HEDIIP NSCS Project -

The Higher Education Classification of Subjects (HECoS) vocabulary

About the New Subject Coding Scheme Project

The New Subject Coding Scheme project was commissioned by HEDIIP under the Standards and Understanding theme. The project aimed to develop a replacement for the Joint Academic Coding System that meets the needs of a broad group of stakeholders and reflects the diverse and dynamic nature of Higher Education in the twenty-first century. The New Subject Coding Scheme Project was undertaken by the Centre for Educational Technology, Interoperability and Standards (CETIS) with partners Aspire Ltd and APS Ltd in two phases between July 2014 and October 2015. The project undertook extensive stakeholder engagement to identify the requirements for the new coding system and developed a coding structure that aims to meet these requirements. The new coding scheme is referred to as HECoS – the Higher Education Classification of Subjects.

The project ran from May 2014 to October 2015. – *this is omitted from this document.*

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

- Andy Youell, Director, HEDIIP
- Dan Cook, Head of Collections Development, HESA
- Dr Christine Couper, Director of Strategic Planning, Greenwich University
- Hannah Falvey, Head of Statistics, HEFCW
- Lesley Donnithorne, HR Manager (Systems, Information and Grading), UWE Bristol
- Mike Spink, Data Architect, UCAS
- Paul Baron, Programme Manager, HEDIIP
- Jenni Cockram, Programme Officer, HEDIIP
- Principal Authors/Editors: Wilbert Kraan and Alan Paull

Contributors: Gill Ferrell, Lorna Campbell, Phil Barker, Adam Cooper, Charlie Paull and Jennifer Denton

About HEDIIP

The Higher Education Data & Information Improvement Programme (HEDIIP) has been established to redesign the information landscape in order to arrive at a new system that reduces the burden on data providers and improves the quality, timeliness and accessibility of data and information about HE.

HEDIIP is funded by the Higher Education Funding Council for England (HEFCE), the Higher Education Funding Council for Wales (HEFCW), the Scottish Funding Council (SFC) and the Department for Employment and Learning (DEL) Northern Ireland.

HEDIIP is hosted by the Higher Education Statistics Agency Ltd (HESA) which is a company limited by guarantee, registered in England at 95 Promenade Cheltenham GL50 1HZ.

Contact HEDIIP

Web: www.hediip.ac.uk
Email: info@hediip.ac.uk
Twitter: @HEDIIP

Acknowledgements – *only appears in this doc*

We would like to thank all colleagues who have taken the time to respond, critique and work with us. The work has only been possible because so many gave us the benefit of years of experience over the course of designing the new subject coding scheme.

About This Report

A report on the development of the Higher Education Classification of Subjects vocabulary, with an account of feedback on it by stakeholders, and how the feedback has been used.

Authorship and Status

Owner	Wilbert Kraan
Principal Author/Editor	Wilbert Kraan and Alan Paull
Contributors	Lorna Campbell, Charlie Paull, Phil Barker, Jennifer Denton, Adam Cooper
This Version	Final for acceptance
File Identifier	HEDIIP_NSCS_PD04_Scheme_2015-11-23.docx

Document History

Date	Person	Notes
2015-07-10	Wilbert Kraan	Internal work-in-progress draft for PMO
2015-07-20	Wilbert Kraan	First version, addressing PMO comments
2015-07-22	Wilbert Kraan	Second version, addressing additional internal comments
2015-08-09	Wilbert Kraan	Third version, addressing addition feedback
2015-08-10	Wilbert Kraan	Minor file references update
2015-08-14	Wilbert Kraan	Minor file references update
2015-08-14	Wilbert Kraan	Minor file references update
2015-10-06	Wilbert Kraan	File references update and correction of HESA “course”nomenclature.
2015-10-08	Wilbert Kraan	Minor file references update
2015-10-09	Wilbert Kraan	Minor file references update
2015-10-15	Wilbert Kraan	Post NSCS Project Board edits to submission process diagram, module inclusion and course nomenclature
2015-11-06	Adam Cooper	Revised according to Advisory Panel feedback
2015-11-23	HEDIIP PMO	Update of cross references
2015-11-27	Wilbert Kraan, Alan Paull	Minor copy edits

Contents

1. Executive Summary

The New Subject Coding Scheme (NSCS) Project was tasked with developing a replacement for the Joint Academic Coding System, a vocabulary used to classify courses, modules and other entities in UK Higher Education. This report describes the second stage of the project, and presents the first iteration of the new scheme, the Higher Education Classification of Subjects (HECoS) vocabulary.

Where JACS3 was a widely used large scheme that gradually developed over time, HECoS is a new vocabulary built from JACS3 with unused terms removed, some classification aids added, and focused on clearly distinguishable subjects.

At the beginning of stage two of the NSCS project, a candidate scheme for HECoS was published, and feedback solicited. The main areas identified for further work were:

- Refine acceptance criteria for terms in order to find the right balance between analytical detail and reliable coding.
- Incorporate the expertise of subject specialists.
- Refine the process for the accession of new terms.

A wealth of comments on HECoS as a whole, as well as on particular terms in it, helped the team to address these and other issues.

One of the areas that attracted a lot of comment was the fact that HECoS, unlike the JACS3 scheme which it succeeds, is designed to work with many hierarchies by separating the terms from the way they are aggregated. This separation allows HECoS to fulfil more diverse functions in a coherent and structured way, but it also means that one widely agreed analytical hierarchy is essential to maintain comparability of analyses. Coordination is also needed to avoid a needless proliferation of hierarchies for similar purposes. This is the subject of a separate report (Recommendations for Subject Based Analysis & Text Mining, Cooper, 2015).

Another important area of comment on the HECoS vocabulary, however, remained the number of terms in it, and the concomitant tension between the vocabulary's provision of fine analytical detail and classifiers' ability to code to such detail reliably and consistently across the whole UK Higher Education sector. Research showed, however, that a vocabulary that is too small can have a similar trade-off in that it becomes increasingly unclear to which broad subject a particular course[1] or module belongs. For that reason, HECoS was constructed around demonstrable distinguishability of a subject as the main criterion.

Most comments were received on the cost to organisations of changing from JACS3 to HECoS. Most of such cost is inherent to any change from JACS3. Nonetheless, a significant continuity with JACS3, improved ease of use through distinguishability and being able to reuse a course and module classification for more than one data return are all designed to help ease the cost of the transition, and save money in the future. Additional measures to aid adoption are outlined in the HECoS Adoption Plan (Ferrell and Campbell, 2015). Finally, by providing more consistent and distinguishable coverage of the wide range of subjects provided in the UK, we tried to ensure that data classified with HECoS will have more value for all stakeholders.

[1] This report uses "course" in the sense of a complete programme of study that leads to an award

2. Identified areas of tension

2.1 Multiple purposes

The uses of subject coding vary greatly, from the allocation of extra funding for Strategically Important and Vulnerable Subjects (SIVS), to benchmarking, to course discovery, to league table construction and more. Because each of these functions require subjects to be aggregated in different ways, different aggregations and hierarchies are currently being used across the sector, although JACS3 was originally designed to have a single hierarchy. At the same time, in order for a common vocabulary, such as HECoS, to have value for the purpose of analysis across the sector, data coded with it needs to be comparable, and this requires an agreed and consistent way of aggregating subjects. Finally, some users also expect subjects to have a hierarchical relation to each other, and appreciate a hierarchical representation to guide term discovery.

In order to balance these countervailing forces, HECoS has been designed to be non-hierarchical as this de-couples the definitions of the subjects of study from the variety of structures which users may need to overlay upon the scheme. Stakeholder consultation has indicated that there is strong cross-sector support for a common (default) aggregation framework to be defined alongside HECoS and for it to be strongly governed. The proposal for establishing this common aggregation framework is outlined in the Recommendations for Subject Based Analysis & Text Mining report (Cooper, 2015). The HECoS Governance Model (Campbell and Ferrell, 2015) also outlines governance procedures to minimise the number of statutory aggregations.

In order to aid term discovery, HECoS is presented using a navigation hierarchy that groups related terms together to help users to locate the appropriate terms they require. These groups cannot be used for coding subject of study, and it is not assumed that they should be the basis for any mapping, aggregation, or association; they exist purely to aid the discovery of terms.

2.2 Size and resolution

The question of the size of the HECoS vocabulary, and thereby the degree of detail it can capture, has been the most constant source of tension since the start of the project. On the one hand, organisations tasked with tracking vulnerable and new subjects would like the greatest amount of detail, while those who supply the data prefer a simple, compact vocabulary with clear distinctions.

To balance these requirements, HECoS focuses on term distinguishability as the main characteristic of the vocabulary. A vocabulary that has too many terms will compromise distinguishability and lead to poor quality data for analysts, unacceptably high coding costs for data providers, or both. A vocabulary with too few terms will not necessarily have greater distinguishability for data providers – since determining what aggregate term a fine grained subject instance belongs to can be just as hard as choosing between several finely distinguished terms – and will not result in adequate data quality for analysts.

The issue is discussed in greater detail in section 5.10 and 5.11.

2.3 Transition from JACS3

The impetus for the NSCS project came from a realisation that JACS3 had a number of issues that needed to be addressed (A roadmap to a new Joint Academic Coding System, Ferrell, 2013):

a) the limit of the existing coding framework has been reached;
b) changes and growth in JACS' range of functions mean it is no longer consistently applied;
c) it does not meet the needs of all of the key sector stakeholders;
d) Higher Education providers sometimes use JACS3 in inconsistent ways;
e) incomplete and misunderstood JACS3 terminology and definitions have led to poor quality data in some instances.

Issue a) has been addressed in HECoS by moving to random six character codes, and b) and c) have led to HECoS' ability to be used in different aggregation structures for different functions. Issues d) and e) were tackled by examining each term for its ability to be distinguished from related terms, and deleting those that weren't used or led to confusion, introducing a few terms where there were gaps and clarifying the definitions of those JASC3 terms that were kept.

All these changes, however, have an implementation cost, and that aspect of the new coding system was by far the most commented on, with twenty-nine HE providers raising the point in addition to HESPA, HESA and five experts. As noted, much of the cost of transitioning away from JACS3 is inherent in the change itself. Any change in subject coding requires adjusting software, changing practices and allowing for differences between new and old data sets. For that reason, HECoS has been designed to minimise the cost of change as well as the cost of classifying. Specifically:

- Several HECoS features help HE providers to reuse a classification of their courses and modules for multiple data returns by:
  - Separating classification from analysis and policy through HECoS' flat list structure.
  - Providing comprehensive coverage at an appropriate level of granularity.
  - Providing the infrastructure and governance to collectively establish one common aggregation framework.
  - Providing the ability to link to other relevant vocabularies.
- HECoS enables a greater range of staff to engage with classifying by moving to a practice based on matching well defined human readable terms, rather than one based on memorising JACS3's tree of fifteen-hundred alpha-numeric codes.
- HECoS should save time, effort and therefore cost by providing a more uniform vocabulary with clearly differentiated terms.
- HECoS' proposed registry for potential new terms (see section 5.8) should provide a cost effective way to track new subjects.
- Significant continuity with JACS3 terms should ease the cost of the change (see section 5.9).
- Guidance built into HECoS such as the related terms and preferred and non-preferred terms should further reduce time, effort and cost.

Furthermore, a programme of targeted adoption support is outlined in the Adoption Plan (Ferrell and Campbell, 2015). The main value of HECoS, though, should come from better, more accurate analyses for all stakeholders, enabled by more consistent and reliable data.

3. Methodology

After the requirements gathering of Stage 1 of the New Subject Coding Scheme Project, two prototypes were merged into one candidate scheme, which was hosted on a vocabulary development site that provided the ability to comment on each term[2]. Through a combination of targeted solicitation of specialists, and wider publicity drives, over thirty-seven[3] respondents, together with the project team, contributed a total of three hundred and twenty-four comments on HECoS vocabulary terms.

Along with the vocabulary development site, the project also hosted outlines of the coding scheme, the Governance Model and the Adoption Plan on a comment site[4]. Forty-three people left sixty-four messages on the comment site relating to all aspects of HECoS.

The project team also received thirty-one email responses and numerous comments and questions during a webinar organised with Jisc that was joined by around one hundred participants.

[2] https://ovod.net/tematres/vocab/

[3] Precise numbers can't be given, because we allowed anonymous contributions.

[4] https://subjectcoding.wordpress.com/

A classification exercise designed to test the coding scheme was conducted during the Student Records Officer Conference (SROC) 2015 conference in York, with a group of over forty participants and at the UCAS 2015 Admissions Conference in Newport with a group of about twenty participants.

Through all of these channels, the respondents represent a cross section of the sector with four agencies, two companies, twenty-three HE providers and a number of experts who contributed as individuals. An overview of respondents is available in appendix C.

The team has also performed several manual and database driven analyses of the new subject vocabulary and relevant datasets such as the HESA student data return in order to apply the acceptance criteria outlined in section 4.1, and spot gaps in subject term coverage.

4. An outline of HECoS

At the beginning of the consultation, the draft HECoS vocabulary had been created by merging two prototypes: one a conservative evolution of JACS3 and the other a more radical flat list of terms whose size could be varied at will. Following extensive consultation, the resulting draft had the following characteristics:

- A flat list, presented with a navigation hierarchy.
- Larger than earlier prototypes, but much smaller than JACS3.
- Focused on subjects of study rather than disciplines as the main entity for classification.
- Focused on courses and modules as the main entities to classify.
- Six character random codes with no leading zeros, though the formal identifiers for the terms are URIs.
- Open licensed.
- Contains related terms and non-preferred terms to aid classification.
- Developed on, and available from a development site at http://ovod.net/tematres/vocab/

During Stage 2 of the consultation, none of these characteristics have changed significantly. However, the focus on distinguishability as a key criterion, as well as the input of subject matter experts, have changed the content of the final proposed versions of the HECoS vocabulary significantly.

4.1 Acceptance criteria for the recommended HECoS vocabulary

The mechanism the NSCS Project used to control the shape and size of the vocabulary, as well as ensure consistency, is via acceptance criteria. These criteria have been used to determine whether a particular term will be accepted into the vocabulary or not.

For the public draft, a term had to:

1. Be part of JACS3 or be required to fill a gap in JACS3 as evident in the overloading of JACS3 codes in HESA returns[5].
2. Have evidence of being used in HESA returns by at least two institutions.
3. Have a definition and scope that is sufficient and comprehensive to allow classification.
4. Be reliably distinguishable from other terms, as evident in the set of courses classified with the term in HESA returns.

[5] Specifically: we found cases where a coherent and distinct cluster of courses was found under an unsuitable JACS3 code, because a more suitable code was missing. e.g. JACS' "developmental and reproductive biology" contained several courses in the study of developing organisms as well as numerous Obstetrics and Gynaecology programmes, for which no code existed.

Following feedback from the sector, this has been amended. For the final version, a term had to:

1. Be part of JACS3 and
   a. Have valid courses or modules coded with it, as evident in HESA returns.
   b. Have a definition and scope that is sufficient and comprehensive to allow classification.
   c. Be reliably distinguishable from other terms, as evident in the set of courses or modules classified with the term in HESA returns.

   or

2. Be required to fill a gap in JACS3 as evident in:
   a. The overloading of a JACS3 code in HESA returns with too many disparate courses and modules.
   b. Clearly identifiable clusters of course and module titles and descriptions in HESA returns for which no single JACS3 code exist.
   c. Suggestions by subject matter experts, with supporting evidence from any source.

As with the draft, the emphasis in the final version of HECoS was on continuity with JACS3, primarily in order to preserve as much backward compatibility with legacy datasets as possible (see section 5.9).

The first step (1.a.) was the elimination of those JACS3 terms that are not used or which are JACS3-internal duplicates. The second criterion was designed to filter out those codes that have definitions or scopes that are too ill-defined to classify with. An example would be "Law by area" (JACS3 M100): in the absence of the specification of said area, it is very difficult to decide what it applies to, and what not. This is especially true for legal subjects, because there was no clearly defined code for just "Law".

The distinguishability criterion was primarily used to tackle those terms that proved to be difficult to choose between when coding. This is evident in courses and modules that could be coded with either one of two subject terms, being spread seemingly randomly between them. Typical cases include JACS3's 'Geoscience' versus 'Earth Sciences'[6]. The difference may be clear to specialists in theory, but which courses and modules are coded using each term is not predictable in practice.

Extending the evidence data to include titles and descriptions of HESA courses was done to uncover any major gaps in subject coverage in JACS3, which is one of the goals of the NSCS Project. Because of the need to preserve continuity with JACS3 where possible, evidence has been scrutinised thoroughly, and use-for or preferred terms established where possible. Such terms function as pointers from commonly used synonyms to the actual term that is used to code. For example, subjects could be defined for 'Applied Social Research', 'Applied Social Studies' and 'Applied Social Science', but all comfortably fit the definition of 'Applied Social Science', so the first two were defined as subjects for which the preferred term is 'Applied Social Science'.

Other clusters have been found where a new subject term has had to be created, because no existing term was adequate. The difficulty with these cases is that the value space for potential new subject terms is very large indeed. In some cases, subject matter experts have been able to alert the team to the need for specific new subjects, e.g. 'enterprise and entrepreneurship'. In cases such as the 'humanities' subject, the gap was clear from the numbers of relevant courses and modules in the data and the absence of a suitable JACS3 code.

In spite of some exploratory text mining efforts, it is likely that some viable subjects remain undiscovered. For this reason, community involvement and a robust term acceptance process is necessary. This is outlined in more detail in the HECoS Governance Model (Campbell and Ferrell, 2015).

[6] JACS3 'Earth Sciences': 'The study of the earth as a unified system; includes earth resources, surface and crustal processes." JACS3 'Geoscience': "The study of the earth sciences, including geological chemistry and physics." 'Earth sciences' should be used for 'geoscience' in HECoS.

One other consequence of widening the evidence base to include the titles and descriptions of courses in the HESA student data return is that some terms required changes in their definition and label. There was very weak evidence of use of JACS3's 'Indian Language Studies', for example, but a search of courses demonstrated a large number of courses in Hindi language studies. HECoS, therefore, now has 'Hindi language' as a term[7].

5. HECoS design goals and the results of the consultation

The draft of the HECoS coding scheme was created with the use of fifteen design goals, each of which summarised a total of forty-four requirements that were gathered from stakeholders during the first phase of the project (Impact Assessment and Requirements Definition, Kraan and Paull, 2014). Over the course of the project, these goals have changed a little as interest shifted from questions such as the support for subjects over disciplines to the shape of the codes. Because these goals summed up requirements and guided the development of HECoS from the start, they remain a good structure to address the responses from the final public consultation.

A full listing of the original set of requirements, their links to design goals and an indication of whether they have been met (yet) is given in appendix D.

Each **topic in bold** refers to a cluster of feedback from respondents, listed in appendix B.

5.1 Supporting many purposes

HECoS' ability to accommodate multiple aggregations by separating terms from their aggregation, arose from an early requirements workshop with HEP representatives. Most stakeholders preferred the flat list structure over a number of alternatives for a variety of reasons (HEDIIP NSCS Structure and Candidate Scheme, Kraan and Paull, 2015):

- Simplicity and ease of use so the coding scheme is no more complex than it needs to be.
- Greater political neutrality because how subjects are aggregated becomes an explicit policy decision, which should reduce the incentive for strategic coding.
- Greater flexibility as more granular taxonomies for specific purposes can be attached 'underneath' the common flat list.
- Easier and more robust coding over time, because subjects do not have to be aggregated at the same time as the subjects are coded.
- A flat list rightly prioritises the subjects themselves and de-emphasises the structure in which they sit.
- A flat list formalises actual, diverse current purposes and eliminates an appearance of uniformity the data can't bear.

Support for a non-hierarchical structure is not universal, however.

Eight HE providers, one agency and five others raised HECoS' **lack of hierarchy** as a concern. Most worried about comparability of analyses if multiple subject analysis aggregations, or none at all, were to be the norm. This was considered to be a great danger and for that reason, the team has started work with the major sector bodies to define a default common aggregation framework along-side HECoS, the details of which are outlined in Recommendations for Subject Based Analysis & Text Mining (Cooper, 2015). This common aggregation framework should also support **benchmarking**, which three HE providers brought up, and the common aggregation framework will also be a topic of discussion in our continued **engagement with league table compilers**.

[7] Hundreds of other languages are spoken in India, of course, but other than Sanskrit, Urdu and English, we did not find much evidence of them being studied in UK HE. Any such programmes or modules could be coded with "South Asian Language Studies", however.

A couple of respondents also questioned whether it is possible, conceptually, to see subjects as being non-hierarchical, and if you can, whether that's workable. While it is true that 'humanities', for example, generally encompasses 'literature' for most people most of the time, it is also true that a course in 'humanities' is a demonstrably different thing from one in 'literature', and the course in 'humanities' may not even contain anything from 'literature'. In short: what may look like a simple 'part-of' relation, can be quite complex in practice. Also, while 'aerodynamics' can be regarded as a narrower term from 'physics' from one perspective, it can also be seen as subordinate to 'aerospace engineering' from another.

While the relativity of relations between subjects may be very real, presenting subject terms without any relations could well be unsettling for some stakeholders, and represent an unfamiliar way of working. For that reason, the navigation groupings will be maintained, even if the distinction between it and the forthcoming common analytical aggregation needs to be made clear to avoid the **navigation and aggregation hierarchies confusion** one agency warned against.

It should also be noted that four HE providers and one other praised the **lack of hierarchy as a benefit**, as voiced earlier in the NSCS Project. Of these strengths, **aggregation transparency** was seen by several respondents as crucial if the policy advantages and reduction in strategic coding are to be realised.

5.2 Making HECoS codes easy to use

Because JACS3's hierarchy is deeply embedded in its codes, HECoS required a new code scheme. From the various options (HEDIIP NSCS Structure and Candidate Scheme, Kraan and Paull, 2015), short codes with no particular meaning proved the most popular solution, but with the proviso that the codes were memorable, or that a switch to a more memorable form be found if needed. In practice, the opaque six digit codes remained and were used exclusively for machine purposes, while people used the labels of the terms themselves.

HESPA, five HE providers and two others raised concerns about **opaque codes** being **problematic**, mostly because there is a fear that the use of the codes could exclude members of staff who didn't deal with them frequently. While this needs further monitoring, experience suggests that humans – expert or otherwise – will use HECoS term labels rather than codes. This will increase the accessibility of HECoS to newcomers.

A crucial aspect in this regard is the way in which various systems will implement HECoS. Using modern user interface functions such as 'type ahead' or 'auto-suggest' to match likely term candidates with what a user has started to type improves usability considerably. We therefore recommend that system vendors integrate such a function in their applications and that lists of the HECoS terms be used in the comparable validation of cell values function in spreadsheets.

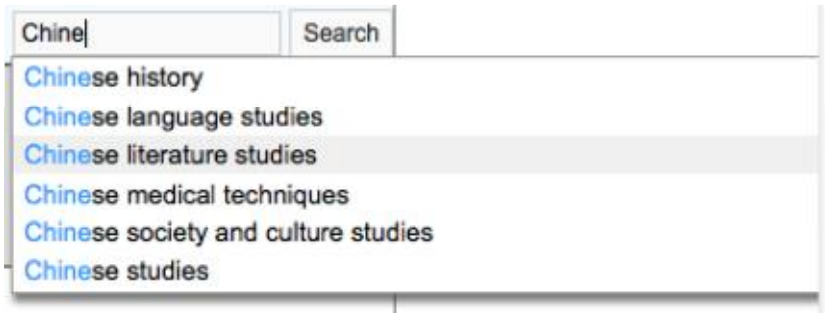

Figure 1 Type ahead function in Tematres

Meanwhile, two HE providers and one expert thought **opaque codes** were **good**, because of their flexibility, and the future proofing they provide. While many IT colleagues and vendors indicated during earlier phases of the project that extending database fields from four to six characters was unlikely to pose a problem, two respondents mentioned the **six character codes** as a **problem** because it will require a small change to some applications.

5.3 Support policy implementation

The purpose of subject coding is to enable policy implementation, and, as such, it has a bearing on quite a few policy areas (HEDIIP NSCS Structure and Candidate Scheme, Kraan and Paull, 2015)

- Funding.
- Monitoring of specific subject areas (e.g. SIVS).
- Widening participation.
- Information for students.
- Progression.
- Accountability (including QA).
- Benchmarking and performance indicators.

There are tensions inherent in serving all these domains, and between those who provide the data and those who use the data to implement policies, as highlighted in section 2.

Generally, the multitude of policies is what prompted the need for a vocabulary that could support multiple aggregations, as well as the need for consensus around the common aggregation framework discussed in section 5.1. Funding and monitoring of specific subjects is what creates the need for comprehensive coverage (section 5.11), balanced by a need for an appropriate level of granularity (section 5.10). Finally, information for students often requires rapid change in the vocabulary to follow the latest trends in provision (section 5.14), tempered by the need for stability for benchmarking (section 5.8).

In all of these areas, the project has attempted to strike the right balance by continuous consultation with the full range of stakeholders from the beginning.

5.4 A unified approach to support a wide group of stakeholders

As noted, the wide range of purposes for which a subject coding vocabulary is used by its various stakeholders means that while the terms themselves can be shared, their aggregations will need to vary.

However, this does not imply that respondents are happy to support multiple aggregations for the same or similar purposes. In that regard, one HE provider was keen that HECoS should be used integrally by UCAS rather than **mapped to UCAS codes.** One other provider advocated its **use for search** by students searching for courses through UCAS.

5.5 A means of linking to data classified in other frameworks.

There was also a notable appetite amongst a number of respondents to map HECoS to **cost centres**, **ATAS**, **QAA subject benchmarks** and **the REF**. The latter was not universally popular, as one HE provider warned **not to map to the REF** on the grounds that it could be misleading because REF returns are subject to very different policy considerations. Similarly, HESA warned against linking to FE vocabularies such as LDCS because they classify different entities.

The linked data nature of HECoS makes it feasible to define relations with any other vocabulary precisely and easily from a technical point of view. However, as the LDCS and REF cases already make clear, considerable care should be taken with what those relations mean. Because HECoS is designed to classify subjects of study, the relation to other entities such as organisational departments is often complex and partial. This subject is explored in greater depth in the Recommendations for Subject Based Analysis & Text Mining report (Cooper, 2015).

Within the sphere of subjects of study, the project did realise early on that there may be a need to relate to much more fine grained specialist vocabularies that already exist in a couple of specific subject areas. A clear example is the area of work classification in the NHS occupation codes (see appendix D). For that reason, we envisage that

relevant sector organisations can relate vocabularies under their control to HECoS using well-known specifications such as the W3C's Simple Knowledge Organization System (SKOS).

Both HESPA, HESA and one expert saw the value, or even necessity, of such a solution, but emphasised that **specialist vocabulary implementations** needed to make sure that those links were sound, and that they not overlap with HECoS. HEFCE and one HE provider also emphasised the need to **limit specialist vocabularies** in order to keep the burden of data submission in check, in line with the goals of HEDIIP.

Because of the linked data approach that the NSCS project took, these concerns can be addressed directly. At the heart of the approach is the assumption that different communities need to define, structure and maintain their own vocabularies entirely independently of others, but that these communities can nonetheless cooperate with each other by agreeing precise relations between the terms of their respective vocabularies. For that reason, an overlap between a term in one specialist vocabulary and another term in HECoS is not a problem; as long as it is clear to everyone which term belongs to what vocabulary.

Another consequence of this approach is that you can rely on the precise, agreed relations between two vocabularies to save work. For example, if you know that one term in a specialised vocabulary such as the area of work classifiers of the NHS occupational codes is a narrower match of a HECoS term, you can treat a module that was classified by an NHS code as if it were classified by the HECoS term in data returns that expect HECoS classifications. More specifically still, if it is agreed that the NHS' "paediatric dentistry" is narrower match to HECoS' "dentistry", then a module that is about "paediatric dentistry" can be returned as classified with HECoS' "dentistry" automatically.

Whether such an automated mapping is done by the HE provider that sends the data or by a data consumer such as HESA is a matter of policy. This aspect is covered in greater detail in the HECoS Governance Model (Campbell and Ferrell, 2015).

It should be noted that, though HECoS is built on linked data concepts and technology, the crucial separation between vocabularies, and the precise, agreed relations between vocabulary terms can be realised in a range of other technologies just as well.

One other aspect raised by one expert is **internationalisation**; the provision of the HECoS vocabulary in a range of languages other than English. This is referred to in the HECoS Adoption Plan (Ferrell and Campbell, 2015).

HESA recognised the suggested need to make HECoS available as open data, but didn't think that a particular **export** format such as **CSV, XML etc.** needed to be defined. This is also taken into consideration in the Adoption Plan (Ferrell and Campbell, 2015).

5.6 The NSCS should be simple to understand

As with any classification vocabulary, HECoS is designed to enable courses to be classified with the most specific term that fits. This is very simple in principle, but a lot depends on the supporting infrastructure around the coding scheme that helps users to find the right term. This is a matter discussed at greater length in the HECoS Adoption Plan (Ferrell and Campbell, 2015), which includes recommendations such as **more user testing** prior to full adoption and the possibility of an **enhanced search service.** There are some features of the vocabulary itself that can also help. Two of these are the Use For terms (UF, which designate the preferred terms to Use For non-preferred terms) and Related terms (RT). These can help point a user to the term that HECoS prefers among a clutch of similar commonly used ones, or point to a HECoS term that is related but different. Such pointers have been developed for a number of HECoS terms where the data indicated that they could help. The suggestion of HESA and one other expert to **share mappings with UF & RT** will be followed.

One expert suggested a **change** of the **name of the scheme** to 'HECS' to make it simpler to spell, but others indicated a strong preference for an acronym without associations. Since many in the sector are already familiar with 'HECoS' as the name of the new subject coding scheme, we suggest the name stays.

5.7 Be robust

At the start of the NSCS Project, several stakeholders voiced a need for the new scheme to be robust to a changing regulatory and resource environment. To that end, the team has spent a large amount of time making sure that every term in HECoS is used and easy to distinguish, and that the vocabulary as a whole covers all foreseeable subjects at some level of precision. All terms, their usage, their relations to other terms and their definitions have been examined multiple times, with a change decision log of over seven hundred entries on the last iteration alone. A large part of the feedback that prompted these changes were given by requests to **add, merge** or delete **subjects** by the six HE providers, HEFCE and two experts who contributed to the consultation website, and the thirty seven experts who engaged with the vocabulary development site. Their suggestions, in combination with analyses of the HESA student data return allowed us to identify and plug as many gaps as possible. As noted previously, however, it is possible that some viable subject clusters may have been missed, which is why it remains important that there is an equally robust change management process for HECoS.

The objective behind the drive to make the vocabulary more robust in this way is to make subject coding easier, which means that the data generated with it is more robust and accurate, which should benefit all stakeholders. While progress toward this goal requires careful monitoring, we hope that this, along with the benefits of HECoS outlined in section 2.3, will address the **unknown benefit** concern of the change raised by six HE providers and one expert.

5.8 Be stable

There is a clear tension between the needs of subject analysts and admissions officers for a vocabulary that changes as fast as possible and the needs of data providers and analysts who work with time series for a vocabulary that changes as infrequently as possible. In order to manage this tension, the NSCS proposed a process that separates the recording of potential new subject terms from changes to the HECoS vocabulary itself, and the business of cyclical data returns, by establishing a separate new candidate terms registry.

The basic idea is that, when coding a new course or module, colleagues can check HECoS for good HECoS terms first. If no term is quite adequate, they can look among the candidate terms proposed by others in the registry. If no suitable candidate term already exists, a new term can be proposed. Metadata about the course or module in question, and the HECoS code used in the interim allows changes to be introduced in an orderly way.

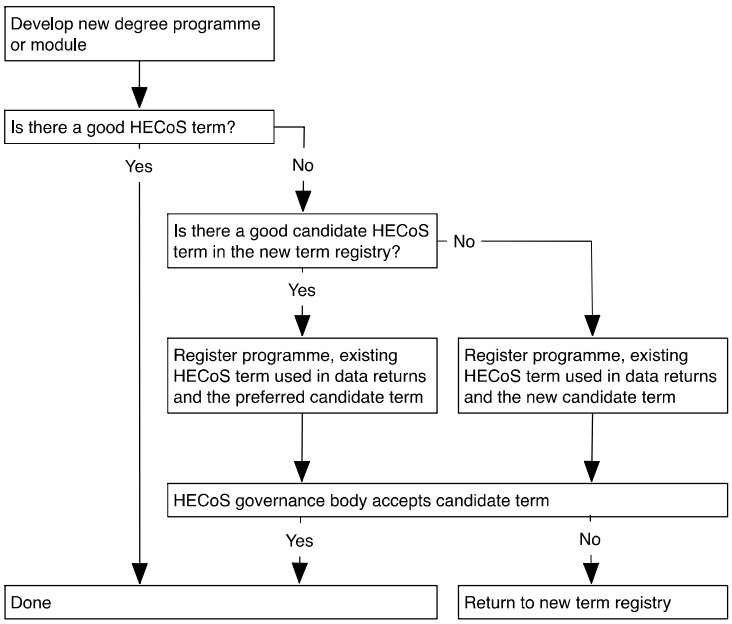

Figure 2 New subject term registry process

What this mechanism does not prescribe are the acceptance criteria for a new term, and the frequency with which they should be added. Consequently, although three agencies, five HE providers and one expert who raised the **stability** of the scheme as a worry were generally happy with this mechanism, opinions about acceptance criteria and frequency of change varied. Management of change is a central governance issue, which is dealt with in the HECoS Governance Model (Campbell and Ferrell, 2015).

5.9 Backwards compatibility with legacy data.

The need to **preserve backward compatibility with time series** was highlighted by three agencies, one company, three experts and thirteen HE providers, and has been a goal for the NSCS Project from the start. Such compatibility could also help preserve the investment in recent one-off analyses.

The preservation goal needs to be addressed at two levels: one is the relation of terms to the legacy JACS3 codes, the other is the relation of JACS3's hierarchy to a new common aggregation framework (see section 5.1). The latter is crucial since most analyses will deal with aggregations rather than individual terms. Fortunately, if stakeholders agree on the comparability of aggregations in the two vocabularies, such a mapping could be relatively straightforward. This is elaborated further in the HECoS Recommendations for Subject Based Analysis & Text Mining (Cooper, 2015).

The need for a **mapping to and from JACS3 codes** was raised separately by two agencies, one company, four experts and nineteen HE providers, not just for analytical purposes, but also to control costs in the transition from JACS3 to HECoS. Because the terms in both vocabularies are necessarily different, a perfect mapping is not possible. However, eighty-six percent of HECoS terms have a close match to an existing JACS3 term, which should make the continuation of the aggregation used by the **KIS** (raised by HESA and an expert as an issue) easier. The degree to which the course catalogues of individual HE providers can be linked in this way will vary, but of all the University of Greenwich's courses, for example, seventy-three percent have a JACS3 code that is a close match to a new HECoS code. By comparison, the University of Edinburgh, with over twice as many programmes, has a close match rate of a very similar seventy-four per cent. This suggests that the danger of costly **recoding problems** noted by one respondent may not be as large as feared.

Having said that, because adjacent terms have shifted and sundry definitions have been clarified, a check of those close matches would improve data quality and consistency considerably. The example cited earlier of the new HECoS term introduced for the two-hundred and seventeen entrepreneurship courses in the HESA student data set that were classified with fifty-nine different JACS3 codes illustrates the potential for improvement. Those courses and modules that don't have a single close match from JACS3 to HECoS will need to be matched to a new HECoS term from among a handful of likely candidates, often with guidance from broader matches or related terms defined within HECoS itself.

The predictability of these links could mean that it may be worth exploring an automated service or script that suggests a HECoS code for a course or module catalogue on the basis of an existing JACS3 code and other data. This matter will be taken up in the forthcoming HECoS Adoption Plan (Ferrell and Campbell, 2015).

5.10 An appropriate level of granularity (in relation to usability and meaningful data)

The size and therefore granularity of the HECoS vocabulary has been another area different stakeholders have held different views from the inception of the project. Seventeen HE providers, HESPA and three experts commented that there were **too many terms.**

The fear is that a large vocabulary would trade increased perceived precision for decreased accuracy. The expectation is that, by having many more terms that are necessarily more subtly differentiated, the chance that an individual coder picks the wrong code increases, and the likelihood of reliably consistent practice across the sector decreases.

Data from clusters of some fine grained JACS3 codes do bear this out. When looking at the JACS3 codes assigned to courses in interior design and interior architecture[8], for example, which course gets assigned to which code appears to be almost random. Though there may well be a difference between the subjects, there is no reason to maintain the distinction between the codes for analytical purposes, because there is no reliable and consistent difference between the sets of courses and modules they classify. For this reason, the terms are **merged** in HECoS, with 'interior design' and 'interior architecture' added as non-preferred terms.

On the other hand, there are also clusters of seemingly equally fine grained JACS3 codes that are coded reliably, even when they are used for much smaller numbers of courses and modules. 'Phonetics', 'Phonology' and 'Phonetics and Phonology', for example, are clearly closely related subjects, but also sufficiently differentiated in label, scope, content and definition to have small but accurate sets of courses classified with their associated JACS3 codes. For that reason, their terms are also in HECoS.

What's more, cases also exist where a *lack* of codes in JACS3 has led to serious inconsistency in the data. For example, as its academic community made clear to us, there are large numbers of courses in 'entrepreneurship' but no specific JACS3 code. As it's unclear what wider term would be appropriate, the two-hundred and seventeen entrepreneurship courses in the HESA student data set were classified with fifty-nine different JACS3 codes of varying validity. Consequently, not only did the subject become invisible from a classification point of view, it also 'polluted' a number of adjacent terms with data points of varying and unknown appropriateness.
What these and many other examples suggest is that the relation between size of the vocabulary and reliability and consistency of classification is not simply inversely proportional. Other factors inherent to the terms and what they classify, i.e. their distinguishability, play at least as significant a role in data quality. For that reason, the NSCS Project has focused on distinguishability as the main criterion over vocabulary size.

5.11 Provide comprehensive coverage of the range of subjects of study available in HE at an appropriate level of detail for its target users

The fact that there is no straightforward inversely proportionate relation between vocabulary size and data quality does not mean, of course, that there is no relation at all. Taken in extremis, the remit of funding agencies to track potential Strategically Important and Vulnerable Subjects (SIVS) means having a subject vocabulary of unbounded size and in permanent flux, because new developments and changing interests are bringing forth a constant stream of new subjects.

Because such a scheme is clearly unworkable, determining how many terms is **not enough terms**, as HEFCE worried, is not straightforward. In order for the tracking of new or evolving SIVS to be meaningful, however, the data on which the tracking is based needs to meet minimal standards of consistency and reliability, which brings us back to the distinguishability criterion outlined previously, as well as the need for training, robust implementation in relevant systems and other adoption support (Adoption Plan, Ferrell and Campbell, 2015).

Feedback in earlier phases of the NSCS Project suggested that, while modules are in scope, the primary focus of the use of HECoS will be in classifying courses. Two experts raised the possibility of **coding modules and courses differently,** on the grounds that modules are different entities with a potentially different even more fine grained set of subjects in current use. Also, while there is some loss of precision, it was felt that the current vocabulary is precise enough for the purposes of coding both modules and courses since it includes codes at differing levels of detail.

[8] JACS3 'Interior design' is defined as "The study of/training in the use of artistic techniques in the planning, designing, equipping and furnishing of residential, commercial and public interior spaces." 'Interior architecture' is defined as "The study of enclosed spaces; design, implementation and materials."

5.12 Increased consistency of application across institutions

Beyond the question of size, consistency of use has been an important design goal from the start by itself. As with other design goals, the distinguishability criterion is key here: the clearer the distinction is between different terms, the more likely it is to be applied consistently in different institutions.

5.13 Be clearly separate conceptually from the JACS3 system

This design goal summed up those requirements that could not be easily met by merely updating JACS3. This includes requirements such as clearly differentiated term definitions, the use of common labels, the elimination of unused terms and supporting multiple subject aggregations. These requirements have been met in the course of pursuing other design goals, and include the distinguishability work on definitions, aligning subject labels to what is used in course and module titles, the reduction of the term count by a third by eliminating unused subjects, and by introducing the flat list structure.

5.14 A framework with scope for evolution

At its most basic, the flexible structure and six digit codes of HECoS enable it plenty of scope for evolution. Its definition in linked data technologies also allows it to be linked to other vocabularies in ways that are precise but flexible, it and allows HECoS to meet the change management requirements of versioning, retiring and status tracking that HESA would like to see. Such content management is readily accommodated by existing linked data practices and standard vocabularies. At the same time, HECoS can be bound in a range of other technologies, and its further evolution is not dependent on a linked data technology base.

More widely, HECoS needs to become as much a process as a product. With the right adoption and governance in place, the UK HE sector as a whole can take ownership of the vocabulary, and evolve it to where we need it to go.

6. Design principles

The design goals summed up requirements from a variety sources, including a set of design principles that guided the development of the NSCS from the start. It is worth revisiting these principles in order to compare the end result with the original intent. In the NSCS work definition, principles 1 to 7 were deemed essential, and 8 and 9 desirable.

6.1 A coding framework that corresponds to recognised good practice

Whilst we have endeavoured to follow good practice in the way we addressed all design goals, this principle was bundled into the design goal of providing ‘a means of linking to data classified in other frameworks’ (see section 5.5) because taking a linked data approach was identified as good practice for a vocabulary such as HECoS (A roadmap to a new Joint Academic Coding System, Ferrell, 2013). This was confirmed by our research of similar vocabularies (Impact Assessment and Requirements Definition, Kraan and Paull, 2014). The linked data consideration was what led to the adoption of the W3C’s Simple Knowledge Organization System (SKOS) as the source format for HECoS. Adopting SKOS gave a wide choice of vocabulary management tools to work with, as well as a number of utilities that allows HECoS to be converted into a range of visualisations as well other formats.

One important recommendation from the initial coding system roadmap (A roadmap to a new Joint Academic Coding System, Ferrell, 2013) was to use persistent Universal Resource Identifiers (URIs) as identifiers for terms in the new vocabulary, and to maintain such authoritative URIs in a web service for the sector.

This has not been done yet, because there is a technical and governance dependency of the solution on the organisation that will administer HECoS in the future. This organisation needs to have an internet domain under its control at a minimum, and preferably have an infrastructure in place to make HECoS available via a web service. This issue is addressed in the Governance Model (Campbell and Ferrell, 2015), currently HECoS still uses temporary URIs that are not authoritative and will change.

In theory, plausible but unresolvable URIs could have been used, but in practice, because URIs look like ordinary web links or Universal Resource Locators (URLs), it was deemed friendlier to make the temporary URIs resolve to a vocabulary management web service under the NSCS team's control.

6.2 A framework with scope for evolution

This principle became a design goal in its own right (section 5.14). The original intent was focused on both the ability to respond quickly to changes in subjects of study being offered, as well as providing a means of coding that has room for expansion.

The ability to respond quickly to new subjects has led to the design of the term registry outlined in section 5.8, which meets the principle, without endangering the competing goal of stability.

The roadmap (A roadmap to a new Joint Academic Coding System, Ferrell, 2013) suggested a six digit code as a solution to the ability of the vocabulary to expand. This was adopted in HECoS, but with the proviso that these codes are semi-random. This was not necessary just to enable the expansion of HECoS, but mostly a consequence of the decision to separate the term codes from the analysis and other aggregations. As a result, HECoS can now evolve in its own way and at its own pace, independent of how it is aggregated for particular purposes.

6.3 A framework at an appropriate level of granularity

This principle became a design goal as well (see section 5.10), and the issue of finding the right balance between potential detail and reliable, consistent coding has remained a constant that was resolved by focusing on proven distinguishability as the arbiter. One aspect originally suggested in the roadmap (A roadmap to a new Joint Academic Coding System, Ferrell, 2013) was to have three rather than four levels, which has been adopted in the HECoS navigation hierarchy.

6.4 An easy transition path from JACS to the proposed scheme

This principle was bundled into the 'Backwards compatibility with legacy data' design goal (section 5.9), and was addressed by taking JACS3 terms as the point of departure for HECoS. Terms were added only if there was strong evidence of a gap. Additional measures to ease the transition are elaborated in the HECoS Adoption Plan report (Ferrell and Campbell, 2015).

6.5 Increased consistency of application across institutions

This principle also became a design goal in its own right (section 5.12), and was addressed by the project's focus on distinguishability, and testing it with the HESA student record return data, as well as a number of coding tests with relevant HE provider staff.

6.6 A coding framework that is consistent with the open data agenda

Along with the 'good practice' principle outlined in 6.1, this principle was bundled into the 'A means of linking to data classified in other frameworks' design goal of section 5.5.

The original intent was to enable linking with similar vocabularies from other jurisdictions, and to enable re-use by others. This was addressed by adopting an open data license, adopting an open data format and developing the HECoS vocabulary on the open web.

As a result, there has been an offer from the open academic data community to host the vocabulary, and assist in translating it and interlinking it to others[9].

6.7 A unified approach to support a wide group of stakeholders

This principle also became a design goal (section 5.4), and was originally concerned with addressing the needs of a wide range of stakeholders from the sector, and gaining their acceptance. This has led to multiple rounds of consultation, and intense involvement and many contributions from a broad range of representatives of the UK HE sector.

6.8 A means of linking to data classified in other frameworks

This principle was desirable rather than essential, but became a design goal (section 5.5) along with the principles of good practice (6.1) and open data (6.6), and was addressed by the same range of measures: the adoption of a linked data approach.

6.9 Backwards compatibility with legacy data

This principle was a desirable, but became a design goal (section 5.9) along with the principle of an easy transition from JACS3 (6.4). The original intent was to preserve the validity of legacy analysis done with JACS3 data, and particularly to preserve time series analyses.

As noted, HECoS started with JACS3 terms, and eighty-six per cent of HECoS terms have a closely matched JACS3 antecedent. For the remainder, users can pick from a limited choice of appropriate mappings.

7. Recommended next steps

The HECoS vocabulary has been developed iteratively over the course of a year. At each stage, representatives from across the UK HE landscape have often not just provided feedback, but have actively helped shape the new subject coding scheme.

Nonetheless, because subject coding plays an important role in the UK HE regulatory framework, tensions do remain about the way in which the vocabulary's structure can serve different purposes, what the right level of granularity or size should be for HECoS, and how the cost of changing from JACS3 to HECoS can be minimised.

In order to manage these tensions, we recommend that:

- A common subject term aggregation framework be collaboratively developed in order to balance the flexibility of HECoS' flat structure with the need for a common yardstick.
- The period prior to HECoS adoption is used to train classifiers and to demonstrate the benefits of the vocabulary by undertaking voluntary 'dry run' data submissions.
- To ease the transition from JACS3 to HECoS by offering automated suggested mappings of JACS3 coded data sets to HECoS coded data sets.

[9] The offer was made by Chris Gutteridge of Southampton University, and one of the initiators of data.ac.uk

For practical adoption purposes, we also recommend that:

- A web service on a permanent domain be established as soon as possible, so that stable, authoritative and derefenceable URIs can replace the temporary ones used with HECoS at the moment.
- System vendors be engaged again to help with cost effective and user friendly HECoS implementations.

The HECoS vocabulary is presented in a separate compact human readable form in a guide described in appendix A. Machine readable versions are also available separately and are described in appendix C.

Further recommendations about a common aggregation framework are available in Recommendations for Subject Based Analysis & Text Mining report (Cooper, 2015). The Governance of HECoS is outlined in more detail in the Governance Model (Campbell and Ferrell, 2015), and the specifics of HECoS' adoption are provided in the HECoS Adoption Plan (Ferrell and Campbell, 2015).

References

Campbell, L. M. & Ferrell, G. (2015) *HEDIIP NSCS Project - HECoS Governance Model*.

Cooper, A (2015) *HEDIIP NSCS Project - Recommendations for Subject Based Analysis & Text Mining.*

Ferrell, G. (2013). *Classifying subject of study; A roadmap to a new Joint Academic Coding System.* Retrieved from http://www.hediip.ac.uk/wp-content/uploads/JACS_Report_2013-07.pdf

Ferrell, G. and Campbell, L.M. (2015), *HEDIIP NSCS Project HECoS Adoption Plan*.

Kraan, W. G. and Paull, A. (2015) *HEDIIP NSCS Structure and Candidate Scheme*. Retrieved from http://hediip.ac.uk/wp-content/uploads/HEDIIP_NSCS_PD03_2015-02-18b.pdf

Kraan, W. G. & Paull, A. (2014) *New Subject Coding Scheme; Impact Assessment and Requirements Definition*. Retrieved from http://www.hediip.ac.uk/wp-content/uploads/Stage_1_report_2014-11-13.pdf

Appendix A. The HECoS vocabulary guide

Under separate cover, a HECoS vocabulary guide (file: HEDIIP_NSCS_PD04_guide_2015-10-09.pdf) offers a terse but comprehensive, easy to navigate, stand-alone listing of the HECoS terms and navigation hierarchy.

Appendix B. Consultation respondents

Below is a list of those who provided comments on the design and structure of the vocabulary itself during phase 2 of the NSCS project, and what issue they raised. Each issue is discussed in section 5.

Issue raised	Respondent
Add Subjects	Exeter University, St George University, QMU, HEFCE, Leeds University, Conservatoire, Rob Walton, Nigel Adams
Aggregation Transparency	HEDIIP Advisory Board, Warwick University, Andrew Watson
Backward Compatibility with Time Series	HESPA, HESA, Greenwich University, UCAS, Loughborough University, University of Canterbury, University of Gloucestershire, UCLAN, University of Wolverhampton, University of St Andrews, Edinburgh University, University of Sunderland, Prospectus.ac.uk, University of York, Sussex University, Sheffield Hallam University, Helen McGarry, Warwick University, Rob Howard, University of Surrey, Plymouth University, Oxford University, University of Sunderland, University of Bangor, Andrew Watson
Benchmarking	University of Worcester, University of Sunderland, University of Bangor
Change Management	Rob Walton, Warwick University, University of Surrey, Leeds Trinity University, NTU, Birmingham University, Amanda Coleman
Code Modules and Courses Differently	Ray Lashley
Cost	HESPA, HESA, Aberdeen, Greenwich, Manchester University, St George University, Northumbria University, University of Gloucestershire, UCLAN, Aberystwyth University, University of Wolverhampton, Napier University, St Andrews University, UWL, Andrew Reynolds, John Fox, Salford University, Edinburgh University, University of Chester, University of Worcester, QMU, Sunderland University, University of York, University of Sussex, Sheffield Hallam University, Helen McGarry, Dr C Hutchinson-Howorth, Warwick University, Robert Howard, University of Surrey, Plymouth University, Leeds Trinity University, Oxford University, University of Sunderland, University of Bangor, Birmingham University, David Singer
Don't Link to FE vocabularies	HESA
Don't Map to REF	Loughborough University
Engagement with league table compilers	Aberystwyth University, Rob Walton, University of Sunderland, University of York
(relate HECoS to) KIS	HESA, Andrew Reynolds
Lack of Hierarchy is problematic	University of Aberdeen, University of Cardiff, HEFCE, University of Wolverhampton, Andrew Reynolds, John Fox, University of Salford, Edinburgh University, University of Chester, Prospectus.ac.uk, Rhodri Rowlands, Newcastle University, Graham Peely
Lack of Hierarchy is a benefit	City University, Simon Walton, Birmingham University, Oxford University, University of Chester
Limit Specialist Vocabularies	HEFCE, St George University
Map to ATAS	Loughborough University, Warwick University
Map HECoS to Cost Centres	HESPA, University of Aberdeen, University of Liverpool, Swansea University, University of Northumbria, University of Gloucestershire, UCLAN, University of Wolverhampton, City University, Andrew Watson

Issue raised	Respondent
Map HECoS to (and from) JACS	HESA, University of Aberdeen, Greenwich University, UCAS, University of Liverpool, Loughborough University, University of Hull, University of Northumbria, University of Gloucestershire, UCLAN, University of Wolverhampton, University of St Andrews, Salford University, University of Chester, QMU, University of Sunderland, Prospectus.ac.uk, Sheffield Hallam University, Helen McGarry, Warwick University, Robert Howard, Leeds Trinity University, University of Bangor, Birmingham University, Andrew Watson, Gordon Rennie
Map HECoS to REF	HESPA, University of Liverpool, Swansea University, University of Northumbria, University of Gloucestershire, UCLAN, University of Bangor
Map HECOS to UCAS Codes	UCLAN
Map HECOS to QAA subject benchmarks	Leeds University
Merge Subjects	David Singer
Navigation and Aggregation Hierarchies Confusion	HESA
Not enough terms	HEFCE
Opaque Codes are Problematic	HESPA, UCLAN, Aberystwyth University, University of Wolverhampton, Edinburgh University, University of Worcester, Helen McGarry, Robert Howard, University of Sunderland
Opaque Codes are Good	Swansea University, University of Wolverhampton, Chris Gutteridge
Recoding problem	University of Chester
Share Mappings with UF & RT	HESA, Robert Walton
Six Character Code Problem	Edinburgh University, Gordon Rennie
Specialist Vocabularies Implementation	HESPA, HESA, Robert Walton
Too Many Terms	HESPA, Greenwich University, Liverpool University, Manchester University, St George University, Northumbria University, UCLAN, University of Wolverhampton, St Andrews University, UWL, John Fox, Salford University, Edinburgh University, University of Worcester, Newcastle University, University of York, Dr C Hutchinson-Howorth, Warwick University, University of Surrey, University of Bangor, Birmingham University, Ray Lashley
Unknown benefit	Aberdeen, Loughborough, Manchester, Aberystwyth, City University, University of Worcester, David Singer
Use HECOS for course searches (UCAS)	Loughborough University

Appendix C. Machine readable versions of the HECoS vocabularies

It should be borne in mind that the URIs in all the machine readable versions are temporary at the time of writing. Stable and authoritative URIs will follow later, but the current ones will be kept dereferenceable as long as possible.

Spreadsheet

An Excel spreadsheet (HEDIIP_NSCS_PD04_HECoS_Vocabulary_2015-10-09.xlsx) with multiple sheets has been made available for general machine processing purposes. The sheets contain:

- The "simple" sheet contains the basics of the vocabulary: a URI to identify each term that can be classified with, a short code derived from the URI, the term itself, its definition, and a scope note (if any).
- The "related" sheet contains all the HECoS URIs and terms as sources mapped to related HECoS URIs and terms.
- The "preferred" sheet contains all the preferred HECoS terms that can be used for classification, and all their non-preferred synonyms, which can't be used for classification.
- The "navigation" sheet contains all the classifying terms and their URIs, and the two levels of navigation groupings that have been defined for them. Beware that a term can occur in more than one navigation grouping, and that the navigation grouping labels can't be used for classification.

XML and RDF

The full HECoS data set is available as both SKOS-RDF and ZTHES XML, both of which can be used in a range of vocabulary management applications. The SKOS-RDF is full five star linked data.

SKOS-RDF: HEDIIP_NSCS_PD04_SKOS_2015-10-09.rdf
ZTHES XML: HEDIIP_NSCS_PD04_ZTHES_2015-10-09.xml

HECoS was developed on a Tematres instance at http://ovod.net/tematres/vocab/. The Tematres site makes each term available in a range of formats. It also has an API, which can be explored here: http://ovod.net/tematres/vocab/services.php

Appendix D. Full Requirements Matrix

The following list represents all requirements as they were gathered during phase 1 of the NSCS project. A 'requirement met' column has been added to indicate what the status of the requirement is at the time of writing.

Note: the IDs of each requirement are not in sequence because some requirements have been merged or dropped in the process of requirements gathering.

ID	Title	Description	Link to design goal(s)	Stakeholder(s)	Rationale	Priority	Type	Grouping	Requirement met
R1	NSCS and JACS3	The NSCS system shall support, and provide guidance on, use of the NSCS with data classified with JACS3, for example for maintenance of time series.	1, 6, 14	analysts	Backwards compatibility for time series and easy transition.	Mandatory	constraint	Uses	Yes
R2	Interoperating with other subject vocabularies	Publishing mechanisms shall be designed so that the NSCS can be linked to and interoperate with other classification systems, including amongst others the Learndirect Classification System (LDCS).	11	SFA, HEFCE, HESA, Jisc	Support linking to other data. Widen usage across the sector. Enables development of services and applications with multiple data sets.	Mandatory	constraint	Uses	Yes
R3	Persisting URIs	A persistent URI for each of its entities shall be included, so that they can be addressed readily by systems using linked data.	2, 4, 5, 11	UCAS, HESA	Good practice. Supports open data.	Mandatory	functional	Codes and structure	To be resolved in the Transition Plan (NSCS Project Stage 3)
R4	Replacing JACS	The NSCS shall replace all current uses of JACS3.	3	UCAS		Mandatory	constraint	Uses	Yes

ID	Title	Description	Link to design goal(s)	Stakeholder(s)	Rationale	Priority	Type	Grouping	Requirement met
R7	Having clear and concise definitions	The NSCS shall have clear and concise definitions for each of its terms. Where possible, these definitions should be drawn from an appropriate authority recognised as such by the HE sector.	1, 2, 10	classifiers, analysts, HEE	Aids understanding for analysis and classification	Mandatory	performance	Content	Yes
R8	Classifying subjects	The NSCS shall be used to classify HE data by subject of study, while supporting aggregation for usage via discipline, including courses throughout their lifecycle.	3, 7, 12, 15	academics	Provides mechanism for using disciplines within a subject framework	Desired	constraint	Uses	Yes
R9	Governing and sector bodies	Governance of the NSCS shall be influenced strongly by specified sector bodies (HESA, UCAS, and others to be determined), by HEPs, representatives of Professional, Statutory and Regulatory Bodies (PSRBs) and other significant stakeholders. There shall be clear lines of responsibility, openness and transparency.	3, 6	HEPs, sector bodies	Provides strong sector representation on development.	Mandatory	performance	Governance	Yes, see governance report (Campbell and Ferrell, 2013)
R10	Providing guidance on coding for specific purposes	The NSCS shall include guidance on how codes are to be allocated with reference to specific purposes. Methods may be different for different purposes.	2, 3, 5, 6, 13	UCAS, classifiers, SFA	Recognises that the scheme sits within a 'service' implementation approach.	Desired	functional	Guidance and help	Partial: additional guidance to be given in the Adoption Plan (Ferrell and Campbell, 2015)

ID	Title	Description	Link to design goal(s)	Stakeholder(s)	Rationale	Priority	Type	Grouping	Requirement met
R11	Providing training recommendations	The NSCS documentation shall include recommendations for training in how to use the scheme.	2, 3, 6, 9, 11, 13	UCAS, classifiers	Recognises that the scheme sits within a 'service' implementation approach.	Mandatory	functional	Guidance and help	To be given in the Adoption Plan (Ferrell and Campbell, 2015)
R12	Comparing courses	The NSCS shall facilitate comparisons between courses by applicants and advisers.	2, 3, 6	UCAS	Recognises importance of subject comparisons at course level	Desired	functional	Uses	Yes
R13	Supporting operational and time series statistics	The NSCS shall enable production of useful operational and time series statistics by HESA, UCAS and others, that are compatible with JACS3-based statistics at JACS3 Principal Subject level (for example student progression rates, staff-student ratios, applications, acceptances, and so on).	3, 4, 6, 7, 8, 9, 10, 11, 14	UCAS, HESA, PSRBs, HEFCE, plus many others	Continuing requirement for planning and analysis of data by subject. Implies major thrust of requirement at roughly JACS3 Principal Subject level.	Mandatory	functional	Uses	Yes, structurally. Additional resources to be proposed in the Adoption Plan (Ferrell and Campbell, 2015)
R14	Supporting regulated professions	The NSCS shall ensure that specific subjects can be catered for, including subjects directly relevant to regulated professions, bearing in mind the requirements for codes and for support in correct usage: 'teaching' and teaching subjects, such that 'teaching' and individual subjects that are taught in teaching courses can be analysed; pharmacy; healthcare sciences; planning;	3, 6, 7, 10, 11	UCAS, HEE, PSRBs, HEFCW	Recognises a particular problem area that could usefully be resolved.	Desired	functional	Content	Yes, addressed in the Recommendations for Subject Based Analysis & Text Mining report (Cooper, 2015)

ID	Title	Description	Link to design goal(s)	Stakeholder(s)	Rationale	Priority	Type	Grouping	Requirement met
R15	Supporting service oriented approaches to publishing on the internet	The NSCS shall be capable of deployment using a service oriented approach.	3, 5, 11	UCAS	Recognises that the scheme sits within a 'service' implementation approach.	Desired	performance	Codes and structure	Yes
R16	Supporting multiple aggregation structures	The NSCS shall support multiple aggregation methods, for example aggregation for league tables, for application statistics, for HEP planning purposes.	3, 6, 7	HEPs, UCAS, HEFCE	Supports more usages than JACS3.	Desired	functional	Codes and structure	Yes
R17	Providing support for course searching	The NSCS shall provide subject classification as the starting point for course search and marketing purposes.	3, 10, 13	UCAS	Supports more usages than JACS3.	Desired	functional	Uses	Yes
R18	Remaining static for an academic cycle	Governance of the NSCS shall enable management of the NSCS as an HE standard that shall remain static for any single specific academic annual cycle.	3, 4, 5, 9	UCAS, HESA, HEPs	Must be stable and robust	Mandatory	constraint	Governance	To be addressed in the Governance Model report (Campbell and Ferrell, 2015)
R19	Facilitating annual reporting and review	Governance of the NSCS shall facilitate annual reporting and review by all stakeholders with a change implementation period of not less than 3 years, with a defined, transparent process for changes, in particular for adding and removing terms.	3, 4, 5, 9, 14	UCAS, HESA, HEPs	Must be stable and robust, but also capable of change	Mandatory	constraint	Governance	To be addressed in the Governance Model report (Campbell and Ferrell, 2015)
R20	Enabling statutory returns	The NSCS shall form the basis of subject-based statutory and regulatory returns by HEPs to sector bodies.	3, 4, 6, 7, 8, 9, 10, 11, 14	HESA, UCAS, HEPs, sector bodies, PSRBs	Continuing requirement for planning and analysis of data by subject.	Mandatory	functional	Uses	Yes

ID	Title	Description	Link to design goal(s)	Stakeholder(s)	Rationale	Priority	Type	Grouping	Requirement met
R21	Supporting student lifecycle comparisons	The NSCS shall support comparisons of what students study and what they progress to doing later, e.g. occupation	3, 6, 11	HEE	HEE has a specific requirement to do this in respect of students who go on to NHS employment	Desired	functional	Uses	Yes
R22	Enabling mapping to NHS occupation codes	The NSCS shall have a mapping to NHS Occupation Codes, in order to support NHS workforce planning.	3, 11	HEE	HEE requirement	Desired	performance	Uses	Yes
R23	Enabling workforce and capacity planning dataset comparisons	The NSCS shall enable comparisons across UCAS, HESA, HEP and HEE data sets for capacity and workforce planning and for quality assessment.	3, 6, 7, 11, 14	HEE, SFA	HEE requirement	Desired	functional	Uses	No; it was found that workforce comparisons require cost centre codes
R24	Enabling disaggregation in health subjects	The NSCS shall enable differentiation between critical health-based subjects.	3, 6, 10, 11	HEE	HEE requirement	Desired	constraint	Uses	Yes
R25	Enable aggregations for NHS planning	The NSCS shall permit aggregation for NHS planning purposes (workforce, capability and quality assessment).	3, 6, 10, 11	HEE	HEE requirement	Desired	constraint	Uses	Yes
R26	Providing more detail in medicine subjects	The NSCS should have more detail in healthcare science and medical specialisms than JACS3 to facilitate usage within the NHS and health professions.	6, 7, 11	HEE, PSRBs	HEE requirement	Best value	performance	Codes and structure	Yes
R28	Providing guidance on the NSCS and KIS	The NSCS should give clear guidance on how it should most usefully be included in the Key Information Set.	6, 7, 10, 11, 14		Covers existing requirement for use of JACS3	Mandatory	functional	Guidance and help	Yes, detailed in the Recommendations for Subject Based Analysis & Text

ID	Title	Description	Link to design goal(s)	Stakeholder(s)	Rationale	Priority	Type	Grouping	Requirement met
									Mining report (Cooper, 2015)
R29	Facilitating datasets that are fit for purpose	Governance shall facilitate the creation, maintenance and usage of authoritative data sets.	3, 6, 9, 11, 14	GPC, HEPs, HEE, other sector bodies	Provides for current and wider usage in analysis via HESA, HEFCE and others.	Mandatory	performance	Uses	Yes
R31	Describing guidance purposes clearly	NSCS guidance shall clearly describe the purposes for which it is designed to be used. It will also cover similar areas for which it is not designed to be used.	1, 2, 3, 6	HEPs, SFA	Competent usage for classification and analysis requires this.	Mandatory	performance	Guidance and help	Yes
R32	Encouraging clarity in the description of data collection purposes	NSCS guidance shall encourage data collectors to describe clearly the purposes for which the subject-classified data will be used.	1, 2, 3, 6	HEPs	Competent usage for classification and analysis requires this.	Mandatory	performance	Guidance and help	To be given in the Adoption Plan (Ferrell and Campbell, 2015
R33	Supporting aggregation of STEM and SIV course data	The NSCS shall support unambiguous aggregation of data for STEM and SIV subjects.	1, 2, 4, 5, 6, 7	HEPs	Important for critical policies using the data.	Desired	performance	Uses	Yes
R34	Supporting more fine grained classification of subjects	The NSCS shall support more fine-grained classification of subjects of study forming a separate coding frame, for example for modules or educational resources. This supports policy interventions.	3, 7, 10, 11	HEPs, HEFCW. HEFCE	Certain functions require module level classification, e.g. HEFCW funding model, reading list creation. HEFCE: nuclear technology, Islamic Studies, and	Desired	functional	Codes and structure	Yes

ID	Title	Description	Link to design goal(s)	Stakeholder(s)	Rationale	Priority	Type	Grouping	Requirement met
					specialisms at Masters.				
R35	Making codes memorable	NSCS codes should be memorable, but should not encourage the use of the code as shorthand for the term itself.	1, 2	HEPs	Facilitates usage by classifiers	Best value	performance	Codes and structure	Partial; the codes themselves are barely memorable, but terms are memorable and unique and should be used instead
R36	Classifying subjects or groups of subjects	Each term in the NSCS shall be a subject of study or a cognate group of subjects of study.	1, 2, 6, 7, 10, 13	HEPs	Facilitates design and maintenance	Desired	functional	Content	Yes
R37	Supporting multiple, combined or interdisciplinary subjects	The NSCS shall include guidance on how to classify multiple, combined and interdisciplinary subjects.	1, 2, 7, 10	HEPs		Mandatory	functional	Guidance and help	Addressed in Recommendations for Subject Based Analysis & Text Mining report (Cooper, 2015)
R38	Excluding unstudied subjects	The NSCS shall not include subjects that are not subjects of study in courses in the UK.	1, 2, 7, 8, 10, 13		Facilitates design and maintenance	Desired	constraint	Content	Yes
R39	Using common labels	NSCS terms shall use labels for subjects of study that are commonly used names within the subject area or discipline.	1, 2, 7, 8, 10, 13		Facilitates design and maintenance	Desired	constraint	Content	Yes
R40	Differentiating term definitions	Definitions of terms in the NSCS should not be confusingly similar.	1, 2, 4, 11, 13	HEPs	Facilitates usage by classifiers	Mandatory	constraint	Content	Yes

ID	Title	Description	Link to design goal(s)	Stakeholder(s)	Rationale	Priority	Type	Grouping	Requirement met
R41	Avoiding leading zeros	NSCS codes should not have leading zeros.		Project team	Facilities technical implementation	Best value	constraint	Codes and structure	Yes
R42	Using a consistent number of characters	NSCS codes shall have a consistent number of characters.	2	HEPs	Facilities technical implementation	Desired	constraint	Codes and structure	Yes
R43	Providing support documents	The NSCS shall have supported documents, such as guidance manuals, subject coding manual, context-sensitive help, scope notes within terms.	2, 3, 9	HEPs	Facilitates uptake of the schema.	Desired	functional	Guidance and help	To be given in the Adoption Plan (Ferrell and Campbell,2015)
R45	Including external definitions of important subjects	The NSCS shall include as terms those subjects of study included in SIV and STEM definitions, and other similarly recognised lists of important subjects. SIVS: chemistry, engineering, mathematics and physics; quantitative social science; and modern foreign languages and related area studies. STEM: anatomy and physiology; biosciences; chemistry; computer sciences; earth, marine and environmental sciences; engineering and technology; mathematics; pharmacy and pharmacology; physics	3, 6, 7, 11	HEPs, HESA, HEFCE, SLC	Important for critical policies using the data, including student finance policy	Mandatory	constraint	Content	Yes
R46	Supporting the classification of subjects not already included	NSCS shall recommend a mechanism to support the classification of subjects of study not currently covered by the schema.	5	HEPs	Facilitates maintenance	Desired	functional	Guidance and help	Yes

ID	Title	Description	Link to design goal(s)	Stakeholder(s)	Rationale	Priority	Type	Grouping	Requirement met
R47	Supporting explicit aggregations of subjects	Ways in which NSCS data is grouped (hierarchies and aggregations) shall be negotiated as part of governance and published.	3, 6, 7	HEPs	Supports use outside HEPs	Desired	constraint	Codes and structure	Addressed in Recommendations for Subject Based Analysis & Text Mining report (Cooper, 2015)
R48	Supporting specific HEP functions	NSCS shall support benchmarking and analytics for internal and external use of HEPs, as well as performance management, competitor/sector analysis and market intelligence.	3, 6	HEPs	Supports use inside HEPs and by HEPs	Desired	functional	Uses	as R47 above
R49	Structural position of Welsh as a subject in the scheme	Ensure Welsh is in a language category, not Celtic Studies.	6, 13	HEFCW	Improves on JACS2 and 3	Mandatory	constraint	Content	Yes