caerphilly.xls
Company: Welsh Assembly Government
Title:
Subject:
Author: kinga1
Keywords:
Comments:
Last Saved By: jenkinsp2
Template:
Version: Microsoft Excel
Revision:
Create Date: 30/Apr/2009 01:39:06PM
Last Revision Date: 07/Sep/2009 10:21:34AM
Last Print Date: 07/Sep/2009 10:21:21AM
Number of Pages: 0
Number of Characters: 0
Number of Paragraphs: 0
Number of Words: 0
Hash: 2DE3129777B81D3EC94E4B736AF37262

Name Of Document: Case_AfterDeduplication\WGdata (3)\172.33.5.196\TestSource\AT S\2. AREA INFORMATION\2008-09\2 CAERPHILLY\Capital Project Funding\capital project funding apllications-final.xls
Company: NCETW
Title:
Subject:
Author: Nicola.Owen
Keywords:

Comments:
Last Saved By: Nicola.Owen
Template:
Version: Microsoft Excel
Revision:
Create Date: 09/Sep/2008 02:25:31PM
Last Revision Date: 22/Sep/2008 09:26:04AM
Last Print Date: 15/Sep/2008 09:31:09AM
Number of Pages: 0
Number of Characters: 0
Number of Paragraphs: 0
Number of Words: 0
Hash: 08F6F655E7164342091E499918911BA4

Name Of Document: Case_AfterDeduplication\WGdata (3)\172.33.5.196\TestSource\AT S\2. AREA INFORMATION\2008-09\2 CAERPHILLY\Estyn\Caerphilly inspection\Area inspection guidance\Area Inspection Guidance en (2).doc
Company: ESTYN
Title: Area Inspection Handbook
Subject:
Author: Nigel Vaughan
Keywords:
Comments:
Last Saved By: Paul.Jenkins
Template: Normal
Version: Microsoft Office Word
Revision: 2
Create Date: 27/Nov/2008 11:11:00AM
Last Revision Date: 27/Nov/2008 11:11:00AM
Last Print Date: 23/Jul/2008 10:23:00AM
Number of Pages: 1
Number of Characters: 63601
Number of Paragraphs: 149
Number of Words: 11158
Hash: 63E67DF2284C40343FC3313195CDDDC6

Name Of Document: Case_AfterDeduplication\WGdata (3)\172.33.5.196\TestSource\AT S\2. AREA INFORMATION\2008-09\2 CAERPHILLY\Estyn\Caerphilly inspection\Distribution letter to 14-19 Learning Network Chair cy.doc
Company: Estyn
Title: Mr Ieuan Ellis
Subject:
Author: Administrator
Keywords:
Comments:
Last Saved By: jenkinsp2
Template: Normal
Version: Microsoft Office Word
Revision: 2
Create Date: 23/Apr/2009 12:21:00PM
Last Revision Date: 23/Apr/2009 12:21:00PM
Last Print Date: 17/Apr/2009 08:35:00AM
Number of Pages: 2
Number of Characters: 1826
Number of Paragraphs: 4
Number of Words: 320
Hash: D9E23B9A1A33B81F231FBE1885ABCB75

Name Of Document: Case_AfterDeduplication\WGdata (3)\172.33.5.196\TestSource\AT S\2. AREA INFORMATION\2008-09\2 CAERPHILLY\Estyn\Caerphilly inspection\PPR - PDP Evaluations.doc
Company: NCETW
Title: Hiya Paul
Subject:
Author: Paul.Jenkins
Keywords:
Comments:
Last Saved By: Paul.Jenkins
Template: e-mail
Version: Microsoft Office Word
Revision: 1
Create Date: 27/Nov/2008 07:44:00AM
Last Revision Date: 27/Nov/2008 08:58:00AM
Last Print Date:

Number of Pages: 1
Number of Characters: 396
Number of Paragraphs: 1
Number of Words: 69
Hash: 1CA90D526651160A8906E8791E019D51

Name Of Document: Case_AfterDeduplication\WGdata (3)\172.33.5.196\TestSource\AT S\2. AREA INFORMATION\2008-09\2 CAERPHILLY\Estyn\Inspection dates (2008-09) 18-07-08.doc
Company: Welsh Funding Councils
Title: Provider
Subject:
Author: Bethan Lewis
Keywords:
Comments:
Last Saved By: Robert.Evans
Template: Normal.dot
Version: Microsoft Office Word
Revision: 2
Create Date: 18/Jul/2008 09:47:00AM
Last Revision Date: 18/Jul/2008 09:47:00AM
Last Print Date: 17/Jul/2008 03:00:00PM
Number of Pages: 1
Number of Characters: 1432
Number of Paragraphs: 3
Number of Words: 251
Hash: 74451C291CA98A40E7685F9BC4750E83

Name Of Document: Case_AfterDeduplication\WGdata (3)\172.33.5.196\TestSource\AT S\2. AREA INFORMATION\2008-09\3 MERTHYR TYDFIL\Area Statistics and Data\FEI data\Merthyr Tydfil College.doc
Company: NCETW
Title: YstradMynach College is one of two colleges located within the County Borough of Caerphilly, the other being Coleg Gwent, Crosskeys Campus
Subject:
Author: Paul.Jenkins
Keywords:
Comments:
Last Saved By: Paul.Jenkins
Template: Normal
Version: Microsoft Office Word
Revision: 2
Create Date: 24/Nov/2008 02:21:00PM
Last Revision Date: 24/Nov/2008 02:21:00PM
Last Print Date: 24/Sep/2008 03:20:00PM
Number of Pages: 1
Number of Characters: 34406
Number of Paragraphs: 80
Number of Words: 6035
Hash: 3C4FAFE69A5E001049A44C83F226E6AA

Name Of Document: Case_AfterDeduplication\WGdata (3)\172.33.5.196\TestSource\AT S\2. AREA INFORMATION\2008-09\3 MERTHYR TYDFIL\Area Statistics and Data\FEI data\University of Glamorgan v2.doc
Company: NCETW
Title: YstradMynach College is one of two colleges located within the County Borough of Caerphilly, the other being Coleg Gwent, Crosskeys Campus
Subject:
Author: Paul.Jenkins
Keywords:
Comments:
Last Saved By: alex.king
Template: Normal.dot
Version: Microsoft Office Word
Revision: 16
Create Date: 30/Sep/2008 01:33:00PM
Last Revision Date: 01/Oct/2008 12:57:00PM
Last Print Date: 01/Oct/2008 11:55:00AM
Number of Pages: 1
Number of Characters: 21350
Number of Paragraphs: 50
Number of Words: 3745
Hash: 03C7B675B0A15A98E7423D7CD278855F

Name Of Document: Case_AfterDeduplication\WGdata (3)\172.33.5.196\TestSource\AT S\2. AREA INFORMATION\2008-09\3 MERTHYR TYDFIL\Area Statistics and Data\General statisitics\Base Data.xls

Company: Welsh Assembly Government
Title:
Subject:
Author: kinga1
Keywords:
Comments:
Last Saved By: kinga1
Template:
Version: Microsoft Excel
Revision:
Create Date: 10/Mar/2009 10:28:44AM
Last Revision Date: 10/Mar/2009 12:44:26PM
Last Print Date:
Number of Pages: 0
Number of Characters: 0
Number of Paragraphs: 0
Number of Words: 0
Hash: 9685A0BDEFE8FED99CEEE970308713DF

Name Of Document: Case_AfterDeduplication\WGdata (3)\172.33.5.196\TestSource\AT S\2. AREA INFORMATION\2008-09\3 MERTHYR TYDFIL\Area Statistics and Data\LA DATA\Merthyr Briefing.doc
Company: NCETW
Title: Merthyr
Subject:
Author: IT Section
Keywords:
Comments:
Last Saved By: jenkinsp2
Template: Normal
Version: Microsoft Office Word
Revision: 8
Create Date: 26/Sep/2008 02:18:00PM
Last Revision Date: 09/Mar/2009 02:59:00PM
Last Print Date: 30/Sep/2008 05:52:00PM
Number of Pages: 1
Number of Characters: 7125
Number of Paragraphs: 16
Number of Words: 1250
Hash: 6FA1EB8F2D913ADC8FD73157DECACE2D

Name Of Document: Case_AfterDeduplication\WGdata (3)\172.33.5.196\TestSource\AT S\2. AREA INFORMATION\2008-09\3 MERTHYR TYDFIL\Dysg\Capital Project Funding\Merthyr College IT facilities Bid.doc
Company:
Title: Institution Name:
Subject:
Author:
Keywords:
Comments:
Last Saved By: Brian.Foster
Template: Normal
Version: Microsoft Office Word
Revision: 2
Create Date: 30/Sep/2008 12:39:00PM
Last Revision Date: 30/Sep/2008 12:39:00PM
Last Print Date: 25/Apr/2008 09:26:00AM
Number of Pages: 1
Number of Characters: 2003
Number of Paragraphs: 4
Number of Words: 351
Hash: 3EB258720806330369A12ACA42233538

Name Of Document: Case_AfterDeduplication\WGdata (3)\172.33.5.196\TestSource\AT S\2. AREA INFORMATION\2008-09\3 MERTHYR TYDFIL\Merthyr\14 - 19\Appendix 2 - Mid Year Monitoring Revenue Dec 08.xls
Company: National Assembly for Wales
Title:
Subject:
Author: devlinm
Keywords:
Comments:
Last Saved By: winiat
Template:
Version: Microsoft Excel
Revision:
Create Date: 04/Nov/2008 02:55:50PM

Last Revision Date: 27/Jan/2009 06:19:57PM
Last Print Date:
Number of Pages: 0
Number of Characters: 0
Number of Paragraphs: 0
Number of Words: 0
Hash: 8452FAE5D6F09261A69E4AA5188EA925

Name Of Document: Case_AfterDeduplication\WGdata (3)\172.33.5.196\TestSource\AT S\2. AREA INFORMATION\2008-09\3 MERTHYR TYDFIL\Merthyr\14 - 19\Appendix 5 Merthyr Tydfil 14-19 Strategy Summary Document (Dec07).doc
Company: Esis
Title:
Subject:
Author: Stuart Broomfield
Keywords:
Comments:
Last Saved By: michaj
Template: Normal
Version: Microsoft Word 10.0
Revision: 2
Create Date: 27/Jan/2009 01:01:00PM
Last Revision Date: 27/Jan/2009 01:01:00PM
Last Print Date: 11/Dec/2007 02:07:00PM
Number of Pages: 1
Number of Characters: 23058
Number of Paragraphs: 54
Number of Words: 4045
Hash: FF585E8D3C0CEC178CC72C86FFE372CC

Name Of Document: Case_AfterDeduplication\WGdata (3)\172.33.5.196\TestSource\AT S\2. AREA INFORMATION\2008-09\3 MERTHYR TYDFIL\Merthyr\14 - 19\Appendix 6b - Example of Quality Framework - Section 3, Self eval provider grid.doc
Company:
Title: Area / success criterion
Subject:
Author: liz
Keywords:
Comments:
Last Saved By: winiat
Template: Normal
Version: Microsoft Office Word
Revision: 3
Create Date: 27/Jan/2009 01:02:00PM
Last Revision Date: 27/Jan/2009 06:10:00PM
Last Print Date: 23/Jan/2009 09:25:00AM
Number of Pages: 1
Number of Characters: 1600
Number of Paragraphs: 3
Number of Words: 280
Hash: 3E044FC84ED51FF10731FA27A99ABBF6

Name Of Document: Case_AfterDeduplication\WGdata (3)\172.33.5.196\TestSource\AT S\2. AREA INFORMATION\2008-09\3 MERTHYR TYDFIL\Merthyr\14 - 19\Copy of Appendix 3B 3C - MT Audit of Combined Curriculum 2008-09.xls
Company: Merthyr Tydfil CBC
Title:
Subject:
Author: michaj
Keywords:
Comments:
Last Saved By: winiat
Template:
Version: Microsoft Excel
Revision:
Create Date: 15/Jan/2009 02:54:02PM
Last Revision Date: 27/Jan/2009 06:14:43PM
Last Print Date: 27/Jan/2009 06:12:43PM
Number of Pages: 0
Number of Characters: 0
Number of Paragraphs: 0
Number of Words: 0
Hash: C732D0DC19DBEAE2ADD3006C6EEB7456

Name Of Document: Case_AfterDeduplication\WGdata (3)\172.33.5.196\TestSource\AT S\2. AREA INFORMATION\2008-09\3 MERTHYR TYDFIL\Merthyr\AGEND4thDec08 ITEM 4a - Adult Community Learning Consultation - WG.doc
Company: Merthyr Tydfil County Borough Council
Title: INTEGRATED ADULT SERVICES ? SERVICE MANAGEMENT BOARD
Subject:
Author: GriffKa
Keywords:
Comments:
Last Saved By: IT Section
Template: Normal
Version: Microsoft Office Word
Revision: 2
Create Date: 30/Dec/2008 02:12:00PM
Last Revision Date: 30/Dec/2008 02:12:00PM
Last Print Date: 20/Nov/2006 01:20:00PM
Number of Pages: 1
Number of Characters: 3158
Number of Paragraphs: 7
Number of Words: 554
Hash: 9F9C032488E50BB1BF1A02E0E43395A1

Name Of Document: Case_AfterDeduplication\WGdata (3)\172.33.5.196\TestSource\AT S\2. AREA INFORMATION\2008-09\3 MERTHYR TYDFIL\Merthyr\Merthyr Learning Steering Group\051208 Project Board Minutes.doc
Company: Merthyr Tydfil CBC
Title: MERTHYR TYDFIL COUNTY BOROUGH COUNCIL
Subject:
Author: MCINTJ
Keywords:
Comments:
Last Saved By: IT Section
Template: 09 Minutes
Version: Microsoft Office Word
Revision: 2
Create Date: 30/Dec/2008 03:03:00PM
Last Revision Date: 30/Dec/2008 03:03:00PM
Last Print Date: 02/Oct/2008 04:01:00PM
Number of Pages: 1
Number of Characters: 4278
Number of Paragraphs: 10
Number of Words: 750
Hash: 2A5C21109B85929EDEF5C11F09E82A4F

Name Of Document: Case_AfterDeduplication\WGdata (3)\172.33.5.196\TestSource\AT S\2. AREA INFORMATION\2008-09\3 MERTHYR TYDFIL\Merthyr\Merthyr Learning Steering Group\4th Dec08Agenda Item 4b - Draft consultation response.doc
Company: NCETW
Title: CONSULTATION RESPONSE FORM
Subject:
Author: LewisClubbeM
Keywords:
Comments:
Last Saved By: IT Section
Template: Normal
Version: Microsoft Office Word
Revision: 2
Create Date: 30/Dec/2008 02:13:00PM
Last Revision Date: 30/Dec/2008 02:13:00PM
Last Print Date: 13/Nov/2008 12:26:00PM
Number of Pages: 1
Number of Characters: 2642
Number of Paragraphs: 6
Number of Words: 463
Hash: 6132D0D18D73353AD627B6F4B1777B09

Name Of Document: Case_AfterDeduplication\WGdata (3)\172.33.5.196\TestSource\AT S\2. AREA INFORMATION\2008-09\3 MERTHYR TYDFIL\Merthyr\Merthyr Learning Steering Group\Agenda Item 9 - 041208ICT and E- Learning Sub Group - Chair's Report Dec 2008.doc
Company: Merthyr Tydfil College
Title: MERTHYR LEARNS: ICT for Learning Sub Group update
Subject:
Author: Gillian Lympany

Keywords:
Comments:
Last Saved By: IT Section
Template: Normal
Version: Microsoft Office Word
Revision: 2
Create Date: 30/Dec/2008 02:15:00PM
Last Revision Date: 30/Dec/2008 02:15:00PM
Last Print Date: 29/Sep/2008 07:54:00PM
Number of Pages: 1
Number of Characters: 3014
Number of Paragraphs: 7
Number of Words: 528
Hash: 207C482073A99246CEBA9E9BE951C17A

Appendix G: Network performance data

TCP Client-Server				
	10MB	**40MB**	**80MB**	**Average**
16	2.5	1.9047		*2.2023*
32	3.3333	2.8571		*3.0952*
64	3.3333	3.6363	4.4444	*3.8046*
128	5	4.4444	5	*4.8148*
256	10	5	5.3333	*6.7777*
512	5	5.7142	7.2727	*5.9956*
1024	10	8	8	*8.6666*
2048	10	13.3333	10	*11.1111*
4096	10	10	11.4285	*10.4761*
8192	10	10	11.4285	*10.4761*
16324	10	13.3333	11.4285	*11.5872*
32768	10	10	11.4285	*10.4761*
65536	10	13.3333	10	*11.1111*
131072	10	10	11.4285	*10.4761*
262144	10	13.3333	11.4285	*11.5872*
524288	10	10	11.4285	*10.4761*
1048576	10	10	10	*10*
2097152		13.3333	11.4285	*12.3809*
4194304		10	11.4285	*10.7142*
8388608			11.4285	*11.4285*

TCP LINUX Client-Client				
	10MB	**40MB**	**80MB**	**Average**
8	1.428571	0.666667	0.266149	0.787129
16	2.5	1.052632	0.512348	1.354993
32	3.333333	1.428571	0.892777	1.884894
64	5	2	0.92115	2.640383
128	5	2.5	0.927902	2.809301
256	10	2.857143	0.929887	4.595676
512	5	4	0.942063	3.314021
1024	10	4	0.94162	4.98054
2048	10	4	0.909753	4.969918
4096	10	5	0.847242	5.282414
8192	10	4	0.828226	4.942742
16384	10	5	0.876424	5.292141
32768	10	4	0.911743	4.970581
65536	10	5	0.918021	5.306007
131072	10	4	0.917515	4.972505
262144	5	4	0.915919	3.305306
524288	10	4	0.924556	4.974852
1048576	10	4	0.918105	4.972702