Guidelines on Performing an Institutional Data Audit

Contents

1 Preface

This document provides guidelines for Higher Education institutions wishing to conduct a data audit, and is an output of work conducted by the IDMAPS Project at Newcastle University.[1]

[1] *Institutional Data Management for Personalisation and Syndication* (IDMAPS) is a JISC-funded Institutional Innovation project which aims to improve the quality and reliability of institutional data flows. For more information, please visit the project website at http://research.ncl.ac.uk/idmaps.

[2] http://creativecommons.org/licenses/by-sa/3.0/.

2 Introduction

2.1 What is a Data Audit?

A data audit is a fact-finding exercise carried out to identify what data an institution holds, as well as how it is collected, processed, used and stored.

2.2 Why conduct a Data Audit?

Although it would seem logical that an organisation should have thorough knowledge of the data it possesses, this may not always be the case.

Like other organisations, Higher Education institutions rely on a wide range of institutional data in order to conduct their day to day operations. Such data is often diverse, and might include items such as payroll data, HR data, IT service provisioning data (computer accounts, print credits, etc.) and academic data (lecture timetabling, exam results, past exam papers, etc.).

Due to its diverse nature, this data is typically stored across many different computer systems, with varying degrees of interoperability. Often, the institution's data infrastructure will have grown organically: individual systems will have been introduced over time to address specific requirements, rather than as part of a coherent plan.

Such incremental additions may well not have anticipated the development of new systems, technological changes, or additional uses of data. As an example, the rapid growth of the World Wide Web in the late 1990s/early 2000s has lead to a proliferation of tools and services for the Higher Education community, many of which rely on specific elements of institutional data to provide authentication and access control.

Higher Education institutions are therefore unlikely to possess a data infrastructure which not only meets their current needs, but has been designed from the ground up for future extensibility and growth. Worse, their overall picture of how their current data infrastructure actually works may well be very hazy, with a lack of clarity regarding issues such as:

- **What** data is collected and from which source(s).
- **Where** and **how** recorded data is stored.
- **What** the data is used for, and **how** it passes both between systems and to data consumers.
- **Who is responsible** for the data at both an operational and a strategic level.

Conducting a data audit allows an institution to gain a better understanding of their existing data infrastructure, which is necessary before measures can be taken to improve the situation if required.

3 Guidelines for conducting a Data Audit

Based on their experience with the project, the IDMAPS team has adopted the following principles in order to maximise the effectiveness of a data audit. They are only guidelines, and should not be taken as comprehensive: there may well be other factors specific to different institutions which are of equal or greater importance. However, they form a basis from which to work, and are adaptable to different institutional contexts.

3.1 Obtain senior management buy-in

Data management is not simply an IT issue requiring technical expertise, but an institutional issue which has implications for governing procedures. Obtaining the **active support of senior management** is therefore crucial in addressing the wider institutional data management issues which an audit may unearth.

Senior Management support can also provide additional influence to ease contacts with different data providers and consumers, and can help smooth potentially turbulent relationships between those responsible for different and/or competing systems or processes. However, the need for such intervention can be minimised if the data audit process actively engages with stakeholders wherever possible.

If the audit is conducted as part of the process of identifying a solution for an institutional problem, the proposed solution requires senior management support. To achieve this, the project team should emphasise the benefits of a simple (yet robust) solution: security, reliability, and legal compliance. Needless complexity within a proposed solution brings unnecessary risks, and is unlikely to be approved by senior management.

3.2 Promote simplicity

A data audit is likely to be more successful if it is **kept simple**. This allows a clearer understanding of the task by colleagues, and helps to avoid scope creep which could compromise the audit.

Clearly define the **scope** of the audit. There may be particular types of data which an institution specifically does, or does not, wish to examine, such as financial data, personal information, academic record data, or research output data. A distinction may be also drawn between automated data processes and manual entries. Some reasons for specifically including or excluding particular categories of data might include legal restrictions, institutional relevance, or resource limitations. Regardless of the reason, clearly document what is, and is not, being covered by the data audit – and the reason(s) why.

Attempt to **gain a broad understanding** of the current situation, rather than trying to document every last piece of data and data flow in great detail. To do the latter requires significant resources, and can easily become an endless task. The most productive way of discovering the main systems and data flows is to talk directly to the people who manage and/or run them, so arrange meetings with system managers or their delegates.

Find out not only what specific managers are responsible for, but also what other data and systems they depend on and who is responsible for those systems. It is also often helpful to speak directly to the administrators who are the day-to-day users of these systems, as they may be aware of practical issues which are relevant to the audit.

3.3 Engage stakeholders

Identifying and **engaging with stakeholders is crucial to the success of the data audit**. Stakeholder engagement can be time-consuming, but meaningful progress with a data audit depends on such engagement. The time and effort spent on this will pay dividends in terms of the quality, the comprehensiveness, and the ultimate utility of the data audit.

Stakeholders are likely to include:

- Those providing the impetus for the data audit (e.g. Senior Management, Records Management, specific projects which rely on data).
- Those producing and processing the data.
- Those running the systems in which data is stored, and those in charge of the processes through which data is passed between these systems.

Stakeholders are more likely to feel a part of the process (and therefore to work towards the project's goals) if their concerns and questions are addressed, and if they are asked for assistance rather than being dictated to.

Face-to-face meetings are very useful, and are in many cases preferable to e-mails or other methods of communication when conducting a data audit. An impersonal form might help gather some information, but sitting down and talking directly with stakeholders is likely to be more productive, and generate much more useful information. It also allows a complex situation to be described and clarified by the stakeholder at the time.

Preparing certain key questions helps to steer the conversation and keep it relevant to the information being sought. Such questions might include:

- What systems/services does your team provide/manage?
- What do these systems/services do?
- Who uses them?
- Where do they get their data from?
- Where do they pass their data to?
- Are there any manual processes involved?
- For any automatic processes (e.g. exports/imports of data), how often do they happen?
- What, if any, processing is carried out upon the input data by the systems/services?

3.4 Use appropriate tools

In order to manage the information being generated by the data audit, the use of appropriate templates and tools should be considered from the outset.

Paper forms or templates are useful for recording the information uncovered during face-to-face meetings in a structured form. These can be as simple or detailed as the individual situation or institutional context requires, though they should be easy to complete and not unnecessarily long. At a minimum, they should provide a consistent structure around which the questions identified earlier (*see above*) can be framed.

In the IDMAPS project, the audit information was gathered in the Data Integration Analysis form.[3] It is more specialised and detailed than a simple data audit, being tailored towards developing a new data infrastructure, and will therefore not represent an appropriate structure for every data audit template.

[3] A copy of this form is available for download from the project website.

Technical tools can ease the process of keeping track of the data audit as it progresses. The use of a revision control system such as Subversion or Git allows many team members to update and edit documents relating to the ongoing data audit without overwriting each others' work.[4]

3.5 Create diagrammatic representations

During the audit phase, take time to create diagrammatic representations of data repositories and flows as they are identified. Although some text is a necessary to provide context and explain specific details, the clarity provided by diagrammatic representations is hugely beneficial. Diagrams are often far easier to understand than a large passage of text.

In addition, diagrammatic representations actively encourage constructive feedback on the data audit from data providers and consumers. Few are likely to read and provide feedback on a long text document. In contrast, printed diagrams will often generate very useful feedback as they are more likely to be clearly understood by stakeholders.

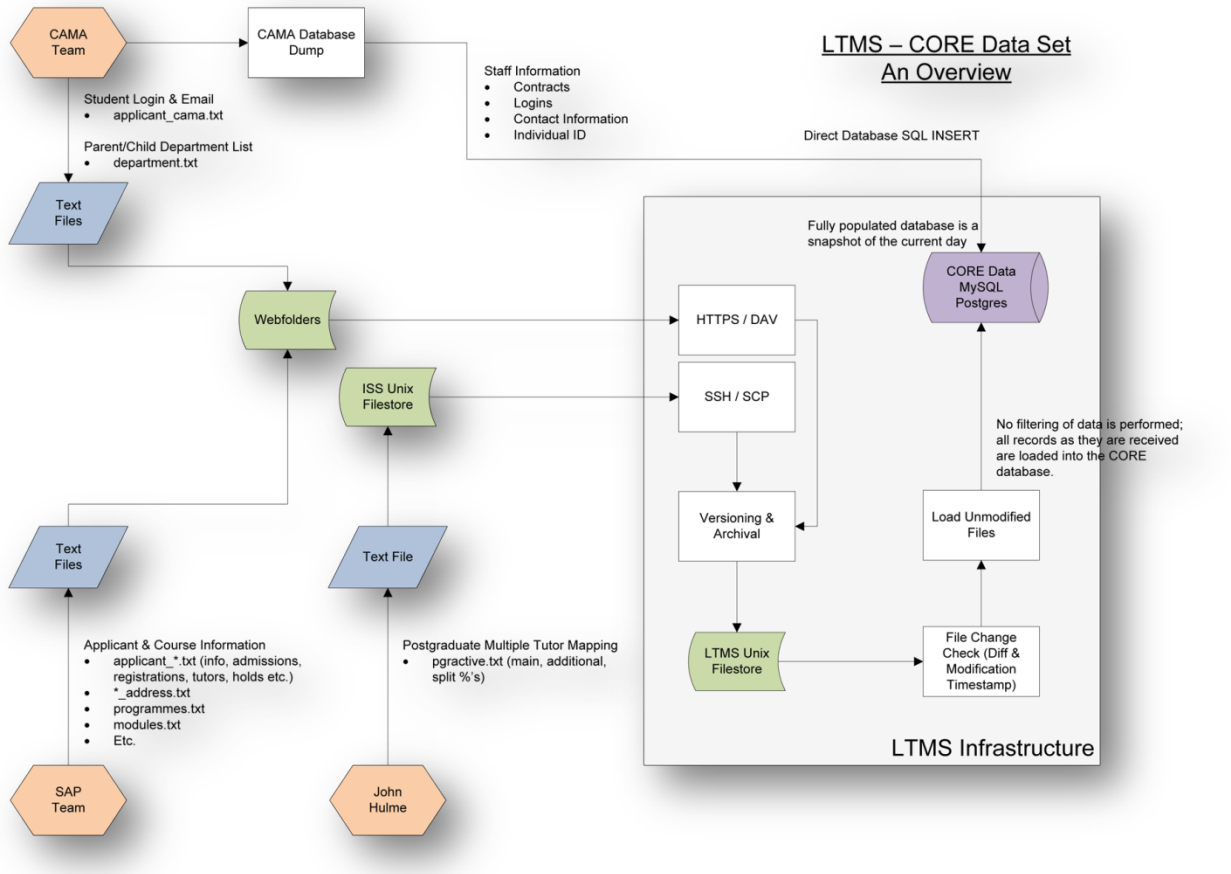

Figure 1: A Diagrammatic Representation of a Data Flow

[4] http://subversion.tigris.org/, http://git-scm.com/.

4 Outputs of a Data Audit

Given that a key reason for performing a data audit is to improve the institution's awareness of its current activities, it is essential to document the outputs of the audit in order to retain this knowledge.

Such documents serve as a hard record of data stakeholders, systems and flows and their interactions, and will form the basis of future project development.

4.1 List of stakeholders and systems

Although the data audit itself should not be over-ambitious, it is worth making the initial list of systems and stakeholders fairly comprehensive. This allows the institution to make informed decisions regarding the relative importance of existing systems and data flows (and their associated stakeholders), and concentrate its effort accordingly.

The following details are generally worth including in such a list:

- Its name(s).
- A non-technical description.
- A technical summary, briefly noting the technologies the system uses.
- Specific details on the groups and individuals associated with the system, including those who are responsible for it (implementers, owners, and managers) and those who depend upon it (users).

Other institution- or context-specific information may also be worth recording at this stage.

Some systems or stakeholders will be obvious, others less so; the relative importance of these will become clearer as the list is completed. It may well be the case that as this list is compiled, other systems and/or stakeholders are uncovered – this was the case with IDMAPS.

4.2 Diagrammatic representation of data flows

As has already been mentioned, diagrams are a highly valuable output of the data audit, as they allow institutions to visualise complex data flows. Such diagrams are most likely to be flow charts, created with tools such as Visio, Kivio, or Dia.[5]

It is important to ensure that any diagrams created are consistent, particularly if they are produced by a team of different people, or over a long period of time. Ensuring consistency will be easiest if all parties standardise early in the process on the use of specific symbols to represent particular objects (such as data flows, data producers, users of data, etc).

[5] http://office.microsoft.com/visio, http://www.thekompany.com/projects/kivio/, http://live.gnome.org/Dia.