Project: Data Language

End Project Report

July 2016

Purpose

The aim of this report is to:

- Review actual project outcomes against those specified in the Project Initiation Document
- Evaluate project deliverables against the original objectives
- Identify residual risks
- Summarise lessons learned

Project objectives

The overall aim of this project was to engage with HE stakeholders to define the Common HE Student Dataset and supporting Data Specifications. The project objectives were therefore to:

- Identify the significant student HE data collectors and their data requirements;
- Define and agree the Common HE Student Dataset;
- Define and agree Data Specifications for the data within the Common HE Student Dataset;
- Adhere to the HEDIIP data principles identified by the Blueprint for a New HE Data Landscape report;
- Transition the Data Specifications and Common HE Student Dataset to HESA for implementation as part of the HESA Data Futures Programme and ongoing maintenance and development.

The desired project outcomes were as follows:

1. A common HE Student Dataset with Data Specifications that meets the needs of the significant student data collectors;
2. Data collectors across the HE landscape that collect data within the Common HE Student Dataset are willing to adopt the Data Specifications;
3. Data Specifications that enable HEPs to provide student data to data collectors in a single format and structure;
4. Consensus on the adoption of the Common HE Student Dataset and Data Specifications by stakeholders across the landscape.

The outcomes set out HEDIIP's stretching long-term goals for the data language, representing the transformative new data language which the programme is building rather than the specific end-state to be achieved by July 2016. The Project Board confirmed that the project should be willing to accept a level of risk in order to ensure that substantial progress was made towards these transformative objectives.

Outcome 1 has been substantially progressed:

- The proposed model meets the needs of HE student data collectors, particularly those currently met by the HESA Student return;
- The proposed model is capable of meeting the needs of curriculum data collectors, with further development. In particular the detailed information about assessment needed by certain PSRBs is not yet achieved;
- The model requires further analysis to ensure that it is capable of meeting the need for data collection in respect of FE in HE provision;
- The proposed model is not capable of supporting the likely needs for TEF data, but initial analysis suggests that it should be capable of development to meet this need.

Outcome 2 has been substantially progressed:

- Existing HESA public purpose customers are willing to adopt the Data Specifications;
- Other data collectors have been engaged as stakeholders, and have expressed general support for the project, but have not yet signed up to adopt the Data Specifications.

Outcome 3 has been substantially progressed:

- Mapping has been undertaken to demonstrate that the current HESA Student, AP Student, Aggregate Offshore, KIS, Provider Profile and ITT returns can be accommodated in the format and structure of the Data Language;
- Mapping has not been undertaken of non-HESA returns. Most collections of aggregate student data should be accommodated within the language. Data collection by PSRBs which includes detailed assessment data and/or detailed data about staff will not be.

Outcome 4 has been substantially progressed:

- Engagement with the informal stages of the consultation was very extensive (88 HEPs sent representatives to project workshops or other events, many of them multiple times), but responses to the formal consultation were very low (845 unique visitors to the consultation website made just 11 formal responses). A high level of awareness/engagement with the project and a low level of response to the consultation implies a deliberate choice on the part of most HEPs not to respond. This can only be taken as indicating that the HEPs do not object to the Common HE Student Dataset and Data Specifications. This implies acceptance rather than consensus.

The project therefore represents a major stepping-stone towards the achievement of the new data landscape envisioned through the HEDIIP programme.

The quality expectations for the project deliverables were as follows:

Product	**Ref**	**Description**	**Quality expectations**	**Outcome**
Key stakeholder list	PD01	A list of the key stakeholders who will be engaged by the project	• Includes the HESA statutory customers • Includes representative HEPs from each UK region, HEIs, FEIs and alternative providers • Includes representative PSRBs identified from the HEDIIP data collection inventory	• Met • Met • Met
Common HE Student Dataset model	PD02	A logical student data entity relationship diagram - that defines the data entities within the Common HE Student Dataset and the relationships between them	• Consensus from the stakeholders engaged by the project (PD01) • Data Model uses a 'Crow's foot' notation method approved by the Project Board • Data model describes all the relationships between the data entities in the catalogue (PD03)	• Acceptance rather than consensus. See consultation feedback • Met • Met
Common HE Student Dataset catalogue	PD03	A catalogue providing the name and description of each entity identified in the Common HE Student Dataset entity model (PD02), to include a controlled vocabulary of key terms that are not necessarily data entities.	• Covers entities identified in PD02 • Consensus from the stakeholders engaged by the project (PD01)	• Met • Acceptance rather than consensus. See consultation feedback
Data Specifications	PD04	Directory of all the logical data items associated with the entities in the Common HE Student Dataset catalogue (PD03) which includes: name, description, type, coverage, logical quality/integrity rules, specification of coding frames/valid values, reason, field length, occurrences, owner	• The specifications cover the metadata included in the HESA Data specification 'Data items' directory • Covers all the data items associated with the entities identified in PD03 • Structure of the specifications is approved fit for purpose by the Project Board • Provides a sound basis for CACHED to progress	• Met • Met • To be confirmed by Project Board • Met

The criteria used to measure whether or not the project is complete are as follows:

- All the specified project objectives have been met.
- All project management and control processes have been completed.
- HESA Data Futures Programme accepts the project deliverables.
- The End Project Report has been completed and approved.

Project delivery

The project was delivered broadly in line with the project approach set out in the PID.

Project initiation ran through to mid-November 2015. During this phase the project plan was developed, a stakeholder list agreed and meetings arranged with key stakeholders.

A period of exploration, development and consultation ran from November 2015 through May 2016. However this period was delivered in a more structured and phased way than originally planned. The period November 2015 – January 2016 was largely devoted to stakeholder meetings with data collectors and users. A period of intense engagement mainly with HEPs followed from February 2016 – May 2016. This engagement mainly featured workshops and conference sessions, with 11 workshops and three conference sessions delivered, but there were also highly valuable visits to a number of specific HEPs. In total 88 HEPs were engaged, many of them repeatedly. We published 5 versions of the model over this period in order to generate feedback. In June 2016 we published a version of the model for formal consultation with the sector. In general, the project experienced nothing like the expected difficulty in engaging stakeholders. HEPs took a keen interest in every aspect of the project and rapidly developed an enhanced understanding of the work being done and the possible implications. However whilst the iterative process was successful in securing engagement, it was not always popular with those HEP colleagues who were engaged and some HEPs expressed a wish for a more structured consultation process. The project responded to this by introducing more structure into our consultation process as the project developed, but the very tight timescales for the project would always have made it impossible to deliver a high volume of consultation if that consultation were also formally structured. It was also notable that responses to the formal consultation document, when it was published, were few in number.

Drafts of the key deliverables – PD02 (Data Model), PD03 (Glossary) and PD04 (Data Dictionary) – were published by 9 May 2016, with PD02 and PD03 published in early drafts much earlier than this. The original deadline was to circulate these to the Project Board on 20 May 2016, so timing was in line with (or even ahead of) expectations but process was not. On reflection, the project team may not have fully appreciated the sheer size of PD04, and the accordingly limited value of putting a draft in front of the Project Board, when the Board would not be able to review or discuss it in any depth.

Responses to the formal consultation were very low (845 unique visitors to the consultation website made just 11 formal responses). The combination of the high level of awareness/engagement with the project generated via the informal and a low level of response to the consultation implies a deliberate choice on the part of most HEPs not to respond. This can only be taken as indicating that the HEPs do not object to the Common HE Student Dataset and Data Specifications. This implies acceptance rather than consensus. The responses that were received cannot be taken as a meaningful sample of opinion in the sector, but they did also reflect this pattern, choosing not to contest the main features of the model, but to raise specific points of detail, or issues about implementation. The evidence therefore suggests that the sector accepts the model as a basis for development in Data Futures.

The project was overseen by a Project Board made up of:

Project Executive
Andy Youell – Programme Director, HEDIIP

Senior Suppliers/project team
Dr Andrew Fisher – Lead supplier
Dan Cook – Head of Collections Development, HESA
Simon Robshaw – Data/Information Architect, HESA
Jenni Cockram – Programme Officer, HEDIIP. Data Language Project Manager

Senior Stakeholders
Catherine Elliott – Education and Skills Manager, Engineering Council
Christine Giles – Deputy Academic Registrar, University of Portsmouth
Hannah Falvey – Head of Statistics, HEFCW
Helen Mansfield – Head of Workforce & Development, HEE North West London
Jayne Rowley – Director of Business Services, Graduate Prospects
Joanna Wagstaffe – Director of Academic Registry, University of Law
Paul Hazell – Assistant Director, Research Information & Enquiry, QAA
Rob Wyn Jones – Consultant, Jisc Special Projects, Jisc
Paul Baron – Programme Manager, HEDIIP

Resources and milestones

Milestone	Planned Date	Achieved Date
Initiation		
Project Initiation Documentation to Project Board	**2015-10-26**	**2015-10-26**
Project Board review Project Initiation Documentation	**2015-11-02**	**2015-11-02**
Exploration, development and consultation		
Interim deliverables for review at sector conferences	**2016-03-16**	**2016-03-15**
Draft PD02, PD03 & PD04 to Project Board	**2016-05-20**	**2016-05-18**
Refinement and handover		
Project Board review draft PD02, PD03 & PD04	**2016-05-27**	**2016-05-25**
Final deliverables to Project Board	**2016-07-01**	**2016-07-08**
Project Board review final deliverables	**2016-07-08**	**2016-07-15**
Project completion	**2016-07-15**	**2016-07-29**
Close		
End Project Report	**2016-07-22**	**2016-07-08**

- The final Project Board meeting was a week later than planned because of Board members' availability;
- Project completion has been delayed by a further week after the final Project Board meeting. Budget is available to produce some additional guidance and resources to explain the model and support its use.

On-going risks

The key ongoing risks are R_DL_005 (the risk that specialist data collection requirements within stakeholder organisations may limit standardisation) and R_DL_008 (the risk that the moving environment of dependent projects such as Data Futures may affect adoption). In particular the decision has been taken to ensure that the first phase of Data Futures is limited to meeting the key needs of existing HESA public purpose customers. This is a sensible decision given the over-riding importance of these requirements; but if this decision is allowed too much influence over the future development of the data language itself (R_DL_008) then the requirements of other stakeholders may not be met (R_DL_005) and the ultimate take-up of the language may not meet expectations. The opportunity for rationalisation of data collections may therefore be lost.

Lessons Learned

The project has achieved its objectives and is substantially complete.

What went well?

- During the early, iterative phase, the project was highly successful in engaging the attention of stakeholders, particularly stakeholders in HEPs, and thereby gained valuable feedback about HEP data practices which improved the quality of the project outputs;
- The project products were delivered and met quality expectations, despite the very tight timescales;
- The make-up of the Project Team brought together different skill-sets and expertise which worked well in delivering the project.
- The project raised the profile of the importance of a common data language.
- The project was able to capture a lot of feedback through the iterative engagement process and colleagues could see that their comments had been taken into account.
- Colleagues found the publication of the consultation timeline helpful.
- The workshop with HESPA colleagues succeeded in providing attendees with a better understanding of the Data Language project and how it relates to the Data Futures work.

What could have been improved?

- While the relatively informal approach to consultation, coupled with rapid iterations of the model, allowed the project to progress at the required speed, it elicited some negative feedback from stakeholders.
- There was insufficient clarity in early project communications to stakeholders. Stakeholders did not all understand that that project was developing a logical model for the whole sector, as opposed to the specification for a HESA data return.
- Presentation of the model – had we known there was demand for the EA version of the model we could have published this much sooner.
- An unplanned week of intensive project team meetings was necessary to work through some of the issues in detail. While the bi-weekly meetings seemed to be at sensible intervals it worked better to extend the meeting slot to include a period for the project team to workshop issues.

Handover and follow on actions

There are three follow-on actions:

- Governance of the model;
- Implementation of the model by HESA in Data Futures. A project to undertake this work is in hand;
- Continued development of the model by the Higher Education Data Landscape Steering Group.

1. Governance

Governance of the model is the responsibility of the Higher Education Data Landscape Steering Group. The Group must define a process for exercising that governance. Whilst it is clearly desirable that the model should be as stable as possible, feedback from the ISB has identified the importance of a governance model that can reach decisions rapidly in order to meet the project deadlines of users of the model (including both HESA and others). The alternative will be that users will modify the model *ad hoc* for their own purposes.

2. Data Futures

A project to define a data collection process and schedule in Data Futures using the data language has already been scoped. This initial use of the data language will concentrate on replacing the key existing data returns required by HESA Public Purpose customers, especially the HESA Student, AP Student and ITT returns. The project has undertaken analysis of the existing HESA returns and has a high level of confidence that the entities and

attributes defined will support these requirements. However, undertaking this further work is likely to involve HESA in extensive and detailed analysis of PD04 (the Data Dictionary) and is likely to identify improvements that are required in proposed coding frames. HESA should work closely with the Higher Education Data Landscape Steering Group to ensure that these requirements are identified and implemented in a timely fashion.

3. Development

The model requires further development in order to meet the needs of further groups of data collectors. The Higher Education Data Landscape Steering Group should allocate resources and agree priorities for this development work in light of the expected value of different areas of development. The expected value is likely to be determined by considering the likeliness of adoption (mainly driven by the willingness of relevant stakeholders to adopt) and the impact of adoption on students, data collectors and HEPs, and the Higher Education Data Landscape Steering Group will be better placed to assess these than the project team. However, areas for consideration include:

3a. Teaching and assessment data

The current model includes limited data about teaching, learning and assessment methods, broadly sufficient to meet KIS requirements. More detailed data (including details of contact hours and individual assessment tasks) are likely to be needed for TEF purposes, and so that a greater range of PSRBs can adopt the language for their own accreditation purposes. This would potentially have important benefits for HEPs and data collectors.

3b. Curriculum data

The current model includes a Course entity which is explicitly not intended to model the administrative curriculum structures in HEPs. The mapping between each HEP's own curriculum structure and the data language is left as a matter for each HEP. Whilst this has the benefit of greatly simplifying the development and presentation of the model, this apparent simplicity has not really be delivered in practice – HEPs still have to meet the challenge of mapping their provision to the model. A number of HEPs (although far from all those involved in the project) have expressed an interest in using the data language for their internal data management. Further work in this area would potentially be of significant benefit to HEPs.

3c Public information for applicants

Data currently published to applicants via KIS are out with the scope of the first implementation of the model in Data Futures. This is a sensible reflection of the need to manage the risk profile of Data Futures. However the model has scope to support a significant improvement in the volume of structured data published for students, applicants and their advisers, for instance relating to accommodation; financial support offers; methods of teaching, learning and assessment; languages of delivery.