Introduction to the HEDIIP Data Model

This paper is intended as an introduction to the HEDIIP logical data model. It discusses:

- What a logical data model is
- What it is for
- How the HEDIIP data model works
- How the HEDIIP data model will be used.

About the Data Language project

HEDIIP has been established to enhance the arrangements for the collection, sharing and dissemination of data and information about the HE system. To achieve this objective, HEDIIP has developed a vision for the future of the HE data landscape. The Data Language project is one of four elements to deliver the new HE data landscape, which envisions more timely, accurate and accessible data at a reduced burden to the sector.

The project has developed a logical student data model for the HE sector, with a common data language. The project began in November 2015 and completed in July 2016. At the end of the project, the data model was handed over to the Higher Education Data Landscape Steering Group (HEDLSG) for continued development and governance. The data model will be used by HESA in the Data Futures Programme.

What is a logical data model?

A logical model is a general depiction of the data – entities, attributes and relationships – that are relevant to a particular subject area (domain); in this case Higher Education students and curriculum. A logical data model is not the definition of a data return or the design for a database, although databases and data returns can be designed using the data model.

A logical data model is:

1. A detailed representation of the sector's data. It is not intended as a high-level overview. It should specify the detailed attributes and relationships of all the entities that are described. For example, the data language specifies how many characters can be used in a module title so that anyone who is trying to collect or store module titles in a computer system knows what size to make the relevant fields.
2. Independent of any particular technology. The data model can be implemented in many different technologies, and is not particularly committed to any technology. In this way it is different from the specification of (for example) a HESA return, which tells you exactly what format and method must be used to transfer the required data.
3. Created using a defined business language. The language is *defined*, so that words like 'course' and 'module' have particular meanings which may not be identical to the way those words are used in other contexts; the language is a *business* language as it relates to the business of higher education: it defines terms like 'course' and 'module' rather than anything relating to IT or programming.

A logical data model is not:

1. High-level and conceptual. The model is not intended to provide a high-level overview of key concepts. It is a detailed map of the sector's data intended for uses (such as specifying data returns) that require such detail. High level conceptual models can be created, and they are useful for other purposes. This is something the HEDLSG might do in future.
2. The design for any specific database (or HESA return). It can be used as the basis for many different implementations.

3. Created using any particular programming language. The data model can be implemented using a wide variety of different technologies.

The logical data model provides a map of the sector's data. Like any map, it is not the territory: it abstracts detail that is not considered relevant. For instance, the data model only allows each person to have one previous surname, whereas in reality a person may change their surname many times. This reflects the fact that a person's previous surname is useful in data collection and exchange, because it is a useful check to ensure a person's record does not get mixed up or confused. On the other hand, there is no identified need to collect or exchange an individual's complete personal naming history within the HE sector. The model aims to describe only the data, which are useful and relevant in the HE context.

The data model consists of entities, their attributes, and the relationships between them. These are widely used terms, but they are not necessarily familiar to everyone, so a brief discussion may be helpful:

- An *entity* is a thing that exists in the model. Entities can be real, physical objects such as people or buildings, but they can also be conceptual objects such as the outcome of a student's undertaking of a module.
- An *attribute* is a characteristic that an entity may have. For instance, the date a person commences a course cannot be an attribute of either the person (who may have started more than one course on different dates) or the course (which may have people starting it on a number of different dates). The HEDIIP model makes the commencement date an attribute of the Student Registration. The attributes which exist in the model are those which are relevant to the domain of the model – higher education curriculum and students – so for instance a person can in reality be either right handed, left handed or ambidextrous, but the person's handedness is not relevant to the domain of the model, so there is no attribute for handedness in the data model's Person entity.
- A *relationship* establishes how two entities can be associated. For instance, a module must be taught in one or more languages, and this is represented in the HEDIIP model by the relationship between a Module Instance and a Module Language.

What is a logical data model used for?

A logical data model is a map of the data in the domain of interest – in this case higher education students and curriculum. A map exists to stop people getting lost, and when different groups of people have copies of the same map it will be easier for them to find each other. It is clear from everyday experience that the current HE data landscape is rather poorly mapped:

- PSRBs which collect data directly from HE providers (which is most of them) often find that the data they collect do not match the data those same providers give to HESA. Because there is not a shared data model underlying HESA and PSRB collection processes, the differences often cannot be explained, and there is therefore no scope for PSRBs to rely on HESA data rather than undertake their own collections.
- Data presented for public information, such as league tables, key information sets and HESA Performance Indicators are difficult to relate to each other even when they ultimately draw on the same dataset, because of differences in approach to concepts like 'course' and 'subject area'.
- HEDIIP's data capability project found that data within HE providers are held, managed and used in silos. This approach brings with it duplication, confusion and additional cost leading to multiple copies of the same data, none of which everyone trusts.

Clearly, a logical data model cannot resolve any of these issues on its own, any more than a map can navigate. It is a tool to help people resolve issues by setting out a common approach to defining the key entities and their relationships.

How the HEDIIP data model works?

The HEDIIP data model is documented in detail. This section of this paper is intended to give a high-level explanation of how the model has been designed. Like any mapping process, the development of the model required decisions about what features need to be shown, and how to show them (road atlases don't usually have contour lines, and contour lines are not the only approach that can be taken to showing heights on a map). The principle in the model is that higher education happens when people undertake the curriculum offered by organisations.

- People are people. This is straightforward, and people are represented by a Person entity in the HEDIIP model.
- Undertaking is also fairly straightforward. Clearly enough when a person studies a course or module they create an academic history for themselves. This is represented in the model through a Module Instance for each module that person has undertaken, a Student Course Session for each course session they have undertaken, and a Qualification Award for each qualification they have gained. The academic history is not just a bunch of events, it needs to be organised into coherent programmes of study. Therefore the Module Instances and Qualification Awards are each associated with a Student Course Session, and the Student Course Sessions are each associated with a Student Registration – an entity which identifies the underlying contract between the person and the organisation which obliges the one to deliver an education and the other (usually) to pay fees.
- The curriculum is less straightforward. The HE sector has traditionally viewed the curriculum in two different ways simultaneously. There is a set of offers made in the prospectus, but also a structure for delivery, and often the two can be quite different (for instance a single 'course' in the delivery structure may be offered under many different titles in the prospectus). The HEDIIP model takes the view of the curriculum that is offered by the prospectus because the prospectus is publicly available, similar across all providers, and forms the basis of a contract with the student who applies for a prospectus offering. There are therefore a series of points in time and space (called Course Deliveries and Module Deliveries) where specific curriculum is offered to applicants, and these offerings are what students go on to undertake.
- Finally organisations are reasonably straightforward. The main organisations relevant to the data language are obviously HE providers, but other organisations are also relevant, such as funders, regulators and partners of those HE providers. The HEDIIP model sets out to be explicit about the roles that each organisation has in the delivery of curriculum so there are Course Roles, Delivery Roles and Curriculum Accreditations which unambiguously identify the roles which any one organisation takes in the organisation or delivery of any one course or module.

How the HEDIIP Data Model will be used

The HEDIIP data model will be published as 5* open data (see http://5stardata.info/en/ if you are not familiar with this concept); it will be available for a wide variety of uses by a wide variety of users which cannot necessarily be predicted in advance. However, there are three immediately planned activities:

- 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. HEDLSG will define a process for exercising that governance, and will also have responsibility for publication of the model in an open format which meets the criteria for 5* open data. The approach to governance will be open and collaborative, with an equal voice for data collectors and HE providers. The model will need to change in response to changes in the sector it is describing, and also in response to the needs and requirements of users, as they develop too so this governance function will be an important one.

2. Data Futures

Data Futures is the HESA project which will transform student data collection in the HE sector. A project to define a new 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. HEDIIP has undertaken analysis of the existing HESA returns and has a high level of confidence that the entities and attributes defined in the HEDIIP model will support these requirements. However, undertaking this further work is likely to identify improvements that are required in the model. HESA will work closely with HEDLSG to ensure that these requirements are identified and implemented in a timely fashion. HESA will also need to maintain strong engagement with stakeholders, especially HE Providers, which have a strong interest in this collection work.

3. Development

Having defined the process for governance, HEDLSG will need to lead the development of the model. The model requires further development in order to meet the needs of further groups of stakeholders. HEDLSG will allocate resources and agree priorities for this development work in light of the expected value of different areas of development.

How to find out more

The complete documentation for the HEDIIP model is available at https://hediip.ac.uk/he-data-language/.