Vision for college HE data ingestion and publication by HESA

Incorporating an outline statistical design and project mandate v.1.4.0

EXECUTIVE SUMMARY – OUR VISION

HESA will ensure that users can access and explore trusted data about all of the UK's HE students and their providers in one place.

175,780, or around 6.6% of the total number of enrolments on HE courses in the UK, are in Further Education (FE) settings. In this document we use the term 'college HE' to describe this concept. UK-wide research by the Office for Statistics Regulation (OSR) has found that 'users would find it helpful to have all higher education data in one place, as information about higher education students studying in further education colleges is not published alongside university higher education data.'[1]

On HESA's appointment as the designated data body in England, the OfS noted that:

> "HESA undoubtedly has a long and successful track record of collecting information directly from providers which is a key requirement, but it has less experience of processing data in bulk that has been independently collected by other agencies such as the ESFA *[Education and Skills Funding Agency]*. This is a high priority area that HESA would need to address. As the designated information body, it would be required to develop its processes for handling data that has been collected by other bodies and would need to improve its knowledge and expertise of these other sources of information." [2]

HESA's journey towards establishing these capabilities is underway, and, for example, HESA already uses student data drawn from external sources alongside its own collections, in the following areas:

- producing the UK wide figures on the totality of HE provision[3]
- conducting the survey fieldwork for the Graduate Outcomes survey[4]
- producing statistical outputs from the Graduate Outcomes survey[5]
- producing the Unistats Open Data[6]

[1] (Office for Statistics Regulation, 2020, p. 31)
[2] (Office for Students, 2018, para. 42)
[3] See: https://www.hesa.ac.uk/data-and-analysis/sb255/figure-3b
[4] See, for example, details of how the ILR data is used for FECs in England: https://www.hesa.ac.uk/innovation/outcomes/providers/information-english-further-education-colleges#contact
[5] Data can be filtered by 'provider type' to show just that data pertaining to FECs/FEIs – see: https://www.hesa.ac.uk/data-and-analysis/graduates
[6] See: https://www.hesa.ac.uk/support/tools-and-downloads/unistats

While the OSR acknowledges successes and progress in this area, its latest report recommends HESA undertake further work with 'further education funding bodies to capture higher education provision delivered in further education colleges in their statistical outputs'.[7] Having identified this area as a priority, this document explains our next steps for tackling this.

We start by identifying specific gaps in the provision of HESA's statistics. The most relevant gaps relate to a lack of provision of more detailed information about college HE students, alongside existing UK-wide student data, in the Student record statistical bulletin, which is a National Statistic, and in the associated Open Data outputs.

Users want to be able to find detailed information about the totality of HE students and provision in the UK in one place. For inquiring citizens, policy influencers, and 'information foragers', this would be best served through the Student National Statistics and Open Data products published on the HESA website. For expert analysts and more technical users this would be better served through the provision of microdata for use under license, deposited within an appropriate controlled environment designed for such purposes, such as direct supply to HESA's statutory customers, and for others via services such as the ONS's Secure Research Service or the UK Data Archive, and *via* products supported through HESA's agreement with Jisc. A data protection impact assessment has been initiated, with the objective of identifying an appropriate legal framework delineating the lawful purposes and data sharing agreements that will be required to finalize this vision statement, and to move the project into its resourcing phase.

In order to fill the gaps in data, HESA must ingest, process, and integrate data drawn from existing administrative FE data collections in England, Scotland, and Northern Ireland, for inclusion in its analyses on a comparable basis with data from its own collections. HESA will therefore aim to establish supplies of student microdata from the relevant data collectors, and develop routines for processing and quality assurance, ingesting this into its analytical framework.

This work will enable HESA to continue to support its existing uses of college HE data and address the gaps we have identified. It will also create substantial public benefit by curating and making available a single consistent, comparable, detailed source of information about all students in HE on a comparable, UK-wide basis. The publication arrangements will enable college HE providers and their students to benefit from existing third-party services and products, as well as increasing incentives for the development of new offerings.

This development project also offers opportunities to reduce burden, first by obviating the need for some of our statutory customers to produce and publish descriptive analyses covering their total HE populations in detail, and second by potentially reducing duplication of the government data collection requirements and private FoI requests on college HE providers. We intend to work with FE colleges, their representative organizations, and our statutory customers across the UK to identify where efficiencies can be found as a result of the successful delivery of this project.

The high-level vision for this project is therefore:

- To ingest college HE data from across the UK by agreement with the relevant governments and funding councils, developing appropriate processes and measures to incorporate it into HESA's datasets

[7] (Office for Statistics Regulation, 2020, p. 32)

- To utilise college HE data in HESA's publications and include it in technical documentation, as far as possible matching and extending the current uses of HESA Student data, to include information about college HE
- To identify what new processes, amendments to existing processes, and resourcing requirements systematic data ingestion will entail for HESA, as this domain of activity expands to meet expectations, and transitions to Business-As-Usual activity.

CONTENTS

INTRODUCTION TO OUTLINE STATISTICAL DESIGN AND PROJECT MANDATE

This is a foundational document for consulting and confirming stakeholders' needs for statistics and datasets about higher education that takes place in settings predominantly serving further education.

HESA aspires to collect and publish the most comprehensive data about higher education in the UK. The Student statistical bulletin is HESA's National Statistics product, which acts as the official annual census of the HE student population. While the enumeration is complete, we are currently missing details about the personal and study characteristics of students enrolled in college HE (except in Wales).

Because data is not currently available in one place, accessibility for users and clarity over what is available is constrained. Although some improvements could be achieved with better signposting, this would be a suboptimal approach.

HESA's basic enumeration of the totality of HE study across the UK is not currently matched with detailed statistics about it, where data has originated from FE colleges (except in Wales). We have counts of college HE students, but lack detailed information about their personal characteristics, and levels and subjects of study.

While we have successfully merged HESA-collected data on HE providers across the UK (including the formerly-titled 'Alternative Providers' in England, and college HE in Wales) into a single statistical bulletin, UK-wide data on college HE collected by other organizations is only available from HESA at a basic summary level. Headline data from collections by the relevant bodies in the devolved administrations is currently combined into a single table within our Student statistical bulletin[8], but is not disaggregated by any other variables. Lack of integrated UK-wide data also hampers the completeness of our Open Data, and the provision of integrated microdata to researchers, regulators, funders, and policymakers.

Comparability of data is constrained where HE data from FE-based data collections is concerned[9] as data standards differ between HE and the different FE collections. Our ambition is now to extend our existing published data to include comparable information about HE in FE settings where HESA is not the data collector.

This Vision Statement explains HESA's plans to integrate and publish comprehensive and detailed information on HE in FE settings across the UK. We focus on data about students. We acknowledge parallel issues about accessibility and clarity of supporting data about the learning environments in FE settings, such as the provision of data on Staff, Finance, Estates, Business and Community Interactions, and so on, but these are outside the scope of this document, and not discussed further here.

[8] See: https://www.hesa.ac.uk/data-and-analysis/sb258/figure-4

[9] The Office for Statistics Regulation have noted "growing concerns about the impact of devolution on the availability and comparability of HE data" (2019, p. 10)

SCOPE – IDENTIFYING NEEDS

We are indebted to the work undertaken by OSR, which has been re-affirmed by informal discussions with known stakeholders and potential users. The evidence for the value of the proposed enhancements is substantial enough to make the case for undertaking the activity overall. The first stage of the project will translate the broad aspirations laid-out here into a detailed design. In considering the kinds of users we wish to support, we here utilize the ONS' user personas[10]. These personas align quite well with the users that HESA typically supports.

For 'inquiring citizens', 'policy influencers', and other 'information foragers', we anticipate serving the majority of their needs through enhancing HESA's core outputs. HESA has existing evaluation and review practices that will enable iterative continuous improvement of its outputs as we learn more about uses of the data, following publication.

We anticipate supporting 'expert analysts' and 'technical users' with the production of processed microdata outputs and supporting user guidance materials such as methodology statements, quality assurance investigations, and metadata resources. This group includes HESA itself, its statutory customers, the ONS' Secure Research Service (under the Digital Economy Act) and our data analytics partner (Jisc). It also includes some of the data collectors for the data sources we seek access to, offering us an excellent opportunity to develop our understanding of the quality characteristics of the data and the needs of these users during the detailed design and initial experimental statistics production process.

HESA also has its own current needs for this data, which would exist even if there were not a compelling external case for development of the collection. This project offers HESA an opportunity to optimize a group of processes that are currently scattered across several teams and managed separately and potentially inconsistently. These are:

- Producing high-level aggregate figures for the 'totality' of HE for the Student outputs
- Obtaining population data for the Graduate Outcomes survey sample
- Obtaining detailed characteristics data for the Graduate Outcomes outputs
- Producing the Unistats outputs.

While the potential savings are probably relatively modest even in the longer term, they are not the main focus of this proposal. Instead, the proposal focusses on efficiency and rationalization of process, appropriate management and stewardship to safeguard and enhance our reputation and data assets, and the development of knowledge and skill in our staff.

We are currently ingesting separate files comprised of the same underlying data, prepared by others, at different times, in different teams, to support different activities. However, there is a substantial qualitative benefit, in that our confidence in these processes will grow as our familiarity with the quality characteristics of the underlying data increases. This could help us to avoid quality issues and reduce the risk of breaches. Our current dependencies on a complex and only partly understood data lineage will also be reduced, allowing us the scope to foresee and avoid quality issues with data that are discovered late in the process (e.g. survey problems with the NI college HE population). This understanding will also have a direct benefit to users, for

[10] The ONS' user personas are described in detail at: https://style.ons.gov.uk/category/writing-for-the-web/personas/

whom we will be able to publish new information about data quality, along with details of our processing arrangements (e.g. derivations), in line with our established practices.

Desiring to minimise the burden of additional data collection on college HE providers who already complete substantial accountability returns for their primary regulators, we will look to re-use existing data. All existing public sources of data about higher education would be compiled into a single UK-wide dataset, with the aim of making possible more comprehensive description, comparison, and evaluation, of Higher Education.

HIGH-LEVEL OUTPUT OBJECTIVES

Our central research question, which the proposed programme of work would seek to answer is: "To what extent is it possible to establish a comprehensive, accessible, clear, and consistent statistical view of the students in higher education in the UK, from existing college HE data sources?"

We therefore aim to produce a new, combined UK-wide dataset, which would (pending legal agreements) be made available through HESA's normal dissemination routes.

We aim for developmental work to be published during 2022, leading to full publication as experimental statistics alongside (or shortly after) the HESA Student outputs in 2023, with full integration pending the removal of experimental status. Microdata is intended to become available from 2024 and with other potential uses following from there.

The main ongoing uses identified for the data are (subject to the necessary agreements):

1. Enhancement of HESA's Student statistical bulletin National Statistics product, and associated official statistics uses
2. Making detailed college HE data sourced from collectors in England, Northern Ireland, and Scotland available as Open Data on a broadly comparable basis to data collected by HESA for college HE in Wales
3. Provision of microdata to our statutory customers
4. Provision of microdata to ONS' Secure Research Service for Digital Economy Act uses
5. Provision of microdata to our analytics partners (Jisc) for use under license in agreed data products
6. Quality assurance activity, including de-duplication, and consistency checking between HESA and ingested third-party observations of attributes where individuals are represented in both datasets, and the production of survey weights
7. Production of relevant metadata and explanatory technical information to support use and interpretation of the data
8. Linking records for longitudinal tracking between FE and HE
9. Production of Graduate Outcomes sampling frame for those FECs subscribing to the survey
10. Production of the Graduate Outcomes outputs (where some linked data is already incorporated in analysis)
11. Exploring the potential for producing the Unistats output directly from integrated ILR data, reducing the requirement for a supply from the OfS
12. An aspiration for the establishment of a (non-/)continuation marker that is consistent across HE, and takes into account FE to HE transfers, using an agreed methodology such as that used in Unistats
13. Exploring the potential for future cost-sharing and integration of data quality assurance and enhancement approaches among our statutory customers.

HESA would aim to retain the data in the long term for statistical and research purposes. We would aim to do this in a manner broadly consistent with the well-established arrangements already in place for the data HESA has collected directly.

DATA AVAILABILITY - SOURCES

In order to achieve this aim, HESA intends to seek access to the following datasets, each of which contains information on HE in FE settings:

- In England, the Individualised Learner Record (ILR) collected by the ESFA, and its counterpart dataset, the Learning Aim Reference Service (LARS)
- In Northern Ireland, the Consolidated Data Return (CDR)
- In Scotland, the Further Education Statistics record (FES)

These sources of data would complement the HESA Student and Student Alternative collections which together hold all data for HE providers in the UK, including those FEIs delivering HE in Wales.

Availability of data will be determined by the establishment of a lawful use under the terms of the Data Protection Act 2018 and the GDPR. For the purpose of this Vision Statement, we sketch the situation as we currently see it as follows.

All uses related to the production of enhanced outputs are public tasks in the public interest, and required for statistical and research purposes, pursuant to the various HE and official statistics legislation under which HESA operates. This encompasses HESA's official statistics production, and the provision of microdata to HESA's statutory customers, and to the ONS under the Digital Economy Act. These uses all align with purposes 1-4 described in full at: https://www.hesa.ac.uk/about/regulation/data-protection/notices
Sharing this data with our analytics partners (Jisc) is aligned with our purpose 5 at https://www.hesa.ac.uk/about/regulation/data-protection/notices. We intend to extend the existing framework of information provision under license, and the incorporation of the new data within existing data governance structures as specified on our website, and in existing legal agreements. Pursuant to this general approach, we will establish agreements delineating appropriate onward use categories with FE Colleges, for the data they have supplied via intermediary bodies. This mirrors an approach we have developed and deployed successfully for the Graduate Outcomes survey.

Exploring and refining this approach to meet the aspirations of our users will be an initial task for the next stage of the project. We will work constructively with all the data collectors supporting this project to achieve this, and to ensure the processing information they provide to data subjects is updated to support these extended uses.

In England, specifically, we consider the proposed enhanced outputs to be publications under s.65 of the HERA. We consider the compilation and making available of the processed data to the OfS, UKRI, and the Secretary of State to fall under s.64 of HERA. We respectfully request that the OfS and the Secretary of State formally co-operate to support HESA in the sharing of information between the governments of the UK home nations under s.63 of the HERA.

OUT OF SCOPE

Attempts to ingest and analyse in-year data collected from FE providers are out of scope for this work at present. As HESA's approach develops, there may be a future benefit in extending the ingestion to include in-year college HE data. We know that the provision of in-year data for college HE is a requirement of future data collection by HESA in Wales, and recent publications by the SFC demonstrates that in-year data is a need for the tertiary sector. Subject to the DfE's

review, some needs for in-year data have been identified by the OfS. HESA's Data Futures programme is responding to these requirements, but college HE is currently outside HESA's collection constituency (except in Wales).

This project therefore complements and reinforces the more flexible collection instruments being delivered by the Data Futures programme, by developing insight and skill in working with data derived from prevailing FE data collections in Scotland (FES), Northern Ireland (CDR) and England (ILR plus LARS). Putative future requirements for handling or collecting in-year college HE data (outside Wales) that are made technically possible by Data Futures, will be more manageable with this skill and knowledge in place.

It would be advisable to re-visit the scope in the 2023/4 academic year, as we are aware that some of the proposed uses within Graduate Outcomes may require data from earlier in the annual collection and quality assurance cycle.

DATA AVAILABILITY - TIME SERIES AND UPDATING

Ingestion of the final signed-off version of each dataset, as used for statistical publications by other producers, is the goal of this project.

HESA aims for the regular ingestion of this data to be a long-term addition to the datasets it curates, and will aim to build repeatable pipelines for the processing and analysis of the ingested data.

In the first instance, we seek finalised data from the 2018/19, 2019/20, and 2020/21 cycles to undertake initial investigations. This should be adequate to capture the majority of current college HE students from their point of entry, and provides an appropriate balance between completeness of the data and an efficient use of resources. We then aim to supplement this with data from 2021/22 and subsequent cycles on an ongoing and regular basis.

OUT OF SCOPE

Once we have established workable routines for the most recent data, we would be in a position to evaluate the costs and benefits of processing historic datasets in the same way. However, this work is out of scope for this proposal.

HESA's (current) future plans also include the publication of in-year student data. Following the successful integration of recent college HE data, we would be keen to explore the feasibility of comparable in-year ingestion of early data returns, including those that have not been finalised, in order to produce comprehensive in-cycle data from both collected and ingested sources. This would be aligned with prevailing timeliness of HE data at the time this becomes a feasible possibility. However, this work is outside the scope of this proposal.

IDENTIFICATION OF COMMON CONCEPTS AND QUALITY EXPECTATIONS

HESA would aim to ingest data from the above sources, and compile them into a single dataset, producing derived values that permit coherence in definitions, regardless of source, where possible. Our goal would be to establish, where possible, comparable standards for concepts that are applicable across UK HE, regardless of location of provision, or the entity providing it.

We anticipate that this will require extensive integration work to:

- review data and assure it for completeness, validity, uniqueness, and consistency.
- classify and recode data
- check the validity of our analysis by producing files and totals to match those supplied to us elsewhere, and to engage in collaborative quality assurance discussions and activities with the data collectors (where desirable for both parties).
- To come to decisions about the appropriate editing of data where quality problems have been identified. This will include working collaboratively with those tasked with the governance of data amendments at all the collectors, and may (potentially) involve checking the validity of our analysis with the original data submitters (FE providers) subject to ethical consideration regarding burden on providers and the proper application of the Code of Practice for Statistics, as well as the availability of a suitably-resourced technology and service platform to support this.
- derive new variables and units that allow comparison across the dataset, and describe their quality characteristics.
- Catalogue the gaps in datasets to support UK-wide data improvements.

We would also need to calculate aggregates and either check against published figures, or where published figures are unavailable, ask for assistance from the suppliers of the data to ensure accuracy and consistency. In the case of the OfS we have already established that we will want to establish a subsidiary data sharing agreement that enables us to share data and analysis in support of consistency, and we anticipate that some similar arrangements will be beneficial in Scotland and Northern Ireland. We will also aim to provide advice to the original data collectors on potential ‘upstream’ quality enhancements, as a result of our analysis.

QUALITY REQUIREMENTS

These are overarching descriptions of the aimed-for data quality during the first phase of delivery, to Spring 2022:

DAMA quality dimension	**Description of high-level quality process**
Timeliness	Utilise finalized data from the CDR in Northern Ireland and the FES in Scotland. In England this will be the ILR R14 data for recent years. The intended output cadence is annual, aligned as close to the Student outputs as feasible.
Consistency	To match data supplied to HESA by the OfS to create the Unistats Open Data product, and explain any differences.

	To provide HESA's own estimates for Student Table 4 consistent with the figures supplied by the data collectors. To provide estimates for various tables in the Student outputs that are consistent with figures published elsewhere, e.g. http://www.sfc.ac.uk/publications-statistics/statistical-publications/statistics-schedule/statistical-publication-schedule.aspx To identify logical consistency of key variables required for outputs.
Completeness	Dataset level: To enumerate a total number of HE students in FE consistent with the OfS', DfE-NI's and SFC's totals, and explain any differences. Attribute level: To identify the extent of missing data for key variables.
Validity	To run a set of cross-tabulations for key variables, with a view to develop a set of validity checks that could be used to support ad-hoc or end-of-cycle quality assurance. To learn more about the validity checks built-in to the original administrative data collections, in order to understand quality features of the data and communicate these to end users.
Uniqueness	To identify a concept equivalent to an Instance from within the ILR data, and de-duplicate, to enable a logically-consistent student population to be derived from the ILR and HESA datasets. To de-duplicate between the HESA and ILR datasets, utilising the most robust values available.
Accuracy	We will rely on the final signed-off versions of the data for this project. Data quality issues discovered in analysis will be raised with the appropriate bodies. A working protocol between HESA and the data collectors will be developed to determine cases where editing (or resubmission) of the data will take place, with reference to existing data governance approaches in place for this, such as the OfS' Data Amendments Panel. Residual quality concerns would be communicated through a process based on HESA's current practice of producing Data Intelligence notes.

Our goal would be to publish experimental statistics once a reasonable level of quality has been attained, and this would be on a similar timescale to the current Student outputs, details to be confirmed.

At the point of output, we would look to observe the following quality characteristics, and note that experimental statistics offer an opportunity to assess the extent to which we have met the needs of users:

ESS quality dimension	Output quality goal description
Relevance	We are aiming to fill a known gap in information provision (as described elsewhere), so we anticipate an improvement in the relevance of HESA data to users. We will consult with users to determine how useful the data is, through the normal means we use for evaluation and review, and including users with interests in the FE sector.
Accuracy and Reliability	These are administrative datasets. We will work with the data collectors to understand how accuracy has been assessed, and ensure this information is available in the user guidance that accompanies the publication. An administrative data quality assessment will be undertaken to support user understanding of the quality characteristics. Assessments of reliability prior to publication will require comparison with similar analyses by the original collectors, to check our analyses match. We will need to attempt to produce files and totals from our own analysis, which match the files and totals ingested; for Graduate Outcomes surveying and outputs, for Unistats, and for the HE totality figures. We will explain our approach to users in a guide that explores the quality characteristics of the data.
Timeliness and Punctuality	Our goal is to run the analysis in parallel with our existing processes for one year, to permit the quality characteristics to be evaluated without jeopardizing our production processes. Our output timescale will therefore aim to follow fairly closely our standard calendar. This means an experimental Student output during 2022/23, with an annual cadence thereafter.
Comparability and Coherence	We will need to compare cross-tabulations of key variables in the HESA Student data and the ingested datasets, to identify significant variations. Since we are intending to use administrative data already published elsewhere, our approach to comparability is largely covered under Accuracy and Reliability. On coherence, we will need to establish the extent of semantic interoperability between concepts across several datasets. Where concepts align, we will aim to integrate attributes from all datasets. Where concepts differ, we will need to make decisions about the extent to which processing can produce derived values that cohere. We will explain our approach in user guidance alongside the Experimental Statistics. However, on assimilation of the Experimental outputs into the main Student output, we will produce a full ‘coding manual’ type publication to support users. We will also produce derived attribute specifications where required, according to our standard metadata practices.
Accessibility and Clarity	There will be user guidance produced to accompany the Experimental Statistics. As our confidence increases, and experimental outputs are assimilated into our standard outputs, we will provide quality information and metadata following our usual practices. We will apply our usual

	dissemination practices, following the prevailing HESA style guides. Data will be published as Open Data and licensed data made available using the prevailing dissemination channels to Statutory Customers and other users, such as our data analytics partners at Jisc.

We are planning a major update to the current quality report on our Student outputs, in line with the timescales for the Data Futures project. That revised quality report will take into account the ingested college HE data covered in this document, along with our new approach to collected data. We will use the ESS quality dimensions and the prevailing version of the ONS' Administrative Data Quality Assurance toolkit as the frameworks for this document.

OUT OF SCOPE

Checking credibility of data with FECs prior to submission to their main data collector (where this is not HESA); or seeking to edit existing values, excepting where an amendment has been approved via an agreed data governance process owned by a statutory customer or HESA, and where a re-submitted data file is available to HESA.

Other data from the FE sector (e.g. Staff or Finance data) is excluded from the scope of this project, but could be considered in future, subject to user needs and data availability where HE-related concepts can be identified, discretely.

This work does not form an addition to the Data Futures programme, but utilization of new technology and standards developed through that programme will be the first preference choice for undertaking the work to share and confirm analysis with college HE providers during the 2023/4 academic year, to maximize efficiencies over the longer term, and to simplify the establishment of college HE ingestion as a regular Business-As-Usual activity.

BUSINESS CASE

Below we look at the indicative roadmap, likely costs, and benefits and beneficiaries, in outline.

INDICATIVE ROADMAP

We believe it is more important to do this work well than to do it hastily. We believe the plan outlined below represents a sufficiently challenging, but achievable timescale.

Stage 0 - preparation: autumn 2020 to spring 2021

- Identify needs for data and policy background
- Develop Vision Statement
- Develop outline statistical design and project mandate (business case)
- Identify research question
- Internal discussion about feasibility (inc. legal)
- Resourcing approach identified
- Decision to proceed to resourcing

Stage 1 - resourcing: spring-autumn 2021

- Initial work with suppliers (ESFA, SFC, DfE-NI; plus DfE and OfS) to design HESA overall approach, gather data dictionaries, documentation etc.
- Data Protection Impact Assessment and identification of appropriate legal framework and initial discussions over DSAs to meet 6 DP principles
- Confirm specification, consult with suppliers/users, refine data flow descriptions and timescales of data availability
- Recruitment to specified mobilization role onboarding and training
- Detailed refinement and estimation of tasks
- Non-human resource requirements specified and procured

Stage 2 - initiation: spring 2022

- Further recruitment plans developed and initiated
- Investigation of existing published materials about data sources
- Detailed design phase for ingestion, processing and analysis
- Application forms for data specifying field-level requirements written
- Legal implement DSAs and update FPNs
- Data supplies arranged
- Legal begin work on framework for onward use and update of processing notices

Stage 3 – ingestion: summer 2022

- Access real data
- Build ingestion system and workflows
- Build/configure analytical framework, software, quality assurance approach and expected approach to producing metadata and technical documentation
- Test and refine design and approaches
- Ingest into warehouse
- Initial exploratory analysis and refinement of analytical goals
- Establishment of team relationships with data suppliers
- Legal continue to progress framework for onward sharing

Stage 4 – exploration: summer 2022 – spring 2023

- Exploratory analysis within Agile Sprints, regular Sprint Reviews
- Quality report developed iteratively
- Definition of derived data specification/standards developed iteratively
- Detailed design for outputs (develop wireframes)
- Detailed design for supporting technical documentation (user guide, methodology statement, quality report, coding manuals) and experimental statistics outputs

Stage 5 – evaluation and transition to BAU – Spring 2023 – Spring 2024

- Publication of experimental statistics in Spring 2023 (no provider-level data at this stage)
- Development of preview process for Colleges based on utilizing the HDP engine
- Establish output data file formatting requirements for onward supply of microdata
- Evaluation of project and learning lessons
- User consultation to support future development
- Publication of experimental statistical outputs in spring 2024 including provider-level data); along with aligned supply of 'signed-off' microdata to statutory customers,
- Put in place legal framework for sharing data with our analytics and data supply partners (Jisc, ONS)
- Further development phases identified and business cases developed where necessary e.g. :
    - Identify pathway for removal of experimental statistics label and additional integration options within Student Statistical Bulletin National Statistics product and associated Ad-hoc statistical release and Open Data products
    - Reproducible analytical pipeline development and streamlining
    - Evaluation of options for enhanced integration of collection and/or quality assurance activities with statutory customers, and increased use of the Data Futures deliverables
    - Investigation of desirability and feasibility of publishing in-year college HE data
    - Investigation of desirability and feasibility of ingesting historic data
- Celebrate and effect transition to BAU operations.

RESOURCING

Considering the scale and scope of this work, it will necessarily draw on skills and knowledge from across the statistical business process, particularly focussing on the processing, analysis, dissemination and evaluation practices, and on the data governance, quality assurance, and metadata management overarching processes.

A successful business case was developed and submitted for approval by the HESA Board.

BENEFITS AND BENEFICIARIES

In considering the resourcing for this work, it may be instructive to consider who benefits from it, and how. The following table summarizes this:

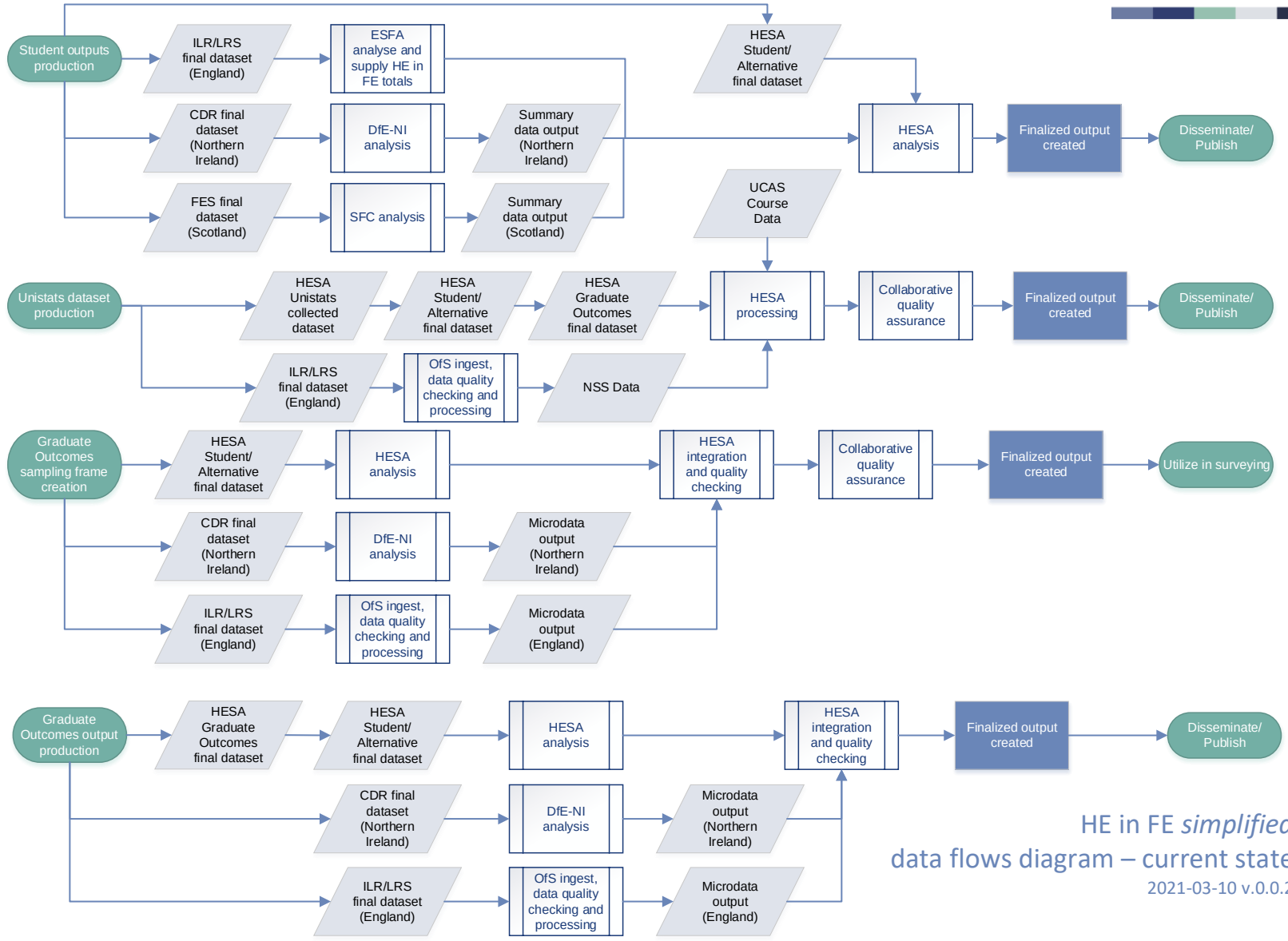
The OfS	Currently undertakes this work. HESA could offer analytical support as a quality assurance partner. OfS has signalled a strong desire for HESA to demonstrate progress in this area, from the point of appointment as DDB, and has reiterated this recently in a report to the OSR. Interest in project's potential for long-term efficiency and productivity gains.
The DfE/ESFA	This project would offer an example of the usefulness of ESFA data, and offer potential benefits in HESA being able to provide quality assurance insights that could help improve data collection practices. HESA and ESFA have worked together in the past, and this project could provide an opportunity for exploring areas of future collaboration for mutual advantage. DfE have previously indicated support for the vision of having all UK HE data in one place.
The SFC, DfE-NI, and the DfE	This activity could potentially obviate the need for some existing publications, which could open up the possibility of creating efficiencies, either reducing their analytical costs, or concentrating their time on higher-value outputs. It may also provide quality assurance insights that could help improve data collection practices. Potential interest in project's potential for long-term efficiency and productivity gains.
College HE providers	Policy analysis indicates that there is likely to be an increased demand for data about and within the FE sector in future. The FE sector in England currently enjoys cross-Party support under the skills agenda. In Scotland a more coherent and integrated approach to data collection is being explored as a strand of the HE review. In Wales an integrated approach to HE data has been established already and the forthcoming HER Bill provides an opportunity to deepen the use of data in driving better social outcomes. The provision of better data is likely to be a necessary cost, and potential benefit from, this attention, and FE providers are likely to find benefit in improved data and associated capabilities. Now is a good time to address the provision of valuable statistics and data to the FE sector in support of educational missions.

Data Analytics partner (Jisc)	Policy analysis indicates increasing needs for research into FE, requiring access to tailored datasets for researchers, currently provided under license from HESA under an agreement with Jisc. Jisc also appears to identify the FE sector as a target for improved services. This project would offer a unique and powerful opportunity to extend established products developed for the HE sector to the FE market. The development of a Heidi Plus service for FECs, and enhanced tailored datasets offerings would be examples of value creation with this data.
Other licensed microdata services (ONS and UKDA)	Researchers of HE and FE may also wish to deposit analyses and data based on with secure services such as the UK Data Archive (powered by Jisc). Legal arrangements should not preclude this development. Under the Digital Economy Act, the ONS is developing a sharing framework for government users, utilizing its secure research service (SRS). We already anticipate supporting this development in due course, and would look to include our processed microdata drawn from administrative collections in FE into the SRS.
HESA	First, HESA has already been asked to undertake this work by the OfS, and the OSR's reports implicating HESA also requires an appropriate response. Second, HESA is already undertaking elements of this work in a way that is not joined-up, is probably costing more than it ought, creating some additional risks, and it does not result in optimal outcomes. This project enables us to rationalize our resources in an appropriate way. Third, this project is on the critical path for desired improvements to the HESA data outputs relating to Graduate Outcomes, Student, and Unistats. Fourth, this project develops capacity where it is most sorely needed for future service development, and represents an investment in HESA's future viability.
The OSR	Has invested in a systemic review directed at making positive change to the statistical arrangements for post-16 data. HESA undertaking this work addresses concerns identified by our regulator and demonstrates the effectiveness of the OSR's approach.
End users	We have identified elsewhere in the document that inquiring citizens, policy influencers, and 'information foragers' would find it useful to have information about all of UK HE in one place, and they will benefit through being able to access more detailed information about the totality of HE through the HESA website. Expert analysts and technical users will be able to access microdata under license *via* services provided by Jisc and

	others. All users will be well-supported by a suite of published materials produced by HESA and made available via its website.

HESA

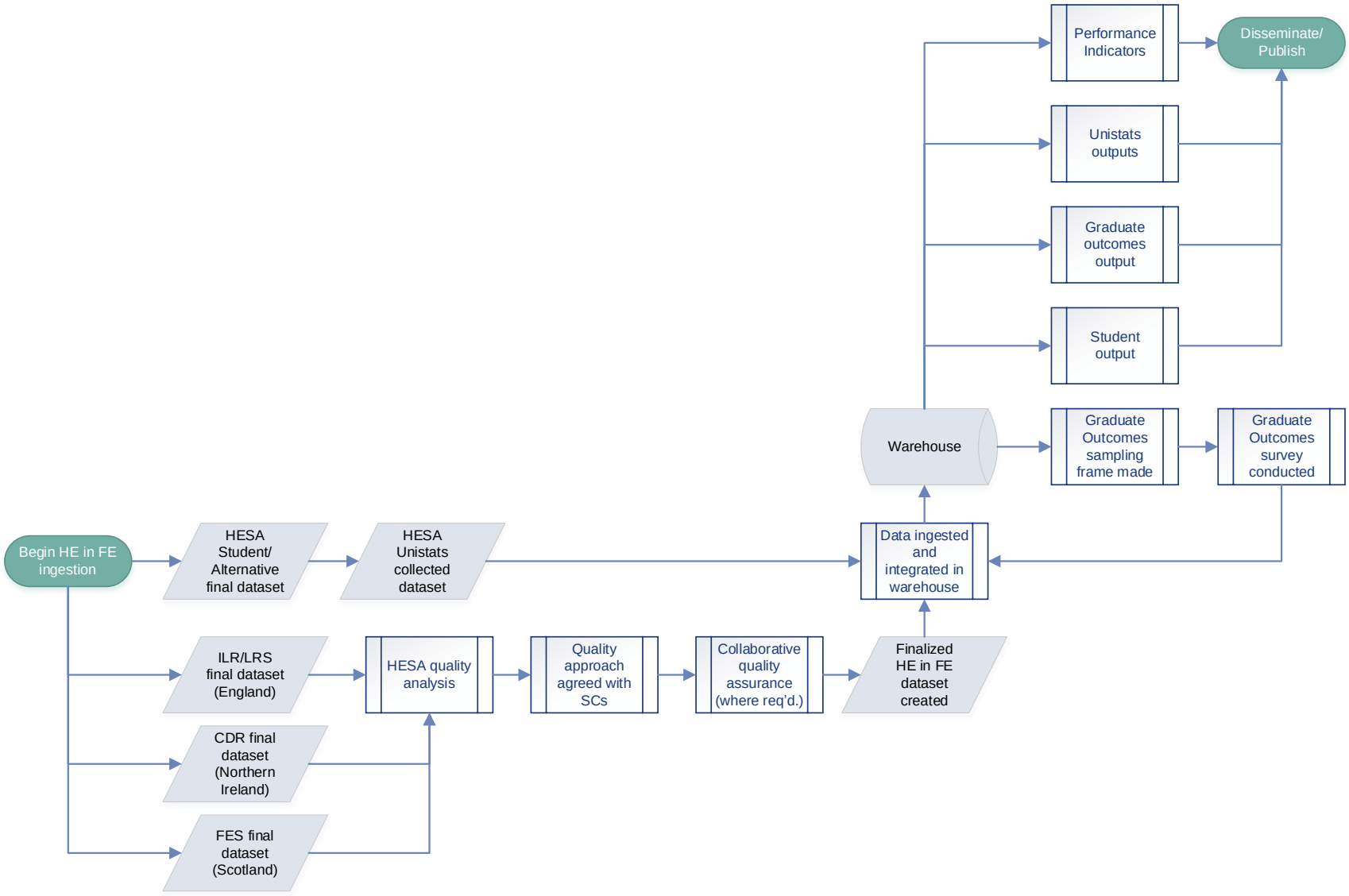
COLLEGE HE SIMPLIFIED DATA FLOWS DIAGRAM – CURRENT STATE



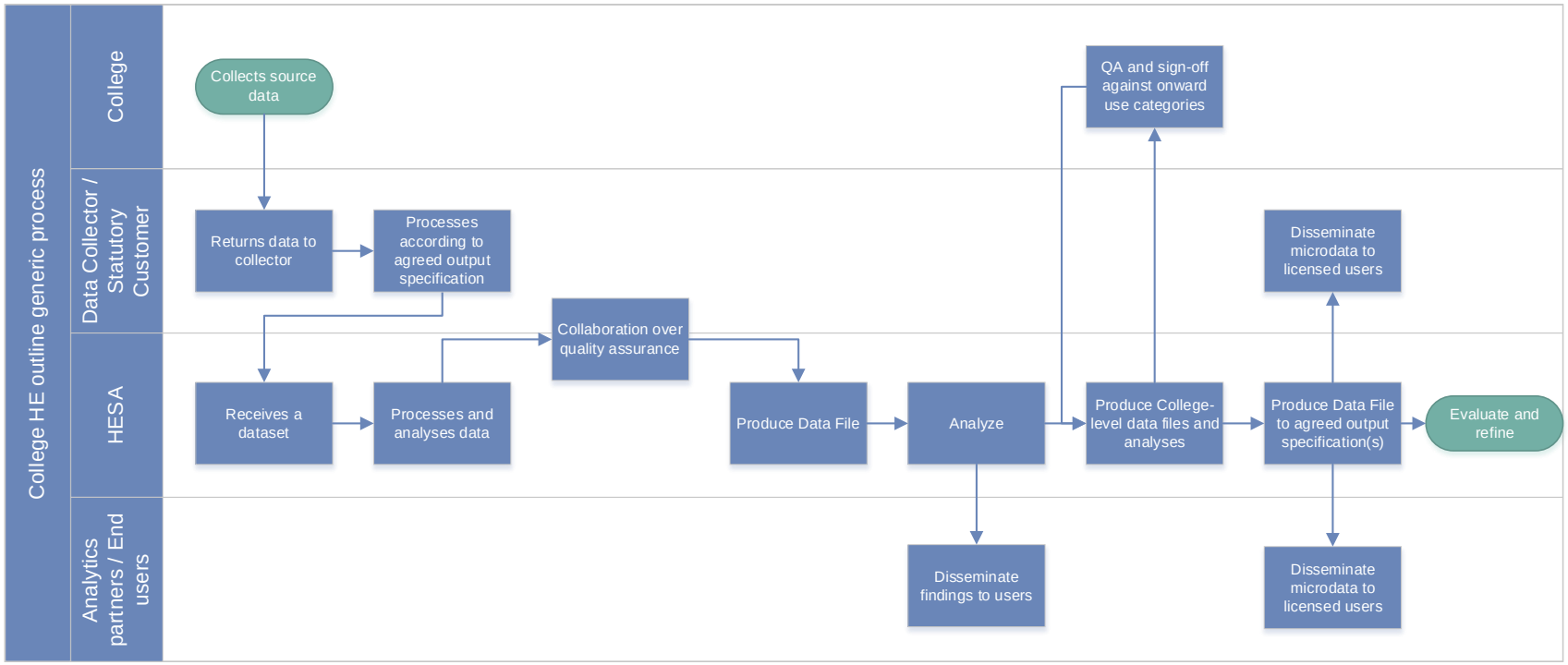
HE in FE *simplified*
data flows diagram – current state
2021-03-10 v.0.0.2

Protective Marking - PUBLIC

COLLEGE HE SIMPLIFIED DATA FLOWS AT MATURITY – FUTURE STATE



DRAFT DATA FLOWS PROCESS



REFERENCES

Office for Statistics Regulation. (2019). Public value of statistics about post-16 education and skills in England. 22. https://www.statisticsauthority.gov.uk/publication/exploring-the-public-value-of-statistics-about-post-16-education-and-skills-in-england/

Office for Statistics Regulation. (2020). Exploring the public value of statistics about post-16 education and skills - UK report (Systemic Review Programme No. 2; Post-16 Education and Skills). Office for Statistics Regulation. https://osr.statisticsauthority.gov.uk/wp-content/uploads/2020/07/Exploring-the-public-value-of-statistics-about-post-16-education-and-skills-UK-report.pdf

Office for Students. (2018). Designation of a quality body and a data body. https://www.officeforstudents.org.uk/media/1325/bd-2018-jan-24-designated-bodies-paper.pdf