Guide to Completing the Data Integration Template

Contents

1 Preface

This document is an output of work conducted by the IDMAPS Project at Newcastle University.[1] It provides guidance on completing the Data Integration Template.[2]

It has been made available under a *Creative Commons Attribution-Share Alike 3.0 License* to the wider Higher Education community in the hope that our experiences will prove useful to other institutions undertaking similar activities.[3]

Any references to third-party companies, products or services in this document are purely for informational purposes, and do not constitute any kind of endorsement by the IDMAPS Project or Newcastle University.

2 Introduction

2.1 About this guide

This Guide is designed to be used in conjunction with the Data Integration Template, which must be completed by anyone requesting data from the Institutional Data Feed Service (IDFS).

It gives a step-by-step explanation of the Data Integration Template, as well as providing example answers for each step where appropriate. Application Providers who have a request for data from the IDFS can follow these examples when completing the Data Integration Template.

2.2 About the Data Integration Template

The Data Integration Template provides a standardised structure through which data requests can be made to the IDFS, and ensures that every data request is supported by comprehensive documentation.

The Data Analysis and Integration Process consists of four phases, each with four defined steps. Every step has a corresponding question within the Data Integration Template.

The Template should be added to as the data request is processed, and contributors should use their own judgment when determining how fully a step can be completed at the time of writing.

It may be the case that some sections can only be partially completed at a given point in time: where this is the case, it should be clearly indicated on the Data Integration Template so that it can be completed at a later date.

The end result should be a complete document which provides an accurate and comprehensive record of the data flow, including technical details of its operation and procedural information which will assist the tracking of information usage within the University.

[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] The Data Integration Template is available to download from the project website.

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

Including clear diagrammatic representations of systems and data flows is beneficial, and will likely reduce the need for verbose answers.

If Application Providers have any questions regarding the Data Integration Template, they are advised to discuss them with ISS staff. The speed and prioritisation of IDFS development work undertaken by ISS on behalf of Application Providers will be greatly increased if the latter provide the former with accurate and detailed information.

The examples used in this Guide are of an imaginary application, “AccessCardApp”, which is based on Newcastle University’s Smartcard system. The examples are as generic as possible so as to be widely applicable across different institutions, and demonstrate the type of answers and level of detail which will be most beneficial to the Data Analysis and Integration Process.

2.3 Definitions

Some words have specific definitions within this document, and are therefore defined below for clarity:

Administrator	The individual/group who administer an Application.
Application	Any application which requires institutional data.
Application Provider	The individual/group who provide an Application.
Guide	This document (*Guide to Completing the Data Integration Template*).
Template	The Data Integration Template.
User(s)	The end users of an Application.

3 Phase A: Requirements Analysis

This phase is to be completed by ISS and the Application Provider. The purpose of this phase is to clarify the current situation/provision and determine the specific needs of the Application Provider.

3.1 Step 1: Describe the application

Provide a brief written summary of the application's remit and function.

Remember to include:

- The purpose of the Application.
- The benefit(s) it brings to the University.

Example 1.1: Written summary of application's remit and function

> AccessCardApp is the system which manages the issuing of University access cards to staff, students, visitors, contractors and certain members of the public.
>
> It can permit card holders to access to specific areas of the campus based on their user category (for instance Staff, Student, Out of Hours), or based on finer-grained door access control systems administered by "Access Control Administrators" around the University.
>
> Access card issuing and replacement is performed by the Library. During the annual registration period, they are issued to students by a registration team largely consisting of ISS staff.

In addition to the written summary, provide the following information in a summary table.

- The categories of Users and Administrators of the application (e.g. all staff, all students, librarians, lecturers, community engagement etc).
- Identify the data for which the system is authoritative in the institute.

Example 1.2: Summary Table

System Administrators	*ISS staff ("Infrastructure Support team"), Library Front desk staff*
System Customers	*Staff, Students, Visitors, Contractors, Lay Library Members*
Data for which the Application is Authoritative Source	*AccessCard chip number (i.e. chip number of current Access Card, not the same as AccessCard number), Users AccessCard photos.*

3.2 Step 2: Describe existing integrations

Provide a brief written summary of existing data feeds to and from this application (if any exist).

Include any data consumed from sources outside of ISS, or hand-keyed. It may be useful to draw a diagram summarizing the data flows at this stage.

This information is used to gain a clear picture of the existing data practices, and to help assess whether data provision can be improved. A secondary purpose is to identify and record current data shortfalls, such as hand keyed data, so that this can be remedied at a future date.

Example 2.1: Description of Existing Integrations

AccessCardApp receives data from the Campus Management and HR systems.

From HR, it receives a file of new staff members to be added to the system on a daily basis, and a file of expired users for deletion on a monthly basis.

From Campus Management, it receives a file of new student users on a daily basis; deletion of expired students is performed manually for those students whose course end date has been reached.

Data imports and account expiry are triggered by manual intervention by “Infrastructure Support Team” staff (generally A Person).

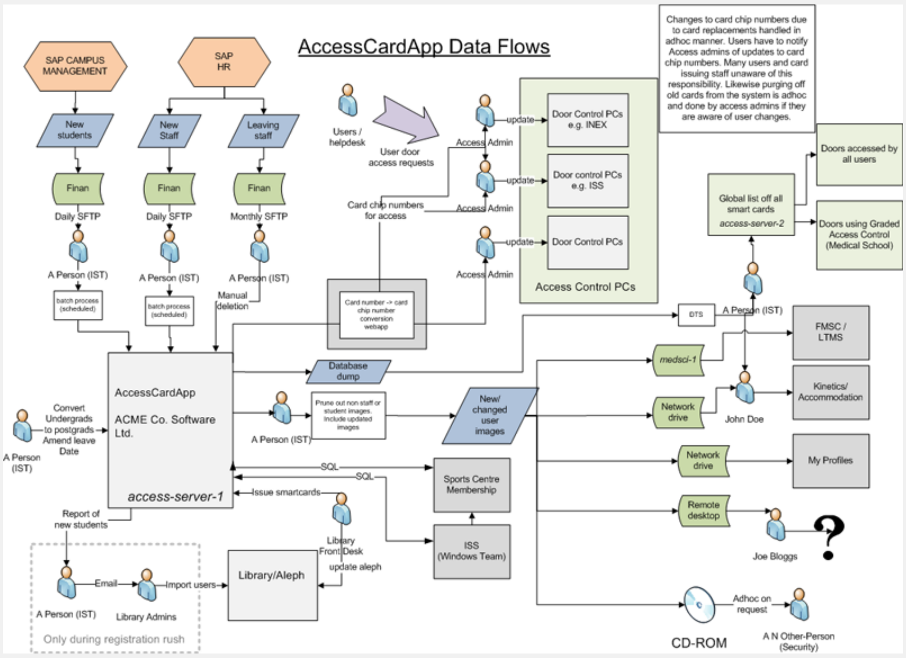
AccessCardApp also feeds new students’ smartcard numbers to the Library Management System about. Staff data is manually inputted when staff members physically visit the library to enrol.

AccessCardApp also feeds data to the coarse-grained door access control systems controlled by the server Locksmith. This mechanism is used to grant access to Medical School.

Although Access Card numbers are used with the fine-grained door access control systems administered by “Access Control Administrators”, there is no formal data feed mechanism. This is instead handled by user interaction and helpdesk request.

AccessCardApp also feeds images of Users to the FMSC, Computer Science, and Accommodation Systems by direct feed, and to A. N. Other-Person in Security on request via CD. For a summary of these flows, see the diagrammatic example below.

There is no automated feed of expired or lost cards into the door access control systems: it would be desirable to update access control on card replacement or user expiry.


AccessCardApp Data Flows
SAP CAMPUS MANAGEMENT
SAP HR
New students
New Staff
Leaving staff
Finan
Finan
Finan
Daily SFTP
Daily SFTP
Monthly SFTP
A Person (IST)
A Person (IST)
A Person (IST)
batch process (scheduled)
batch process (scheduled)
Manual deletion
Users / helpdesk
User door access requests
Card chip numbers for access
Card number -> card chip number conversion webapp
Access Admin
Access Admin
Access Admin
update
update
update
Door Control PCs e.g. INEX
Door control PCs e.g. ISS
Door Control PCs
Access Control PCs
Changes to card chip numbers due to card replacements handled in adhoc manner. Users have to notify Access admins of updates to card chip numbers. Many users and card issuing staff unaware of this responsibility. Likewise purging off old cards from the system is adhoc and done by access admins if they are aware of user changes.
Doors accessed by all users
Global list off all smart cards
access-server-2
Doors using Graded Access Control (Medical School)
A Person (IST)
Convert Undergrads to postgrads Amend leave Date
AccessCardApp
ACME Co. Software Ltd.
access-server-1
Database dump
A Person (IST)
Prune out non staff or student images. Include updated images
New/ changed user images
DTS
A Person (IST)
FMSC / LTMS
medsci-1
Network drive
John Doe
Kinetics/ Accommodation
Network drive
My Profiles
SQL
SQL
Sports Centre Membership
Issue smartcards
ISS (Windows Team)
Remote desktop
Joe Bloggs
?
Report of new students
Library Front Desk
update aleph
Email
Import users
Library/Aleph
A Person (IST)
Library Admins
Only during registration rush
CD-ROM
Adhoc on request
A N Other-Person (Security)

AccessCardApp Data Flows: such diagrams must be large enough for all detail to be visible.

3.3 Step 3: Describe data involved in proposed integrations

Clearly describe the data that this application requires to perform its function.

Include definitions of the data created, or added to, by the application itself. This should include all data which the application provides and other applications depend on or could benefit from.

This information is used to help define enterprise data feeds and flows, and to identify which system (or combination of systems) is authoritative for specific institutional data.

Where appropriate, the empty tables supplied in the Template may be useful in recording this information.

A descriptive example of both Data Consumption and Data Production templates are included below as Example 3.1.

Example 3.1: Commented Example Tables of Data Consumption and Production

Data Consumption					
Authoritative Source System:	The authoritative source of the data (if known) *e.g. SAP HR*				
Intermediary Source System(s):	External data processing systems that pass on the processed data *e.g. CAMA*				
Source Data Structure:	*e.g. fixed-width file import, .csv import, SQL dump*				
Source Field Name	**Field type**	**Width**	**Nullable**	**Destination Field Name**	**Data processing and comments**
e.g. Age	*int*	*3*	*YES*	*UserAge*	*Processing (Note1)* *Comment (Note2)*
etc.	*etc.*	*etc.*	*etc.*	*etc.*	*etc.*
Notes: 1. *Age is converted into months as this is what the application expects. The conversion is performed by a Visual Basic scripted data import tool.* 2. *Often age is missing from the data feed, in which case – since the application requires an age field –the figure of 2400 months (200 years) is used as all users are suitably old to use the application.*					

Data Production				
Data Destination	*e.g. .csv file dump, insert record into database table*			
Field Name	**Field type**	**Width**	**Nullable**	**Description and comments**
e.g. tutor id	*varchar*	*20*	*YES*	*The campus username of the students tutor where known*
etc.	*etc.*	*etc.*	*etc.*	*etc.*
Notes: *None*				

In addition, the completed AccessCardApp tables are provided below as Example 3.2, without the descriptive comments.

Example 3.2: AccessCardApp Data Consumption and Production Tables

Data Consumption					
Authoritative Source System:	SAP HR				
Intermediary Source System(s):	-				
Source Data Structure:	AccessCardStarters.txt – fixed width, comma-separated value file.				
Source Field Name	**Field type**	**Width**	**Nullable**	**Destination Field Name**	**Data processing and comments**
Staff number	varchar	9	NO	PREF	Employee number with S and multiple 0s to pad to 9 characters
surname	varchar	38	NO	PLASTNAME	
firstname middlename	varchar	55	NO	PFIRSTNAME/ PMIDDLENAME	Middle name is converted to initial when printed
date of birth	varchar DD/MM/YYYY	10	NO	PDOB	Converted to DBTime Stamp (16)
expiry date	varchar MM/YYYY	7	NO	PEXPIRYDATE	Converted to DBTime Stamp (16)
staff card category	integer	1	NO	PTYPEID	3 = Staff
department/school code	varchar	4	NO	PDEPT	
Notes: *None*					

Data Consumption					
Authoritative Source System:	SAP HR				
Intermediary Source System(s):	-				
Source Data Structure:	AccessCardLeavers.txt – unrecognized format, see Note 1.				
Source Field Name	**Field type**	**Width**	**Nullable**	**Destination Field Name**	**Data processing and comments**
Staff number	varchar	8	NO	PREF	Employee number without the S but still padded with multiple 0s to 8 characters
surname	varchar	Variable, delimited by CR/LF	NO	PLASTNAME	Added directly after employee number
Notes: *1. Format is the SAP payroll number (including leading zeros but not the S) with the surname sandwiched on at the end of those records marked as LEAVERS on SAP by HR during the previous month. For example:* *00000123Person* *00000943Smith* *00000283Jones*					

Data Consumption					
Authoritative Source System:	SAP CAMPUS MANAGEMENT				
Intermediary Source System(s):	-				
Source Data Structure:	sap2smart<*data+timestring*>.txt *e.g. sap2smart081125_0800.txt* Fixed-width, comma separated value file.				
Source Field Name	**Field type**	**Width**	**Nullable**	**Destination Field Name**	**Data processing and comments**
Student Number	varchar	9	NO	PREF	Multiple leading 0s to pad to 9 chars
surname	varchar	38	NO	PLASTNAME	
firstname middlename	varchar	55	NO	PFIRSTNAME/ PMIDDLENAME	Middle name is converted to initial when printed
date of birth	varchar DD/MM/YYYY	10	NO	PDOB	
expiry date	varchar MM/YYYY	7	NO	PEXPIRYDATE	
postgrad or undergrad	integer	1	NO	PTYPEID	1 = PG, 2=UG
department/school code	varchar	4	NO	PDEPT	
***Notes:** None*					

Data Production				
Data Destination	AccessCardOutput.txt Comma separated value file.			
Field Name	**Field type**	**Width**	**Nullable**	**Description and comments**
Smartcard chip number	varchar	8	NO	Crucial to access control, as it is chip numbers (not users/card numbers) that are granted access.
Smartcard user image	BLOB	n/a approx 6k	YES	Can be much larger than 6k
Notes: *None*				

3.4 Step 4: Describe the high-level flow of data

Provide a high-level description of the desired flows of data into this application.

This should include:

- The preferred data consumption method
- The frequency of data transfer
- The transport methodology used.

Examples might include a flat-file nightly feed, direct database querying, LDAP querying, web service querying, etc.

The purpose of this step is to identify what the optimum data flow would be. This may help identify a shortfall in the institutional infrastructure (for instance, whether better quality user data is required in LDAP or in a central, live, query-able database).

A diagrammatic representation summarizing the data flows may be useful during this step. This may take the form of an annotated, or otherwise modified, copy of existing diagrammatic representations such as that found in Step 2 (Example 2.1).

Example 4.1: Desired High-level Data Flows

Not Applicable, already in place.

4 Phase B: Integration Work Required

This phase is to be completed by ISS, based on their experience and knowledge of the institutional data infrastructure and the customer's responses to Phase 1.

The purpose of this phase is to determine the amount of work required to implement the data feed, and allocate the resources appropriately.

4.1 Step 5: List new/existing enterprise data required

List the use of enterprise data fields and data processing rules, both new and pre-existing, that will be integrated to provide data for this Application.

Example 5.1: List of Existing/New Enterprise Data for Integrations

Not Applicable, already in place.

4.2 Step 6: Define data feeds used to implement integrations

Provide a definition of the data feed(s) that will provide the required data to the application.

Even if this step involves the re-use of an existing data feed, this should be recorded with the same level of detail and accuracy as a new data feed. This ensures that all parties are aware of where data is being passed to and from, which will help maintain a documented and understandable flow of data over time.

If data processing is required to produce the output data, the details of this processing should be documented in this step.

Example 6.1: Define Data Feeds

Not Applicable, already in place.

4.3 Step 7: Detail transport methodology

Provide detail of how data will be transferred to and from the application.

This should include:

- Mechanism
- Technology (e.g. SFTP, SQL)
- Frequency of transfer.

In addition, include details of the procedures for logging, error handling and correction, and failure-alert procedures.

Example 7.1: Detail Transport Methodology

Not Applicable, already in place.

4.4 Step 8: Work estimation and scheduling

Identify available resources, estimate the work involved with each aspect of the project, and prepare a project plan.

The work plan may well be produced as a separate document, but a brief overview should be included here. This overview should provide a high-level description of the work to be undertaken, the people involved with this work, and the expected delivery date.

The expected completion date should be communicated with the customer.

Example 8.1: Work Estimation and Scheduling

Not Applicable, already in place.

5 Phase C: Test, Deploy, Monitoring

This phase is to be completed by ISS and the customer. Once a security assessment is completed, test data feeds will be created and evaluated for suitability. Live feeds will then be implemented based upon the findings of the test feeds, and monitored over the lifetime of the feeds.

5.1 Step 9: Security Assessment

This step is essential: without proper security no live data will be sent.

Conduct a security assessment of the proposed application platform and transport methodology.

Ensure that the security responsibilities of the customer are clear and understood.

Commitments to suitable security precautions and procedures should be received in writing and logged.

Example 9.1: Security Assessment

> Already in place.
>
> The ad-hoc nature of purging of expired and lost cards is not optimal, and may present a risk.

5.2 Step 10: Test

Set up test data feed(s) and test application integration.

This step should ascertain the fitness for purpose of the feeds, and whether they enable the application to perform as desired.

Example 10.1: Test

> Not Applicable, already in place.

5.3 Step 11: Deploy data feed

Set up, configure and document the system and processing to provide the required data feed as a robust production service.

In particular, any processing carried out should be well documented and made available to the customer.

Example 11.1: Deploy data feed

> Not Applicable, already in place.

5.4 Step 12: Monitoring and log

Set up and document monitoring systems to automatically monitor the data integration process.

These should alert both the IDFS and the Application Provider of any failure.

It may be appropriate to log each data event at the application, as a minimum logging of statistical information (such as the number of user adds/deletes) is required.

Where appropriate, set up audit logs as well.

Example 12.1: Monitoring and log

Since most of the data feed procedures are carried out manually at the instigation of Infrastructure Support Team staff, monitoring is effectively performed by user interaction.

Infrastructure Support Team staff manually correct anomalies and errors in the process as they perform daily imports. No audit logging requirement has been expressed.

6 Phase D: Service Responsibilities, Boundaries, Documentation

This section is to be completed by ISS and the customer. It sets up the ongoing policies and procedures which will govern the use of the data feeds.

6.1 Step 13: Legal compliance, Data protection, FOI procedure

Outline the data protection responsibilities and procedures associated with the application.

Provide contact details of those responsible for data protection and freedom of information issues relating to the application.

Detail any further legal or policy compliance issues.

Example 13.1: Legal Compliance, Data Protection, FOI procedure

Falls under existing ISS FOI and Data Protection procedures.

6.2 Step 14: Define support structures

Clearly define the division of responsibilities for Application User support and IDFS technical support, documenting contact details for each.

Clarify that User enquiries will be dealt with by the Application Provider, and provide relevant contact details (such as e-mail addresses or support websites) which can be passed on to the ISS Helpdesk.

Document contact details of IDFS representative for any technical queries which might be made by those providing data to the Application or consuming data from it (including the Application Provider).

Provide detail of the support boundaries: for instance, clarifying that the Application Provider is responsible for queries about Application's functionality and ISS is responsible for fixing errors or applying changes to core data.

Example 14.1: Define Support Structures

Support structures are well defined. The library handles card updates replacements and issuing. Helpdesk receives access control requests and passes them on to Access Control Administrators for each of the door control PCs. The Access Control Administrators also receive direct requests for door access.

The Infrastructure Support Team provides application level support. While feeds for image data are ad hoc, current customers understand the need to contact the Infrastructure Support Team with requests, though potential new customers will struggle to identify who they need to ask without knowledge of ISS.

6.3 Step 15: Shortfalls

Provide information about any identified or possible shortfalls in the solution which is being provided by IDFS.

The purpose of this step is to alert any subsequent review process to potential improvements which could be made in the data provisioning architecture.

This step will also help to inform the institutional risk log.

Example 15.1: Shortfalls

> The lack of automated user deletion can lead to a poor purging of defunct user accounts, as this process is heavily reliant on staff interaction. This poses an institutional risk which could be improved with future developments.

6.4 Step 16: Documentation

IDFS must update their own documentation to reflect any changes made by introducing the Application, whilst the Application Providers must also update their own documentation.

These updates should include accurate changes to any diagrammatic representations of data feeds. Application Providers understanding of data roles, responsibilities and the architecture itself must be sufficiently detailed and well-documented.

Example 16.1: Documentation

> IDFS documentation has been updated to reflect the changes made in providing this Application. The diagrammatic representation of the Access Card data flows (AccessCardApp.png) has been changed.
>
> The ISS Infrastructure Support Team has updated their documentation; the status of the Library's documentation is currently unknown.