HEDIIP NSCS Project -

Recommendations for Subject Based Analysis & Text Mining

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 Scheme that met the needs of a broad group of stakeholders and reflected 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 APS Ltd and Aspire Ltd. 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.

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

- Andy Youell, Director, HEDIIP
- Dan Cook, Head of Collections Development, HESA
- 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
- Cetis, as senior supplier

Principal Authors/Editors: Lorna M. Campbell and Gill Ferrell.

Contributors: Phil Barker, Adam Cooper, Alan Paull, Wilbert Kraan.

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

About This Report

This report is part of the work of the Higher Education Data and Information Improvement Programme (HEDIIP) New Subject Coding Scheme Project (NSCS Project). It comprises deliverable PD05, as defined in the work specification. Although it is set apart from the proposed new classification scheme (HECoS, the Higher Education Classification of Subject) and proposals for its governance and adoption plan, it is closely related to those aspects; it builds upon the classification scheme to be published alongside this report and the consultation responses to the draft adoption plan and governance model, and will inform the forthcoming deliverables concerned with adoption and governance.

Authorship and Status

Owner	Adam Cooper, Cetis
Principal Author/Editor	Adam Cooper, Cetis
Contributors	-
This Version	Final for acceptance
File Identifier	HEDIIP_NSCS_PD05_SubjectAnalysis_2015-11-23.docx

Document History

Date	Person	Notes
2015-07-03	Adam Cooper	Adapted from the discussion documents created for the meeting 2015-06-30 (See Appendix 4)
2015-07-07	Adam Cooper	Revisions completed, comprising: - modifications arising from discussion meeting - incorporating feedback from NaCTeM (text mining) - addition of appendix with simple text mining examples - formatting to HEDIIP style
2015-07-10	Adam Cooper	Revised according to peer review by Lorna Campbell First version sent to HEDIIP PMO for acceptance
2015-07-17	Adam Cooper	Revised following review by HEDIIP PMO. Minor corrections and clarifications.
2015-07-23	Adam Cooper	Revised following review by HEDIIP PMO. Minor corrections and clarifications.
2015-08-21	Adam Cooper	Corrections to HEDIIP house style. No content change.
2015-10-13	Adam Cooper	Modified according to 2015-10-12 Project Board comment.
2015-11-06	Adam Cooper	Revised following Advisory Panel comment.
2015-11-23	HEDIIP PMO	Updating cross references

Contents

1. Executive Summary

Subject Based Analysis (SBA) refers to the analysis of data in which the records have been classified using HECoS, with statistics presented at the level of aggregations over several different subjects. SBA is an essential component of the management of Higher Education at all levels from HEPs to government, but current practice is not coherent across the sector, which means that statistical summaries cannot be accurately matched. This hinders effective strategic management, and policy formation and implementation, and leads to media/public confusion. Improving SBA requires common rules for: 1. what constitutes a subject grouping and; 2. the way in which balanced/joint studies (e.g. joint maths and physics, or maths with physics) are apportioned in statistical work. In addition to statistical analysis, which is the focus of the work specification for the NSCS Project, subject groupings are also used for a range of other purposes such as: for eligibility criteria, funding allocation, in guidelines and informal process support. Hence there is a wider scope of opportunity for transparent and managed approaches to subject grouping. 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.

Outline Recommendations Relating to Subject Based Analysis

1. Base a set of standard cross-sector aggregation rules on the KIS/Unistats subject groups and with apportionment for major/minor and balanced studies.
2. Integrate governance of the standard aggregation rules with governance of the HECoS classification scheme under the aegis of the body recommended by "The Blueprint for a New HE Data Landscape" (KPMG, 2015), as detailed in the Governance Model report also produced by the NSCS Project (Campbell & Ferrell, 2015).
3. Ensure that subject group definitions, which may be owned by individual Core Sector Bodies or government bodies, and maintained by them in addition to the standard set, are effectively disseminated alongside HECoS. These include: HEFCE SIVS, HEFCW ASCs, mapping to QAA subject benchmarks, FCO ATAS.
4. Provide support to these bodies by drafting subject group rules with them as part of the adoption process, in the interest of consistency and efficiency.

Text mining fulfils a different role in understanding the landscape of subject of study. In the context of this report, Text mining is concerned with insights which may be gained about the subject of study from an algorithm-driven analysis of module or programme descriptions, or other text assets related to a course. Text mining is intrinsically not 100% accurate, but becomes a relevant approach when the quantity of data is such that human analysis is not feasible or when the lower levels of accuracy are an acceptable trade-off for reduced human effort. It is likely to be particularly applicable whenever it would be necessary to inspect a large number of course descriptions to answer a question, for example to explore provision of specialised topics or skills which are cross-cutting, but it may also be used for recommending HECoS codes, quality control, or for automated trend analysis.

At present, the principal practical obstacle to useful text mining is the absence of a suitable body of text to analyse. This should be of a consistent style, rich in keywords and of sufficient length. A modification of the data collection requirements imposed on HE providers would enable programme specifications, which are usually of the necessary quality, to be gathered into a repository for text mining. Stakeholder feedback indicates that there is little appetite for such an initiative; although it is conceivable that the benefits would outweigh the cost, there is a lack of evidence to motivate action.

Outline Recommendation Relating to Text Mining

1. Promote a small pilot and encourage experimentation, isolated from impact on business processes, to build evidence for the value of text mining for management and administration.

2. Introduction

2.1. Structure of this Report

This document begins with short introductory sections to give definition to the character of subject based analysis, and text mining, and to outline the methodology followed in developing this report. It then deals with subject based analysis and text mining separately. For each of these, we outline the problem space before presenting a set of proposals for a way forward, comprising aims and associated actions which could form part of the adoption of HECoS within the wider HE data landscape programme. It will be a matter for HEDIIP, its sponsors, and stakeholders to decide in what ways, if any, to progress the proposals. The proposals do, however, presume adoption of HECoS. The purpose of this document, and the remit for the NSCS Project, is to capture what we believe is a broadly-supported way forward founded on good information management principles.

2.2. Clarification: Subject Based Analysis

We use the term Subject Based Analysis (SBA) to refer to the analysis of data where subject of study is explicitly recorded, and for which statistics are reported according to groupings of HECoS terms. For the purpose of this work, we are primarily concerned with SBA where the statistics are published or shared with other organisations since it is for those cases that a common approach is highly desirable. The same aggregation rules are valuable for HEP-internal statistics too, because institution-level decision-making will be more effective when sector-level data can be included in the decision-making process.

This report will also consider a number of matters related to subject based analysis which are not analytical *per se*, but which stakeholders have identified as being important use cases for HECoS, and which are in some ways similar to SBA. These are candidates for consideration in the interest of an effective and efficient data and information landscape.

SBA is an essential component of the management of Higher Education at all levels from HEPs to government.

Key Features

Practical SBA requires defined rules for:

1. what constitutes a subject grouping;
2. the way in which balanced/joint studies are apportioned in statistical work (e.g. joint maths and physics, or maths with physics);

In practice, analysis may take different approaches such as counting individuals or creating statistics using full-time equivalency, and different approaches to aggregation over the level of intended outcomes are likely. For example, UCAS admissions processes deal with applications to full-time undergraduate provision so the business process defines the appropriate treatment of the data on these dimensions. It is not realistic to have a single approach across the HE data and information landscape, given the variety of processes in which the data is used. Hence, this report is concerned with driving common practice in respect of aggregation over subject of study only.

2.3. Clarification: Text Mining

Text mining is a research topic, an academic research tool, and a technology used at scale in business contexts. A near-synonym for “text mining” is “text analytics”, which is often used in the business world and is somewhat less mysterious.

In the context of this report, text mining is concerned with insights which may be gained about the subject of study from an algorithm-driven analysis of module or programme descriptions, or other text assets related to a course.

In principle, this could include teaching and learning materials, or other supportive content, but such resources are generally unlikely to be available outside the institution and do not form part of the processes with which HEDIIP is concerned; they will not be considered to be in scope.

Text mining is intrinsically not 100% accurate, but becomes a relevant approach when the quantity of data is such that human analysis is not feasible or when the lower levels of accuracy are an acceptable trade-off for reduced human effort. It is likely to be particularly applicable whenever it would be necessary to inspect a large number of course descriptions to answer a specific question. Text mining may be applied in a fully-automatic way or may be used as a supporting tool, for example, to filter out a manageable subset of documents for inspection, to extract document sections for attention, or to offer keyword prompts. It should also be borne in mind that indexing and query technologies, which are so widespread that they would not normally be called "text mining", often rely on similar computer science.

2.4. Methodology

The genesis of this report involved the following stages:

1. Work in Stage 1 of the NSCS Project established requirements, which were documented in project deliverable Impact Assessment and Requirements Definition (Kraan and Paull, 2014). That document set out some design principles arising from the requirements which also comprise constraints on approaches to subject based analysis. Stage 1 of the NSCS project involved extensive consultation with stakeholders from sector bodies, Higher Education Providers, PSRBs, etc., which is described in PD02.
2. Desk research was undertaken on current practice in subject based analysis and to investigate suitable source text for text mining. Appendix 1 summarises SBA sources.
3. Stage 2 public consultation, which took place from February to May 2015, was based on publication of draft versions of the subject classification scheme (HECoS), an adoption plan, and a governance model. These elicited comments pertaining to subject based analysis, which are summarised in Appendix 3.
4. Discussion documents for each of text mining and subject based analysis were developed by synthesis of the desk research and responses made during the public consultation. These characterised the problem space and presented draft proposals.
5. On June 30th 2015, a meeting of analysis and planning stakeholders took place to discuss the draft proposals. The draft proposals were broadly accepted, but a number of necessary changes were identified. Appendix 4 indicates the participants and the principal conclusions. In other respects, this deliverable follows the same approach as in the discussion document referred to in #4, above.
6. The National Centre for Text Mining (NaCTeM) was consulted to verify technical aspects of the desk research on text mining. We are grateful to Ioannis Korkontzelos from the University of Manchester (the host of NaCTeM) for his assistance. In addition, some small scale text mining was undertaken to demonstrate core ideas (Appendix 5).

3. Subject Based Analysis

3.1. Terminology: Mapping, Hierarchy, Aggregation, Association

A number of terms are in common use when talking about approaches to grouping subject of study, but they are sometimes used for different concepts. In charting a way forwards for subject based analysis and related issues, it is important to discriminate between different aspects so that the discussion and recommendations are understood in relation to the problem space. We hope that drawing distinctions will avoid errors in assuming comments relate to a different problem than is intended.

In this document, we use the following terms with these specific meanings:

Mapping refers to the relationship between elements in two classification schemes when the usage is functionally the same, for example, when both are intended for use in classifying subject of study. Mappings may sometimes be simple 1:1 relationships or may be imprecise and require a case-by-case review by a human being. A mapping between JACS3 and HECoS could be created, but with the caveat that it is only valid when both schemes have been used to classify subject of study. This document does not deal with mapping, but it has been identified as an important tool for HECoS adoption and a JACS3 to HECoS mapping is proposed in the NSCS Project Adoption Plan (Ferrell and Campbell, 2015). During HECoS adoption, the HE Data Governance Body may identify the need for more mappings to be created to eliminate adoption obstacles.

Aggregation refers to the grouping of classified entities for a given purpose, for example the computation of standardised statistical reports (hence we will often use "statistical aggregation" below). Different aims for statistical reports might indicate different aggregations. Common practice has historically been to use the JACS3 hierarchy as a default statistical aggregation approach but, over time, aggregation rules have been defined which diverge in different ways from the hierarchy (see Appendixes). Statistical aggregations are critical features of understanding and managing Higher Education provision, at all levels from HEPs to government, and the principal purpose of this document is to address that need in the context of HECoS. We will refer to the specification of a given aggregation as an aggregation rule, and note that practical statistical aggregation requires rules which also cover attributes other than subject of study.

Association is used when none of the above forms of relationship apply, for example when the relationship is akin to a mapping but the function of the classification differs. For example, the HESA Academic Cost Centre is defined[1] in a way which is quite different to the concept of subject of study, so we would describe a collection of statements about the relationship between cost centre codes and HECoS to be an association and not a mapping. "Mapping" may be commonly used for this kind of relationship but we prefer to use (and define, within the scope of this document) different terms because the inferences which can reasonably be made when the two classifications have a different function are not the same as when they do (e.g. in the current example, because the different definitions of cost centre and subject of study must be accounted for). Failure to appreciate the difference in purpose between, for example, cost centre and subject of study is to risk turning data into misinformation.

Hierarchy, when referred to a classification scheme, is reserved for a strict relationship between terms in which terms may have any number of narrower terms but only one broader term. HECoS is a non-hierarchical scheme for classifying subject of study, whereas JACS3 was a hierarchical scheme. Hierarchy may also be used to describe an aggregation scheme also, when it means that larger groupings are strictly computed as sums of lower level groups.

Note on Structure in HECoS

Although HECoS appears to be presented as a hierarchy, related terms have been grouped together only to help users to locate appropriate terms. These groups cannot be used for coding subject of study, neither is it assumed that they should be the basis for any mapping, aggregation, or association; they exist purely to aid the discovery of terms. HECoS is non-hierarchical by design because this de-couples the definitions of the subjects of study from the variety of structures which users may need to overlay upon the scheme. By doing so, it allows for a HECoS term to be located by several different routes, which is particularly useful for subjects which do not sit under traditional academic categories.

[1] https://www.hesa.ac.uk/component/studrec/show_file/12041/a%5E_%5EACCENTRE.html

3.2. Problem Definition

Three issues have been identified with the status quo of subject based analysis:

1. The results of SBA from different parts of the HE sector are not comparable, with the result that sector statistics as a whole are incoherent. It is not possible to get a consistent picture[2]; this hinders effective strategic management, and policy formation and implementation, and leads to media/public confusion.
2. The definition of subject groups may change over time, as what is considered strategically-important[3] by policy-makers changes, and new subject groups become relevant. This may arise from changes in the nature of what is taught, leading to changes in HECoS, and from emerging policy concerns (e.g. the recent interest in Big Data stimulated interest in the level of statistical and data-handling skills among graduates). This amplifies the previous point; it is neither efficient nor beneficial to manage and disseminate these changes in an uncoordinated way, as at present.
3. Newspapers (e.g. The Times, The Guardian) and the Complete University Guide, OECD and others, use differing sets of subject headings (and probably different interpretations of how subjects should be grouped under these), leading to further public confusion and difficulty in relating data from the media to official data, including Unistats.

The change from JACS3 to HECoS affords an opportunity to reform an undesirable situation and to increase the strategic and operational value of SBA. Increasing use of data for decision-making will surely make it more important that high quality statistics from a variety of sources can be robustly integrated.

In addition to analysis, which is the focus of the NSCS Project work specification, subject groupings are also used for a range of other purposes:

- as components of eligibility criteria, e.g. for student finance, or funding allocation to HEPs;
- as criteria in the Academic Technology Approval Scheme[4], which is used to impede the flow of expertise that might pose a national or international security risk;
- in guidelines and informal process support, e.g. to indicate the likely match between course coding and relevant Quality Assurance Agency for Higher Education (QAA) Benchmark Statements.

Hence there are opportunities for transparent and managed approaches to subject grouping. A consistent approach across the mix of analytical and non-analytical uses, so far as consistency is compatible with business requirements, stands to reduce effort in data handling systems and so these non-analytical cases are included in this report.

3.3. A Proposed Way Forward for Subject Based Analysis and Allied Issues

It is expected that the specific activities outlined below would form part of an integrated adoption process.

Consultation revealed some over-arching stakeholder requirements (Appendix 3), particularly from HEPs, for: a controlled environment in which proliferation and rapid change are countered, and; an organised and centralised approach to dissemination. The following sections address these requirements in different ways according to the character of the mapping/aggregation/association and our assessment of their ownership. The over-arching principle is that transparency and an organised and centralised approach to change control and dissemination should be developed, such that a small core set of definitions of subject groupings becomes the *lingua franca*; the governance arrangements described in the NSCS Project Governance Model report, (Campbell & Ferrell, 2015) indicate that this should be achieved through the same consensus-driven body, the HE Data Governance Body, as

[2]This occurs across many areas. For example, SFC and HEFCE definitions of STEM differ, with HEFCE making more use of the JACS3 subject line. UCAS and HESA take different approaches to combined studies. Groupings used for general purpose statistical publications vary.

[3]For example, HEFCE research and intervention in relation to strategically important and vulnerable subjects (SIVS): http://www.hefce.ac.uk/whatwedo/crosscutting/sivs/

[4]https://www.gov.uk/academic-technology-approval-scheme.

governs HECoS and other key aspects of the HE data and information landscape, consistent with the vision outlined in "The Blueprint for a New HE Data Landscape" (KPMG, 2015). In respect of SBA, this body should be concerned both with directly managing some of these structures, and with recognising others from appropriate sources, to guide stakeholders towards trusted information. The following recommendations make clear which structures should be managed and identifies some which should be recognised, at least during the introduction of HECoS. Appendix 2 outlines some additional recommendations pertaining to implementation, which particularly relate to the dissemination of subject group definitions.

Assumptions

The proposed way forward relies upon four assumptions:

1. The idea of **a common aggregation scheme being used as the default option for public statistics is broadly supported and Core Sector Bodies are committed to working towards the creation of a joint approach as part of the adoption plan for HECoS**. Public consultation comments (Appendix 3) and consensus at the meeting of analysis and planning stakeholders indicates this is a reasonable assumption (Appendix 4).
2. The scope of a common approach is limited to definition of: a) what constitutes a subject grouping (at various levels); b) the way in which balanced/joint studies are apportioned (e.g. joint maths and physics, or maths with physics).
3. Where governance is indicated, this is assumed to be the remit of the 'independent collective and consensual governance body' envisaged in The Blueprint for a New HE Data Landscape (KPMG, 2015), as detailed in the Governance Model report also produced by the NSCS Project (Campbell & Ferrell, 2015).
4. This governance body, or HEDIIP in the interim, would oversee the process of agreement on the details outlined in this document.

Standard Cross-sector Statistical Aggregation Rules

These would be the default aggregation rules used for sector statistics and benchmarking, without prejudicing the need for some sector bodies to additionally produce statistics based on alternative rules, according to business need. The use of alternative rules should, however, be strongly discouraged unless there is a clear requirement for divergence; it should be for the 'independent collective and consensual' process of governance to act as a clearing-house for the views of the HE sector at large as to what is, and what is not, reasonable, and to be a venue for peer pressure to be applied.

The results of SBA must dove-tail with the Unistats/KIS information services. Unistats/KIS has been recently developed, is in active use, and was the subject of consultation by HEFCE within the last year[5]. Consequently, during the consultation on SBA the use of Unistats/KIS categories was proposed in a draft report (see Appendix 4). No contrary views were received by the NSCS Project.

Proposals

SA1.
The Standard Cross-sector Statistical Aggregation Rules should comprise definition of: a) subject-groupings of HECoS codes, and b) a standard approach to apportionment of, for example, combined or joint studies.

SA2.
The subject groups should be based on Unistats/KIS categories at all three levels, following a review of any known issues[5] and verification that Unistats level 3 contains an appropriate level of detail for general-purpose sector

[5]Consultation undertaken by HEFCE between 15 December 2014 and 13 February 2015 on publication thresholds and aggregation for Unistats/NSS concluded that "while the majority of respondents agreed that the current subject groupings would benefit from revision, there were no compelling arguments presented for implementing changes in 2015 and the majority favoured deferral of changes to 2016." http://www.hefce.ac.uk/lt/unikis/aggregation/

statistics[6]. The apportionment for major/minor and balanced studies should follow the current HESA approach, which is also reflected in funding council statistics.

SA3.
Consideration should be given to a formal alignment of Performance Indicator labels with level 1. Alternatively, deviations should be clearly documented.

SA4.
A draft set of rules should be created as an early task in the adoption process, alongside the creation of a JACS3-HECoS mapping[7], with a view to these being officially sanctioned through the governance process, and subsequently committed to by core sector data processors, at the earliest opportunity.

SA5.
The HECoS governance process for new term accession and term deprecation should include a mechanism for maintaining aggregation rules and the default aggregation should be version controlled and disseminated under the same governance regime as HECoS, with changes being SUBSTANTIVE in the terminology of the Governance Model.

SA6.
Adoption should be synchronised to an academic cycle agreed in the landscape governance process.

Academic Technology Approval Scheme

This is an important regulatory instrument of the Foreign and Commonwealth Office (FCO) and it is also of interest to Defence Intelligence. It is a formal association, in the terminology outlined above, that would be realised as a list[8] of HECoS associated with a single concept of "requires ATAS approval". The FCO is not an integral part of the HE data and information landscape, and yet it would be a source of burden on HEPs should the migration of ATAS to HECoS not occur in line with broader adoption. This suggests that a proactive and supportive approach to FCO would be sensible, in which central (HEDIIP) resource is used to expedite change in FCO.

Proposals:

ATAS1.
The initial association/mapping should be drafted as part of the work to create the default aggregation and agreed by consultation with the FCO.

ATAS2.
The owner of the adoption process should negotiate with FCO on the timetable for change-over.

ATAS3.
The authoritative version of the ATAS criteria is presumed to be published at www.gov.uk and maintained by FCO (as at present), but clearly linked-to from the HECoS dissemination channel.

ATAS4.
The HECoS governance process for adding or deprecating terms should include a reliable communication mechanism with FCO.

---

[6] Unistats level 3 contains 108 headings at present. Given a standard mapping of HECoS to that set, stakeholders not requiring high levels of specificity could express idiosyncratic aggregation schemes in terms of level 3, rather than at HECoS term level. This compares to 165 headings at JACS principal subject (letter+number) level, which is often used for this purpose.

[7] Development in parallel is in the interest of consistency of approach at different points in the data lifecycle.

[8] Strictly speaking, several lists, since ATAS discriminates differently according to level of study (e.g. undergraduate, masters by research).

QAA Benchmark Statements

QAA Benchmark Statements currently contain indicative lists of JACS3 subject groups that they are likely to apply to. These are indicative mappings which HEPs, in particular, may find useful for quality checking of data or for internal analysis. They are unlikely to be useful for cross-institutional comparison because of the freedom which exists for academics to select the best benchmark statement according to the affinity of teaching groups with discipline-related norms.

Proposals:

QAA1.
An indicative mapping from HECoS to QAA Benchmark statements should be drafted under the aegis of HEDIIP, as part of the same activity which defines the default aggregation, in consultation with QAA.

QAA2.
Based on the indicative nature of the current wording in QAA Benchmark Statements, it is assumed that this mapping is not business-sensitive for QAA. Subject to QAA consent, and bearing in mind possible changes in the regulatory landscape of HE, the mapping should be maintained, and published, by the custodian of HECoS using an informal process (e.g. delegated to a secretariat rather than requiring formal approvals, and not under a strict version control).

League Tables etc.

This includes the Complete University Guide, Times/Sunday-Times, and Guardian. Some respondents in the public consultation indicated a preference for an imposed approach to league table providers, but we have assumed that this is not practical, since the providers are understood to require freedom to differentiate their offerings.

Proposals:

LT1.
Alert these organisations to change as part of the adoption plan and engage them in discussion about the benefits, in terms of data journalism, of converging on the Standard Cross-Sector Statistical Aggregation.

Policy-oriented Subject Grouping

This sub-section is concerned with subject groups defined and used by funding councils.

Subject based analysis *prima facie* includes analysis of strategically important and vulnerable subjects (SIVS[9]), as undertaken by HEFCE. SIVS are also used to drive support for specific subjects, although in a rather limited way. HEFCW makes more extensive use of subject of study for funding via its definition of Academic Subject Category (ASC). Although ASC is for funding allocation, rather than for publishing SBA, it is sensible to include it in scope since there is *de facto* analysis undertaken at HEP and funding council level. Neither SIVS nor ASC definitions correlate with more general-purpose subject groupings, reflecting their purposeful rather than descriptive character. We believe it is not acceptable to bring ASCs or SIVS under the same governance regime as HECoS because their definition is so close to policy-making. This introduces both a likelihood of quickly changing definitions, and a desire for a locus of control within policy-making bodies.

[9] STEM (Science, Technology, Engineering, Mathematics) is a category in the HEFCE SIVS list, but STEM is differently defined by SFC. The HEDIIP NSCS Project Specification includes, for the work on subject based analysis (see Appendix 1), as an example “a process for developing a common definition of STEM”. At this point in time, however, we understand that the idea of converging on a single definition of STEM does not have stakeholder support; the proposals reflect this by outlining some steps which provide near-term value and which may ultimately lead to convergence, but which do not presume a common definition as the objective.

In spite of the above, there are a few reasons to consider policy-oriented subject groups in the context of HECoS:

- There is effort required to re-cast the definitions from JACS3 to HECoS, and this would be efficiently undertaken as part of the work to develop the mappings, other aggregations, and associations.

- HEPs and others may wish to use the definitions, in terms of HECoS codes, to undertake their own analysis. This will be aided by the publication of SIVS/ASC definitions in a form which is easy to use. The consultation process revealed a desire for mappings in use to be accessible to sector stakeholders and it is conceivable that better dissemination of definitions would help to reduce needless divergence and clarify where differences exist.

- Keeping the SIVS definitions up to date as HECoS changes will be particularly important, since changes are likely to occur most in strategically-important and vulnerable areas.

Proposal:

P1.
Encourage owners of policy-oriented definitions to: a) describe their policy-oriented subject groups using the same conventions[10] as the Standard Cross-sector Statistical Aggregation rules, and b) publish them (or references to them) from the same central location.

P2.
Subject to P1, the mappings from HECoS terms for funding council definitions (HEFCE SIVS, HEFCW ASCs, SFC Subject Areas) should be initially drafted in collaboration with the funding councils, as part of the same activity which defines the default aggregation, before hand-over to the definition owners.

P3.
The HECoS change process should include a reliable communication mechanism with definition owners such that addition and deprecation of terms in HECoS can trigger an update, if required.

Associations

Two associations have been identified through consultation as being of significant interest to HEPs: REF Units of Assessment and HESA Cost Centres. Following the terminology outlined above (Section 3.1), these would relate HECoS to entities of different logical type. There is a danger that creating and publishing associations elevates them to a status which is not justifiable. If they were to be created, maintained, or published alongside HECoS, strong caveats should be applied to make clear that these are informal and non-authoritative guides. For example, it should be emphasised that a published association for HESA Cost Centres is not a short cut to selecting a cost centre code, and that the definition of Cost Centre should be followed.

Although the NSCS Project team appreciates the point of view expressed by members of HEPs during the consultation process, we are concerned about the down-stream impact which may arise from misinterpretation or misuse. Hence, we believe it is very important that the decision to create such associations should not be appropriated from the organisations for which the data is part of core processes.

Proposal:

A1.
Associations between HECoS and Cost Centre or HECoS and UoAs should NOT be created under the aegis of HEDIIP. The decision to create an association, and the details of that association should be the realm of HESA or REF panels.

[10]So that HEPs and sector bodies can easily digest the definitions into their data processing systems.

3.4. Key Issue: Time-series Continuity

The issue of time-series continuity has been raised by HEPs many times during the NSCS Project consultation activities. Two aspects have been identified: the continuity across a transition from JACS3 to HECoS; and the continuity arising from changes to aggregation rules. The selection of the Unistats/KIS headings (SA2) coupled with assignment of HECoS terms to these subject groups in parallel with work to develop of JACS3-HECoS mappings (SA4), is intended to ameliorate the situation. Furthermore, the capture of data about continuing students in normal data collection processes will allow publishers of official statistics to compute weighting factors across the break-point. This would be outside the remit of HEDIIP. The HESA/Jisc HEIDI+/labs initiative may have a role to play in both defining and delivering data services to bridge the transition.

The *Adoption Plan* (Ferrell, G. and Campbell, 2015) considers time-series continuity, along with other transitional issues.

4. Text Mining

Text mining is a research topic, an academic research tool, and a technology used at scale in business contexts. Care must be exercised to discriminate between accounts of text mining as the object of research and accounts which apply well-established text-mining techniques for exploring text corpora. For the purpose of this report, we only consider text mining which can be undertaken using widely-available technologies and which would be accessible to someone with existing data-analytic skills, but who may need to acquire text mining expertise.

Taking account of these restrictions of scope, we adopt the following view on text mining for subject of study as an adjunct to HECoS:

1. It would be likely to use “bag of words” approaches in which the text is split into words and word order neglected. Common words (known as “stopwords”) are usually eliminated and groups of words with the same stem are often represented by the stem (e.g. “pharmacology” and “pharmacological” would both be represented as “pharmacolgi”). Bag of words approaches may use a simple word count, a binary yes/no for the appearance of a word, or may weight frequency according to the inverse of some function of the number of documents containing the word (known as term frequency inverse document frequency, or TF-IDF[11]), which emphasises the distinctive words. Bag of words approaches allow the text to be treated as numerical data using standard statistical tools and, although the process appears crude, it has been shown to work adequately in many situations. Bag of words approaches are technically-accessible. Appendix 5 gives some simple examples of using bag of words text mining.

2. It requires access to a large number of documents[12] each containing a sufficient volume of text, and of a similar genre. What comprises “large”, “sufficient”, and “similar” are inter-related and depend on the objectives. This requirement arises from the statistical nature of bag of words text mining and the variety inherent in free-form text.

Bag of words text mining is frequently applied to, for example: the identification of keywords to describe documents, automatic classification of the topic, clustering of documents according to similarity (or, conversely identifying anomalies), and the identification of statistically-significant changes in term frequency (including new terms) over time. Some of these objectives are best approached using a “training set” of documents which have previously been classified. Automatic classification typically does this, whereas clustering and keyword extraction are typically undertaken without that kind of prior knowledge, although they may apply some simple theoretical principles to decide which words are important according to the information which is sought.

[11] See https://en.wikipedia.org/wiki/Tf%E2%80%93idf

[12] “Document” is used in a loose sense to refer to a self-contained unit of text.

4.1. What Problems Might Text Mining Help Us With?

Before exploring the problems which text mining might help various actors in the HE sector to address, we wish to be clear that there is one problem which is not omitted by neglect: the assignment of HECoS codes to programmes or modules[13]. Text mining can be used for automatic classification of documents, given a suitable training data-set of pre-classified documents. We do, however, consider it to be highly unlikely that HE Providers would accept an approach which did not involve professional decision-making, even if a technology could be developed which was objectively as accurate as the current data suggests human coders are.

The following aims have been identified:

1. Providing decision-support at the time of HECoS code assignment to courses. This could use the content of a programme specification (or section thereof) and either raise a query if an assigned code appeared to be incorrect, or actively recommend a small number of likely HECoS codes.

2. Quality control of the use of HECoS. Similarly-coded courses should have similar descriptions, judged according to the distribution of keywords. Text mining could be used to bring cases of possible miscoding to the attention of a human being for review. This would be particularly useful while HECoS is being initially adopted.

3. Validation of HECoS. Outliers detected in the process indicated for #2 might be indicative of a deficiency in HECoS; the miscoding might have been forced by a scheme defect. This would necessarily only apply at the start of HECoS adoption and would require it to have provisional status and the mechanisms for revision and re-coding to be in place.

4. Un-coded subject analysis. Policy-makers and funders, in particular, may be interested to understand the extent to which Higher Education is delivering specialist skills which are below the level at which a programme or module would be classified. Changing political agendas mean that the subjects of interest cannot be specified up-front. This indicates that a post-facto method such as text mining is likely to be particularly useful. These questions might address cross-course skills, for example “to what extent are Big Data skills being covered across disciplines?” Alternatively, they might be tightly-focussed, for example “what is the supply of graduates with knowledge of nuclear waste management?”

5. Automatic keyword extraction could be used for two purposes:

    a. Keyword search (e.g. in Unistats or other public-facing course discovery and information services) could be enhanced using a lexicon generated from keyword extraction, especially if the keyword to HECoS code linkage is captured. This could usefully be published as open data and used for auto-completion, “see also”, search expansion, etc.
    b. Semi-automatic trend analysis in which the popularity of keywords, and the emergence of new keywords, is annually determined using a consistent method and the results openly disseminated.

4.2. Key Issue – Source of Suitable Text

Two potential sources of text have been identified for analysis: course descriptions as used in prospectuses, and programme or module specifications. Returning to the earlier statement that suitable text would be found in a “large number of documents each containing a sufficient volume of text, and of a similar genre”, we identify the genres as being an advertising description or a programme specification, and note that these two documents are stylistically quite different.

[13]We will often use “course description” as a brief, if rather imprecise, term to indicate programme or module descriptions when not wishing to be precise.

Concerning prospectus/advertising descriptions, we have some doubt that the text will be of sufficient length from some providers, and doubt that they will be rich enough in keywords to address many of the aims of text analytics listed above. *Prima facie* we would expect that programme and module specifications would be likely to be a better source of text for the aims outlined because they are more likely to contain specialist terms, given the audience and purpose for which they are created.

The QAA requires HEPs to publish programme specifications and the fact that the QAA also defines the content of programme specifications means that there is some sector-wide standardisation. Standardisation is beneficial for text mining, which might otherwise be compromised by differences in intent, leading to differences in the results of text mining. Unfortunately, in respect of programme specifications, a requirement to publish does not mean that they are easily and reliably discoverable.

The KISCOURSE table in the Unistats data download contains three fields which contain URLs: ASSURL for assessment information, CSEURL for the course page, and LTURL for teaching and learning methods. The Unistats data download URLs variously and inconsistently resolve to programme specifications, online prospectus pages appropriate for the course, or to non-specific prospectus or information pages. Accesses to some URLS returns an empty response or a "page not found" place-holder. Even if the course URLs did reliably resolve to web pages describing a single course, it would still be necessary to extract the description from among the other content on the page.

In summary: at present, there is not an available source of suitable text for mining.

4.3. Towards a Methodology for Text Mining

The specification for this report indicated it should contain a "proposed methodology for text mining". Consultation with stakeholders has indicated that there is a lack of readiness for such. Although text mining is becoming quite widely used in diverse areas of scholarship in higher education, there is relatively little appreciation of its potential among those responsible for policy and management at institutional or sector level. There is a "chicken and egg" problem. It is difficult to convince these stakeholders that the effort required to make the necessary corpus of text available is worthwhile in the absence of the results it would enable.

In principle, the easiest way forward to establishing such a corpus would be to revise the KIS data collection[14] to locate (by URL) the programme specification for each course. These could be processed to create a programme specification repository[15] as open data.

It is believed to be highly likely that, in a climate of cost control and given the raft of other changes in the HE data and information landscape, HEPs would resist attempts to require them to make programme specifications available in the ideal form for text mining; even though this would be a marginal additional effort for many HEPs, it is yet another change to manage. It is not thought that any sector body would attempt to impose such a requirement. Hence, while the project team can imagine text mining being put to good effect, and believe there is sufficient intangible RoI from the creation of a programme specification repository, we propose a less demanding approach to boot-strap evidence that might motivate such an innovation in the future.

[14]https://www.hesa.ac.uk/index.php?option=com_studrec&Itemid=232&mnl=15061 .

[15]They could be retrieved into a centralised depository recording year, outcome level (foundation, bachelors, masters), institution, and HECoS code(s). A better option than this whole-document approach would be a more structured format which is machine-readable, i.e. to separate programme aims/objectives, intended learning outcomes, and module titles, etc. in a way which enables reliable automatic processing. The XCRI specification would be a suitable base, but an alternative based on CSV and some usage conventions may be more attractive to HEPs.

Proposals:

TM1.
Undertake proof-of-concept study, using web search engine methods with manual intervention (to establish a representative but incomplete set of programme specifications), to demonstrate the use of text mining to address defined aims, compared against existing methods used.

TM2.
Work with one or more HEPs who maintain their programme specifications in a database to quantify the benefits of access to component parts – especially programme aims/objectives, intended learning outcomes, and module titles - over whole document treatments.

TM3.
HEDIIP should encourage the XCRI-CAP[16] National Roll-out (Jisc) to explore demonstrator scenarios involving text mining for subject analysis. National roll-out of XCRI[17] would *de facto* make it part of the data and information landscape for information about courses in Higher Education. There may be opportunities to shape the adoption to increase potential utility for text mining, and to exploit adoption for the aims outlined in this document.

5. Acknowledgements

We are grateful to Ioannis Korkontzelos from the University of Manchester (the host of NaCTeM) for his assistance in providing feedback and suggestions on the subject of text mining. We also appreciate the constructive support of the HEDIIP Programme Management Office, both as a sounding-board for the approach to this deliverable, and for facilitating engagement with stakeholders.

6. References

Campbell, L.M. and Ferrell, G., (2015). *HEDIIP NSCS Project Governance Model.* (NSCS Project Deliverable PD06)
Ferrell, G. and Campbell, L.M. (2015), HEDIIP NSCS Project HECoS Adoption Plan.

KPMG, (2015). *The Blueprint for a New HE Data Landscape*.

Retrieved from http://www.hediip.ac.uk/wp-content/uploads/HEDIIP_Data_Landscape_Report.pdf

Kraan, W. G. & Paull, A. (2014). *New Subject Coding Scheme; Impact Assessment and Requirements Definition.* (NSCS Project Deliverable PD01/02)

[16]The name XCRI-CAP derives from eXchanging Course Related Information – Course Advertising Profile. It is an XML standard for expressing course advertising information.
[17]The national roll-out of XCRI-CAP is for postgraduate courses and is an initiative of Jisc and Prospects. It might be useful since it means that a growing number of HEPs will be making structured course information available over the next 2-3 years. It is conceivable that this will spill-over to wider adoption for undergraduate provision. XCRI-CAP is also used in the technical specification for the Higher Education Achievement Report (HEAR). XCRI-CAP could easily be built upon to express programme specification information in a structured way (e.g. separate programme aims, assessment, teaching and learning methods, component module titles, etc.) for automated processing.

Appendix 1 – Summary of Variety in Current SBA Practice

UCAS

UCAS publishes statistics at JACS "subject group" (subject area) and "subject line" (principal subject) level (e.g. A, and A1, respectively). Courses coded as combined are given separately, as subject groups and subject lines in their own right (e.g. "Combined Sciences", and "Combs of engin/tech/building studies", respectively18). Foundation degrees are included except where stated.

Unistats and National Student Survey

NSS responses are quantified against a three level scheme, which is different from, and mapped to, JACS3 codes[19]; subject of study for "KIS Courses" data uses level 3 of this scheme. "Combined" appears as a single entry at all three levels. NSS level 1 almost aligns with JACS3 subject area: JACS3 F, L, N are each split across NSS levels, while JACS3 Q, R, T are combined into one.

HESA

HESA[20] uses a system of "apportionment" for balanced ("and"), major/minor ("with"), and triple subject combinations. 19 Subject areas are defined, which both split and combine JACS3 subject area (letter code), with an additional aggregation for all science subjects. Initial teacher training is dealt with as a special case.
HESA publishes Performance Indicators (PI) for HE on behalf of the UK HE funding bodies (SFC, HEFCW, HECFE, DELNI). PI subject headings largely map directly to JACS3 subject area (letter code), with the exception of combining Q, R and T (as NSS).

SFC

The Scottish Funding Council Statistical Bulletin[21] uses the same apportionment method as HESA for combined awards. SFC defines three subject groupings on the basis of JACS subject area (letter code) and subject lines: "controlled", "STEM", and "other".

HEFCW

HEFCW uses subject of study in its funding allocation formula on the basis of Academic Subject Categories (ASC), which are currently defined in terms of JACS3. An apportionment approach is used for major/minor and balanced combinations. Circular W14/40HE[22] defines the mapping from JACS3 to ASCs. The ASC mapping is broadly a set of groups of JACS subject areas (letter codes) with some exceptions at levels 2 and 3. It does not align well with UNISTATS.

[18]http://www.ucas.com/data-analysis/data-resources/data-tables/subject/applications-choices-acceptances-and-ratios-subject-group-2013
[19]The lookup table is published at https://www.hesa.ac.uk/unistats-dataset
[20]Definitions used in HESA published statistics are at: https://www.hesa.ac.uk/intros/studefs1213#sub
[21]http://www.sfc.ac.uk/PublicationsStatistics/statistics/higher_education_statistics/statistical_bulletins/stats_bulletins.aspx . Bulletin Appendixes describe the approach to SBA.
[22]http://www.hefcw.ac.uk/documents/publications/circulars/circulars_2014/W14%2040HE%20Higher%20Education%20Students%20Early%20Statistics%20Survey%202014_15.pdf

HEFCE

HEFCE defines, and publishes[23], a set of subject areas for strategically important and vulnerable subjects (SIVS) based on a combination of JACS3 subject area (letter code) and principal subject (letter+number) and maps these to cost centre codes also. The HECFE definition of STEM differs slightly from the SFC definition in the treatment of subject areas C and D.

SLC

SLC appears not to publish any SBA.

Non-Sector Publications

Three non-sector publications are widely used: The Complete University Guide (CUG), The Times and Sunday Times Good University Guide, and The Guardian University Guide.

These are largely constructed as aggregations at principal subject (letter+number) level, but differ between providers. CUG and particularly The Guardian omit a number of JACS3 codes.

HESA currently maintains a mapping from JACS3[24], which provides the detail.

[23]http://www.hefce.ac.uk/media/HEFCE,2014/Content/Analysis/Supply,and,demand/coverage-and-definition.xlsx

[24]https://www.hesa.ac.uk/dox/informationProvision/Subjectmapping_Sept2014.xlsx This is the September 2014 version; the mapping is periodically updated to keep pace with change.

Appendix 2 – Implementation Issues

This section contains some matters which would be relevant to implementation but which fall outside the focus of the body of this document, which is at a higher level of objectives and process.

Improved visibility of the various mappings/associations/aggregation-rules which relate to HECoS, and provision of them in user-friendly formats[25], can support convergence of practice with the Higher Education sector beyond what may be required by regulatory bodies. This convergence will lead to an ambient improvement in the data and information landscape, supporting more efficient data-related processes and promoting more effective use of data. The core idea here is that the more findable and usable the definitions of subject groupings are made, the more they will be used.

Proposals:

IMPL1.
Disseminate a range of formats suitable for different stakeholders. Avoid PDF as being the sole, or definitive, format since that requires re-keying for processing.

IMPL2.
Use standardised formats to communicate subject grouping rules. This could usefully include both a formal representation using semantic web technologies as well as documented conventions for using CSV for easy loading into multiple technologies.

IMPL3.
Disseminate the various mappings/associations/aggregation-rules through a central portal (by hosting master copies or using external references as appropriate), whether they are formally governed or not.

IMPL4.
An archive of current and historical versions, with dates of applicability, should be maintained, and change clearly communicated.

IMPL5.
The standard cross-sector aggregation scheme should be transparent to users of HECoS at the point of use, so far as this is possible, since users want to know how coding decisions influence down-stream analysis.

IMPL6.
Explore the provision of services through the HESA/Jisc HEIDI+/labs initiative to support HEPs in transitioning from JACS3 to HECoS. E.g. support for recoding, QA, bridging time-series data.

[25]It is assumed different formats would be considered “friendly” by different stakeholders. We need to accommodate users with a preference for printable documents through to database administrators or software developers who might prefer a more arcane format.

Appendix 3 – Consultation Summary

This section summarises those points made during the stage 2 consultation from Feb to May 2015 which specifically relate to, and have been contextualised within the scope of, this report on subject based analysis and text mining.

	Issue	Raised by*	Comment
1.	Mapping from JACS3 to HECoS (explicitly or presumed to be related to assignment of codes)	Many	Not in scope for this document. The adoption plan will specify requirements for a migration aid which relates JACS3 to HECoS.
2.	Consistent subject [aggregation] hierarchies are important for benchmarking across the sector (HEP-level strategy and planning and cross-sector bodies). Comparability is key. Improved benchmarking and HEIDI mentioned. Multiple hierarchies should be discouraged.	Manchester U, Northumbria U, Gloucestershire U, Edinburgh Napier U, City U, HESA, HESPA, UCAS U Greenwich, Northampton U, Birmingham U, Salford U, U of York, Sheffield Hallam, Edge Hill U, U of Warwick, Goldsmiths, Oxford U, U of Sunderland, Bangor U, U of Birmingham, UEA, Prospects	Identified as the key driver in this document.
3.	Concerned changes will have an impact on trend data (time-series)	Canterbury U, Gloucestershire U, St. Andrew's U, HESPA (implied), UCAS, U of Greenwich, U of Sunderland, Prospects, Liverpool Hope U, U of Warwick, Goldsmiths, U of Surrey, Plymouth U, Oxford U, U of Sunderland, Bangor U	An approach to mitigating this issue is proposed.
4.	Concern over lack of hierarchy leading to divergent practice.	Cardiff U, St Georges, London, HEFCE, Northumbria U, UCLAN, Aberystwyth U, Edinburgh Napier U, U West London, HESPA, Aberdeen U, Birmingham U, U of Chester, Newcastle U, U of York, U of Warwick, U of Birmingham, ANG	The absence of a hierarchy in the consultation scheme (HEDIIP NSCS Structure and Candidate Scheme) reflected a separation of concerns and a focus on the definitions foremost. The process to achieve the default aggregation rules which would address this need is the purpose of this document.
5.	Should be centrally imposed hierarchies for use by 3rd parties	Aberystwyth U, Edge Hill U, (implied by many others)	We have avoided “imposition” in favour of combined landscape governance and consensus.
6.	There should be clear responsibility for hierarchies (strong governance)	Manchester U, St Georges, London, HECFE, UCLAN, Edinburgh Napier U, U West London, HESPA, U Greenwich, Edge Hill U, Bangor U, U of Birmingham, ANG	Indicated in proposals.
7.	HESA should take the lead in defining hierarchies	St Georges, London, Aberystwyth U, HESPA, Sheffield Hallam	Proposal refers to the recommendations of the KPMG landscape study as the assumed blueprint for governance.
8.	Desire for allocation of HECoS terms to UoA &/or Cost Centres	Swansea U, Northumbria U, Gloucestershire U, UCLAN, City U, Liverpool U, HESPA, Aberdeen U, U of York, U of Birmingham, UEA,	In scope of this document, although not originally identified in the work-plan. The recommendation is that these allocations are not created under the aegis of HEDIIP.
9.	League tables identified, including need to engage with these providers and the problem of misleading results. One comment proposed governance body veto of league table aggregations.	Northumbria U, Aberystwyth U, HESPA, ANG, Salford U, U of Sunderland, Newcastle U, U of York, Sheffield Hallam, U of Sunderland	This document recommends a light-tough engagement. See also #5
10	Aggregation at greater detail than the broadest JACS3 groups (e.g. for DLHE) is desirable in order to make meaningful decisions.	Canterbury U	The default aggregation rules proposed accommodate this, but it is ultimately a matter for data processors to meet the need for statistics.
11	HECoS flat list makes burden for owners of custom aggregations when new terms are added because a conscious decision is required	HEFCE	This is not explicitly dealt with in this document but it could be avoided if custom aggregations are expressed in terms of the third level of the default aggregation hierarchy, which we

	Issue	Raised by*	Comment
			propose is maintained alongside the process for addition or deprecation of terms. We also expect that the governance process will avoid rapid change.
12	It would be useful to cross-index HECoS to QAA benchmark statements	Leeds U	Included in this document.
13	Not in favour of parallel running of HECoS and JACS3, in favour of synchronised adoption in one academic cycle.	HESA	This report does not recommend an extended period of parallel running. Synchronised adoption is recommended.
14	Need to communicate conceptual distinction between HECoS navigation structures and groupings for statistical aggregation. Call for guidance on structures.	HESA, Aberdeen U	Noted. Distinctions outlined in this document but dealing with this issue implies much more than a document, requires thorough orientation of a spectrum of guidance material.
15	Publish aggregation groupings as part of the vocabulary	HESA	Included in proposals insomuch as a central dissemination channel for HECoS and aggregation rules is proposed. It is not proposed this would be the exclusive source of all aggregation rules in use in HE.
16	Benefits from transparency in aggregations in use. E.g. central directory/register should be public and accessible. Some references to e.g. “open data template” HEIDI mentioned	HESA, U of York, U of Surrey, U of Birmingham UEA	Included in proposals, including explicit mention of aggregations/associations that would not be under the same governance regime as HECoS.
17	NSS levels 1, 2, 3 recommended. Complex procedure around KIS identified; strong case to minimise impact.	HESA, Newcastle U	Adopted as a proposal.
18	STEM agenda referred to	Prospects	Included in this document, noted as being problematic for a centrally-managed approach.
19	Role of subject groupings in IAG; guidance queries tend to be broad	Prospects	Not explicitly considered. Following KIS/Unistats groupings, which are very broad at levels 1 and 2, gives some connection to IAG.
20	ATAS mentioned	U of Warwick	Included in this document

** - affiliation not given is indicated by “ANG”*

Appendix 4 – Analysis and Planning Stakeholder Meeting

On June 30th 2015, a meeting of analysis and planning stakeholders took place to discuss proposals for text mining and subject based analysis that had been developed through a combination of desk research and synthesis of responses made during the public consultation which took place from February to May 2015 (which is summarised in Appendix 4).

In addition to NSCS project team members, the meeting involved:

- Andy Youell – HEDIIP
- Paul Baron – HEDIIP
- Richard Puttock – HEFCE
- Jonathan Waller – HESA
- Christine Couper – Greenwich University. Member of HESPA Executive
- Gordon Anderson – SFC
- Hannah Falvey – HEFCW
- Mark Corver – UCAS
- Michael MacNeill – DEL
- Steve Riddell – SFC

Key points of agreement:

1. A common approach to aggregating published statistics across differently-coded subjects was supported by all participants. This would not preclude organisations from conducting analysis and publishing statistics using alternative approaches in addition to the common approach.

2. The scope of a common approach should be limited to “key features” 1 and 2 (as outlined section 2.2.).
3. HEPs and HE Core Sector Bodies are not sufficiently motivated to take collective action on text mining.

Proposals which have been removed in moving from the draft proposals on SBA, as presented to the meeting, to this document (PD05):

1. It was agreed that, while some HEPs may operate JACS3 and HECoS coding in parallel, and that this could be of benefit in quality assurance during the transitional period, this was not something that should be required of them.

Appendix 5 – Simple Illustration of Text Mining

Since text mining is not widely applied in the management of HE, this section is intended to give an outline of what the results of simple text mining look like.

In the absence of a repository of programme specifications, web search engines were used with URLs intended to generate pages of search results which contain only programme descriptions. It is possible to use this approach to get search results which mostly contain programme descriptions using a URL of the form:

```
https://www.google.co.uk/search?q=%22programme+specification%22+site%3Abolton.ac.uk+filetype%3Apdf&num=100&start=100
```

This is not reliably complete (some specifications may not be recalled) and demonstrably not suitably specific (irrelevant documents are retrieved, for example guidance on producing programme specifications). This approach also mixes specifications with a different qualification level, modules and programmes, and takes no account of years of applicability. The URL given also assumes publication in PDF format (which improves the relevance of the search results but the assumption is not valid across HEPs).

Data used

This approach was used to gather specifications from: University of Bolton, City University London, and The University of the West of England. There was some manual intervention to filter irrelevant documents after which 407 documents were submitted to a text mining script written in R and using the tm package. 29 documents were found to be scanned images and could not be processed. After stemming and stopword removal, 8277 terms were identified, of which 3069 appeared only once. Acronyms/abbreviations, typographic errors and module codes are evident.

Experiment 1 – documents containing a term

To look for pharmacology/pharmacological (etc), all documents containing the stemmed form “pharmacolog” were identified. As a specialist term, we would expect this to be quite reliable in identifying relevant programmes and the document filenames were found to be:

```
AdvDipCPDV300·pdf
APM021-Enhancing-Critical-Care-Skills.pdf
ClinicalPracticeMSc.pdf
IndependentPharmacistPrescribingConversionCourse.pdf
NMM024-Medicines-Management.pdf
NMM083-Practice-Based-Module.pdf
NMM110-Independent-Supplementary-Non-Medical-Prescribing.pdf
NNMWIF-PG-Dip-Midwifery-78-week.pdf
NonMedicalPrescribingHE6.pdf
NursePrescribingProgrammeV150.pdf
PSAHNR-MSc-APHSC-Nursing.pdf
PSNACN-Advanced-Nurse-Practitioner-Adult-Child-Neonate.pdf
PSNUAP-MSc-Nursing-Advanced-Nursing-Practice.pdf
PSOPNU-MSc-APHSC-Ophthalmic-Nursing.pdf
USMIDB-BSc-Midwifery.pdf
USNSNS-USNSCC-USNSHD-USNSNE-USNSEM-USNSIN-BSc-Nursing-Studies-Top-up.pdf
USNSNS,-USNSCC,-USNSHD,-USNSNE,-USNSEM,-USNSIN,-USNENT,-USNLMT-Bsc-Nursing-Studies.pdf
```

It is evident that this list contains module specifications, as expected from a web search engine approach to gathering source documents. It is also evident that pharmacology is a subject of study in courses that would not be classified as pharmacy using JACS or HECoS.

This approach is effectively the same as indexing and search and for which the opportunities for statistical analysis afforded by text mining are wholly unexploited.

Experiment 2 – finding distinctive terms

Term frequency inverse document frequency (TF-IDF) statistics were computed and the 0.2% most distinctive terms, according to this measure, were inspected.

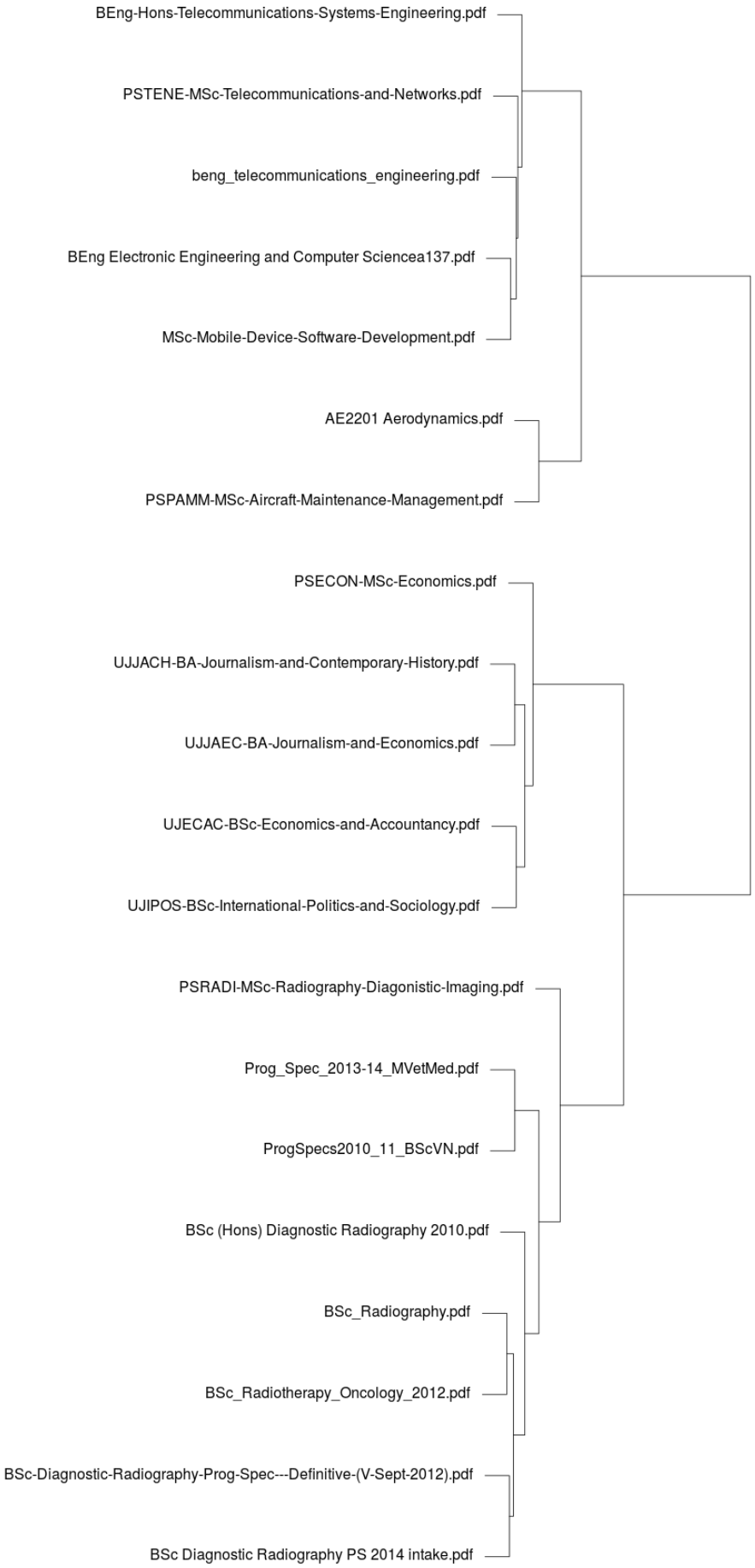
Example 1.
The document with the most distinctive terms across the whole set is, not surprisingly, a module specification. It is for a module on aerodynamics and contains these distinctive terms (in stemmed form): aerodynam, aircraft , cruis , drag , flight , flowfield , inviscid , isbn , land , layer , lift , longitudin , stabil , takeoff , vehicl , viscous , wing. With the exception of "isbn", these appear to have accurately captured appropriate keywords.

Example 2.
A programme that had been coded B150 (generally used for human biology, although not in the current JACS3 scheme) was seen to contain distinctive terms (in stemmed form): biolog, clinic, diseas, dispens, eye, goc, lens, ocular, optic, optometri, optometrist, patient, preregistr, registr, street, vision, visual. The programme is, in fact a BSc Optometry, and the mis-coding could be automatically identified on the basis of clustering since this programme specification shares many similar distinctive terms with courses properly coded as optometry.

Experiment 3 – clustering on distinctive terms

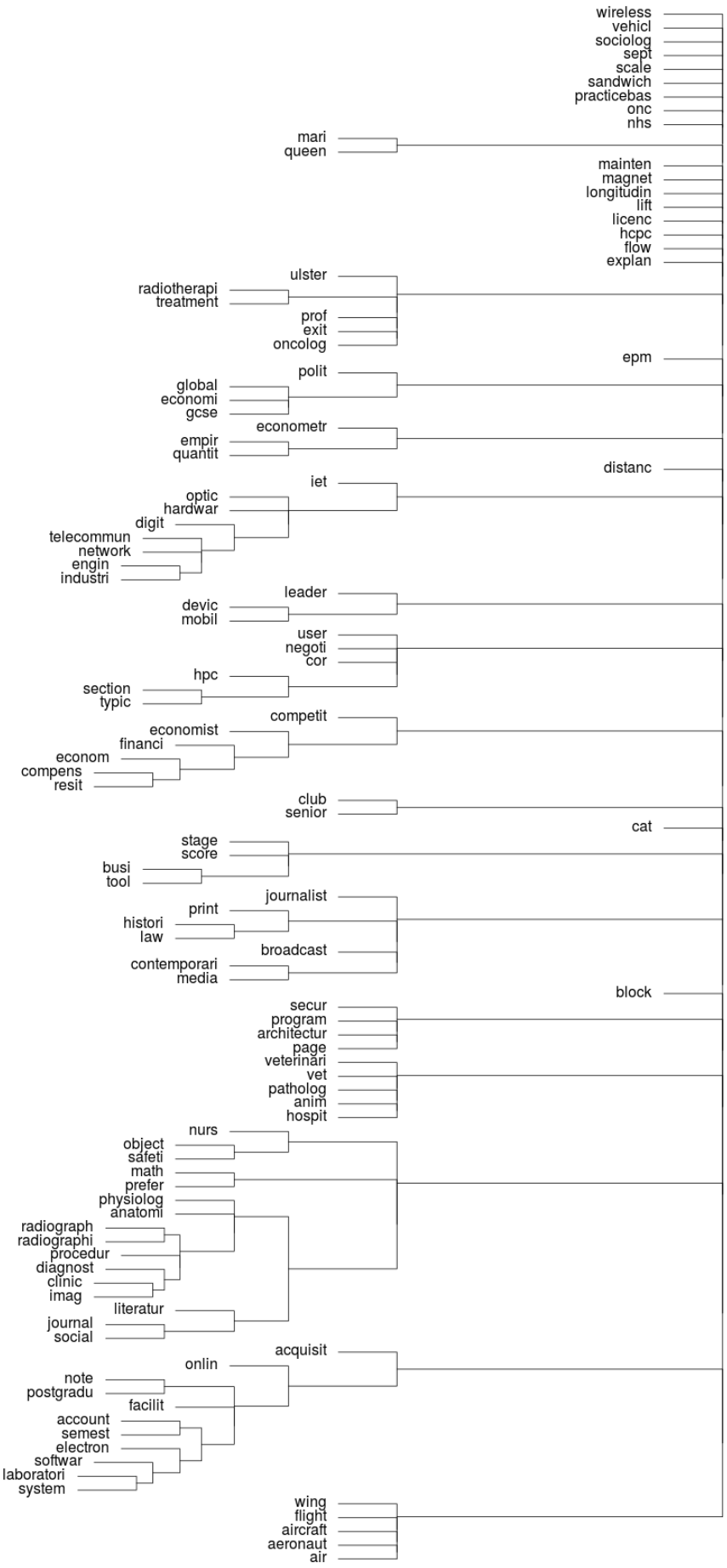
For this experiment, a small sample of 20 documents was used, using the previous set plus the results of a web search for programme specifications dealing with diagnostic imaging, radiography, radiology and telecommunications. The aim of creating the set was to have a range of similarity and difference in subject of study. Term frequency inverse document frequency statistics were computed and the 0.5% most distinctive terms, according to this measure, were used. Pair-wise document similarity was computed according to the number of distinctive terms in common. A hierarchical clustering algorithm (hclust in the R Core) was used to automatically group documents according to similarity (strictly, a distance measure was used, computed as 1/similarity). The same approach was used to identify how terms are related according to their co-occurrence in documents.

[Diagrams on next page]



Dendrogram of similar programme specifications.

The length of the horizontal lines indicates extent of difference. To the left the difference is between documents, and moving right it is between clusters of documents. Notice that Aerodynamics and Aircraft Maintenance Management are correctly associated, but are also connected more distantly to a cluster dealing with a different branch of engineering, while engineering subjects are clearly separated from the rest of the tree. The clustering appears to have also correctly separated veterinary courses from others in a broadly similar domain. Courses in journalism have been similarly separated from, but remain close to, related subjects of study. Clustering has also correctly associated the BSc Radiography with Radiotherapy and Oncology; the former does indeed include therapeutic uses, whereas the other radiographic courses have a diagnostic focus.



Dendrogram of distinctive terms.

This diagram shows how terms (in their stemmed forms) are associated through co-occurrence in documents. Terms on the far right are un-related. The associations largely match intuition from the document titles. A larger sample of documents would give a more reliable result; it can be seen that the clustering has failed to associate econometrics with economist.