# A TAXONOMY OF PLAGIARISM IN COMPUTER SCIENCE

**Mike Joy[1], Georgina Cosma[2], Jane Sinclair[1], Jane Yin-Kim Yau[1]**

[1]Department of Computer Science, University of Warwick
Coventry, UK

[2]Department of Business Computing, P.A. College
Larnaca, Cyprus

*m.s.joy@warwick.ac.uk, g.cosma@faculty.pacollege.ac.cy, jane.sinclair@warwick.ac.uk, j.y-k.yau@warwick.ac.uk*

## Abstract

Many on-line resources exist for testing students' knowledge of plagiarism, however few of these cover both text and source code plagiarism in a comprehensive manner to encompass all types of plagiaristic activity of relevance to computing students and academics. In order to provide suitable resources it is useful to identify and categorize aspects of text and code plagiarism so that, for example, quizzes can be generated which ensure coverage of each important topic. This paper reports the results of a taxonomic analysis of data collected from sources relating to plagiarism, including existing on-line quizzes and previous research, in order to inform the construction of a quiz generation system which covers all areas of plagiarism relating to a computing course.

The principal aim of this research was to identify a taxonomy of issues relating to student (and academic) plagiarism, so that a resource could be built which can accurately assess a student's understanding of what plagiarism means and how it can be avoided. Such a resource would target computing students, and cover source code topics in addition to the generic plagiarism issues of importance to students in other disciplines. The taxonomy reported here allows us to construct representative question sets for use in such a resource, and to present formative material to students which addresses their individual misunderstandings.

Our methodology for constructing the taxonomy initially involved collecting data from two types of source. The first consisted of on-line interactive resources, such as student-focused plagiarism questionnaires which were created for testing students' knowledge of plagiarism and for providing informative feedback to students based on their responses. We identified 23 which were publicly accessible, and which together contained 268 questions. The second type of source data is represented by published work, and included books on plagiarism and on "academic misconduct", and conference and journal publications, some of which focus on source code plagiarism.

The quiz data were analyzed using facet analysis, in order to identify discrete categories into which the questions might be classified. This provided a comprehensive overview of what types of question have been used for testing students' understanding of plagiarism in a generic context. The other data were then used to refine the classification by incorporating the major issues which currently are important to computing students and academics.

The resulting taxonomy consists of 6 categories (Plagiarism and copying, Referencing, Cheating and inappropriate collaboration, Ethics and consequences, Source code plagiarism, and Source code documentation) subdivided into 23 subcategories.

At the time of writing, an online tool has been written and contains both tutorial material and over 200 questions arranged according to the above categorization. Although primarily the tool generates quizzes relating to source code plagiarism, it can be adapted to generate quizzes relating to other topics.

***Keywords -*** Computer science education, plagiarism, taxonomy, source code

## 1 INTRODUCTION

Plagiarism is a growing concern in universities. Material can be obtained and copied from various sources including the Internet, essay and source code banks, and text books. In a recent survey by Nadelson [20], 72 academics commented on issues surrounding academic misconduct in a large University, and reported 570 suspected incidents, involving 460 undergraduate and 110 graduate students. Of these, 173 were 'accidental/unintentional plagiarism' (134 involving undergraduate

students and 39 graduate students). In addition, a large number of incidents were reported where academics suspected that students had submitted papers copied from the Internet. Other forms of academic misconduct reported were 'purposeful plagiarism', 'class test cheating' and 'take home test cheating'.

Plagiarism can relate to material which is not text, and in the computing disciplines there is a particular concern with students copying computer programs (source code) which they then submit as their own work in programming assignments [3].

Recent studies suggest that incidents of plagiarism are increasing and that the problem is endemic [12, 13], and on-line resources have become available which allow computing students to hire expert coders to complete their programming assignments [15]. Such opportunities make plagiarism easier for students, and concerns about the ease of access to on-line material have been expressed in a number of studies [17, 21].

Much work has been done to alert students and to advise teachers as to how plagiarism may be avoided and prevented, and instances of it detected. In addition to books on the subject [2, 11] there is a growing corpus of web resources produced by individual institutions. In the UK, for example, the JISC Plagiarism Service (jiscpas.ac.uk) provides advice and a portal to the Turnitin® plagiarism detection software, and advice at a discipline-specific level is provided by the Higher Education Academy subject centres (www.heacademy.ac.uk).

Many of these resources exist to test students' understanding of plagiarism, in the form of on-line quizzes that provide feedback to students based on their responses. The common aim of such material is to educate students on plagiarism related issues and reduce the prevalence of plagiarism. However, the scope of these interactive resources is variable, and few address the issue of source code plagiarism in any depth. Furthermore, many of the quizzes fail to cover important areas, such as self-plagiarism, and other discipline-specific issues. There is evidence that these are the topics which students find most confusing [16], and this suggests that there would be a benefit if there were to be more coverage of those topics in such quizzes.

The principal aim of this research is to identify a taxonomy of issues relating to student (and academic) plagiarism, so that a resource can be built which can accurately assess a student's understanding of what plagiarism means and how it can be avoided. The resource contains a quiz which is designed to work in two modes – *formative* and *summative*. In formative mode, a student will be presented with randomly selected questions relating to plagiarism, and offered substantial feedback on their responses. In summative mode, the student is presented with a *representative* set of questions which cover all important aspects of plagiarism in computing, and their response to those questions measures the student's understanding of the issues. The quiz targets computing students, and covers source code topics in addition to the generic plagiarism issues of importance to students in other disciplines.

The taxonomy reported here allows us to construct such representative question sets, and to present formative material to students which addresses their individual misunderstandings.

## 1.1 Plagiarism and Cheating

Plagiarism in academic institutions is often expressed as copying someone else's work (i.e., another student's or sources such as books), and failing to provide appropriate acknowledgement of the source (i.e., the originator of the materials reproduced). The act of plagiarism is still regarded as an offence regardless of whether it was intentional or unintentional. Hannabuss [14] defined plagiarism as "the unauthorized use or close imitation of the ideas and language/expression of someone else". In the context of academic work, plagiarism can range from the citation of a few sentences without acknowledging the original author to copying an entire document.

Plagiarism can take many forms, and various different definitions and descriptions of plagiarism can be found. The fact that these do not always identify the same activities as being plagiaristic indicates that there can be a lack of clarity about exactly what constitutes an offence. Martin [19], for example, sets out the following six forms of plagiarism.

- *Word-by-word copying*, which involves directly copying sentences or chunks of sentences from other people's work without providing quotations and/or without appropriately acknowledging the original author.

- *Paraphrasing*, which involves closely rewriting (i.e. only changing some of the words but not making enough changes) text written by another author without appropriately citing the original author.
- *Plagiarism of secondary sources*, which involves referencing or quoting original sources of text taken from a secondary source without obtaining and looking up the original source.
- *Plagiarism of the form of a source* is when the structure of an argument in a source is copied without providing acknowledgements that the systematic dependence on the citations was taken from a secondary source. This involves looking up references and following the same structure of the secondary source.
- *Plagiarism of ideas*, which involves using ideas originally expressed in a source text without 'any dependence on the words or form of the source'.
- *Blunt plagiarism* or *authorship plagiarism* which is taking someone else's work and putting another's name to it.

These relate to the forms of plagiarism relevant to student behavior (and in fact Martin goes on to argue in favor of a broader perspective on plagiarism). In institutional guidelines there tends to be general agreement and coverage of the main points here but some aspects (such as secondary source plagiarism) become "grey areas" which are less frequently referred to. Further, even the main topics can be stated in different ways or in ways which leave room for interpretation. Added to this are the areas not mentioned here such as self-plagiarism and (for computing subjects) code reuse. It is little wonder then that research into students' perceptions and attitudes towards plagiarism report that they find definitions of plagiarism "contradictory, unclear and confusing" [5]. Research into the subject-specific area of source code plagiarism [16] shows that students are even more unclear about the boundaries here.

Our intention in producing a taxonomy is to map out the area with particular reference to Computer Science, and to provide a framework within which student resources (such as example banks or quizzes) can be defined. Concrete examples (which students find extremely beneficial) can be related to each specific area of plagiarism, and additional practice in "problem areas" can be provided if necessary.

Plagiarism occurs when a student uses someone else's work and submits it as his own work, whereas cheating refers more generally to actions taken by a student to fraudulently obtain academic credit. The resources which are already available tend not to make a clear distinction between the two concepts. In order for our work to be inclusive, our investigation includes "non-plagiarism" cheating in the taxonomy.

## 2 METHODOLOGY

Our methodology initially involved collecting data from as many sources as was practical. There were two types of sources we consulted.

The first source type consisted of *on-line interactive resources*, such as student-focused plagiarism questionnaires which were created for testing students' knowledge of plagiarism and for providing informative feedback to students based on their responses. We identified 23 which were publicly accessible, and which together contained 268 questions. The sources of these quizzes were Bradford University, California State Polytechnic University (2 resources), Canterbury Christ Church University (5 resources), Cardiff Metropolitan University, Cardiff University, Drexel University, Fairfield University, Howard University, James Madison University, Northwest Missouri State University, Penn State University, Rutgers University, Simon Fraser University, Spring Hill College, St Hubert Catholic High School, University of Maryland University College, University of Southern Mississippi, and Wayne State University.

The second type of source data is represented by *published work*, and included books on plagiarism and on "academic misconduct" [2, 4, 10], and conference and journal publications. Some of these, such as [12], focus on source code plagiarism. Decoo's book briefly discusses software plagiarism at the levels of user-interface, content and source code [4], and a recent study by Cosma and Joy [3] identified the perceptions of academics in the UK as to the plagiarism issues which are specific to the computing disciplines.

The quiz data were analyzed using facet analysis [1], in order to identify discrete categories into which the questions might be classified. This provided a comprehensive overview of what types of question have been used for testing students' understanding of plagiarism in a *generic* context. The other data were then used to refine the classification by incorporating the major issues which currently are important to computing students and academics.

## 3 FACET ANALYSIS

A faceted classification scheme [1] allows each item to be categorized according to a number of different relevant features (or *facets*) which are independent of each other. This allows a greater degree of flexibility than a simple enumerative classification (such as the Dewey Decimal library classification system) in which each item is expected to fit into exactly one classification "slot". Each facet acts as a mini-categorization and the overall classification of an item is the profile provided by the collection of classifications by each relevant facet.

The first step in the process is to identify suitable facets by analyzing the domain. This has been described as a "journey of discovery" [17] since there is no "right" or "wrong" collection of facets, simply those which differentiate the material best for the purpose in hand. In the present case, the analysis clearly suggested that each question could be usefully classified according to *four* facets.

We initially observed that an understanding of plagiarism depends on both *intrinsic* and *extrinsic* factors. Intrinsic factors relate to the *process* of plagiarizing – for example, where students copy material from, how it is sourced and copied, and how the copying is disguised. Extrinsic factors relate the plagiarism activity to external influences, such as other forms of cheating and the reasons why students may be tempted to plagiarize. Whilst the extrinsic factors are important, they relate to the wider educational process, and are not the principal focus of our activity. We therefore assigned extrinsic factors to one facet and concentrated on identifying further facets within the intrinsic factors.

- The first relates to *sources* of plagiarized material, and includes the Internet, books, media (Radio/TV), encyclopedias, lecture materials (slides/notes), and other people.
- The next facet enumerates the types of *actions* which the plagiarizing activity involves. These include copying, paraphrasing, incorrect referencing, using quotation marks, collaborating and translating.
- The third facet relates to the *type of material* which is used in the activity. This is varied, and includes text (free text, source code and source code comments), common knowledge, abstract or concrete ideas, data (specifically statistics, tabular data and mathematical solutions), graphics (figures, diagrams and images), music (lyrics and sheet music), and spoken words.
- The final facet is *extrinsic*, and relates to the context within which the plagiarism takes place. This includes other forms of cheating (such as stealing material and contract cheating [8]), the reasons why students may plagiarize, and ethical considerations.

The data we collected from the other sources were then scanned to identify any substantive issues which were excluded from the quiz analysis, and as a result we were able to determine a taxonomy which excludes source code issues. To this we then added the sources (online messages, such as IM and email), materials (online messages, source code) and actions (such as code editing) which are unique to programmers. These are illustrated in Table 1.

As stated above, the main motivation for this categorization exercise was to support an online educational resource. The following observations are relevant in this context.

Firstly, the above analysis identifies four facets which are independent of each other, but it does not indicate the relevant importance of each (or of the topics to which each relates). For example, a computing student is unlikely to be tempted to translate musical lyrics obtained from a radio station as part of an academic assignment, whereas copying source code obtained from a web site is a greater risk.

Secondly, for obvious practical reasons, it would not be reasonable to use each of the many permutations of source / material / action / extrinsic as a classification for questions within such a tool. Rather, the student user needs to be presented with a small collection of important issues which they can focus on.

**Table 1.** Output of data analysis

| | |
|---|---|
| Sources | Books (incl. encyclopedias)<br>Internet<br>Media (Radio/TV)<br>Lecture materials (slides/notes)<br>Persons |
| Actions | Copying (incl. self plagiarism)<br>Paraphrasing<br>Referencing (using quotation marks, ensuring reference accuracy)<br>Translating<br>Avoidance strategies<br>Code editing |
| Material | Text (free text, source code, source code comments, online messages)<br>Common knowledge<br>Ideas<br>Data (statistics, tabular data, maths solutions)<br>Graphics (figures, diagrams, images)<br>Music (lyrics, sheet music)<br>Spoken words |
| Extrinsic | Hiring someone to do the work<br>Stealing someone else's work<br>Collaborating<br>Ethics<br>Consequences and punishments<br>Other cheating (e.g. faking data) |

The taxonomy was therefore modified to generate 6 categories, subdivided into 23 subcategories, as illustrated in table 2. Each of these sub-categories contains a description, such that each of the permutations is described by one and only one category. Furthermore, for each sub-category it is – in principle – straightforward to articulate quiz questions to identify whether a student has understood the issue that the sub-category refers to.

## 4 DISCUSSION

Through analysis of existing resources we developed a faceted taxonomy to describe different features of plagiarism. Our aim was to partition the domain as a necessary precursor to providing structured support material which would cover the full range of possibilities and ensure suitable coverage of all identified aspects. In particular, our aim was to use the taxonomy as a basis for a “quiz generator” which would use a database of questions. The classifications could be used to generate quizzes in different ways. For example, a general quiz could be automatically constructed to cover all topics while a specialized quiz could be made to provide extra practice with a less well-understood topic or topics.

There is no such thing as a “correct” taxonomy. There are generally many different ways in which classifications could be developed. In general, a faceted taxonomy should allow for unambiguous classification by identifying mutually exclusive, clearly distinctive facets. These facets should be easily recognizable and represent aspects of the domain which are important in the required context. There is inevitably an element of subjectivity in the choices made. It is therefore more appropriate to ask whether a given taxonomy is fit for purpose and provides a useful framework for the work it supports. In addition, the domain to which a taxonomy relates is constantly evolving, and no more so than in our discipline. Hence it is useful to work with an extensible approach such as facet analysis which may be further developed in the future if required.

**Table 2**. Taxonomy categories and sub-categories

| | |
|---|---|
| (1) Plagiarism and copying | Ideas: referencing people's experiences, impressions, ideas and inspirations (which are not stored as a document which can be referenced) |
| | Facts: referencing commonly known facts, such as basic mathematical facts, common geographical and historical facts |
| | Speech: referencing someone saying something (e.g. referencing the words of a TV presenter in a documentary, a story told by a friend, or what was said while interviewing a friend) |
| | Copying: identifying what constitutes copied material (text, figures, images) from various sources of information, and which should be referenced. |
| | Paraphrasing: acceptably paraphrasing text or editing diagrams |
| | Self-plagiarism: referencing work that was previously submitted for academic credit (or publication) |
| | Avoidance strategies: good practice for avoiding plagiarism |
| | Translating text: translating text between languages |
| | Email: copying words from email, IM, or other personal contacts |
| (2) Referencing | Referencing: correctly referencing, placing quotation marks where appropriate, and citing in appropriate formats |
| (3) Cheating and inappropriate collaboration | Collaboration: identifying when it is acceptable for students (or groups of students) to collaborate and sharing work |
| | Purchasing: purchasing academic material such as essays or hiring experts to write essays or source code (contract cheating) |
| | Cheating: Other cheating issues (not necessarily called “plagiarism”), such as falsification and fabrication |
| (4) Ethics and consequences | Ethics: understanding the relevance of ethical behavior, copyright and fair use related to plagiarism |
| | Consequences: consequences (punishments, etc.) when students are caught |
| (5) Source code plagiarism | Adapting source code: adapting (modifying) source code written by other programmers |
| | Open source: using and referencing Open Source code |
| | Copying source code: using and referencing source code written by other programmers |
| | Code generation: referencing source code which has been automatically generated |
| | Translating code: translating source code between programming languages including algorithms written in pseudo code or diagrams such as UML |
| (6) Source code documentation | Documentation: copying comments in source code or other documentation |
| | Designs: copying source code or interface design material and reverse engineering |
| | Testing: copying test data and/or test strategy |

The main purpose for which our taxonomy was designed was to enable a quiz generator to be developed. The aim of this was to generate online materials (tutorials and quizzes) to be targeted at computing university students, in particular those which are undergraduates and/or international students. The rationale for this was that many of these types of computing students lacked a full understanding of what constitutes source code plagiarism, as established previously in [16]. Additionally, there is a current lack of online resources which can help these students learn interactively and/or test their existing knowledge on source code plagiarism as well as self-plagiarism.

At the time of writing, an online tool has been written and contains both tutorial material and over 200 questions arranged according to the above categorization. Although primarily the tool generates quizzes relating to source code plagiarism, it can be adapted to generate quizzes relating to topics associated with other university courses.

The tool has been coded using the Moodle platform (moodle.org). Since Moodle is Open Source, this provides us with a resource which is in principle portable to other institutions and departments. In formative mode, the tool consists of 6 “mini lessons”, corresponding to the six categories in Table 2, each of which contains a (short) discussion, a set of FAQs, and an interactive quiz with questions focused on that category. In summative mode, a set of questions is randomly generated such that at least two are taken from each of the six categories (this figure is configurable), and the student is required to answer all of them. A high mark (the exact value is left to the discretion of the teacher) should therefore provide evidence that the student has understood the issues involved in plagiarism in the computing disciplines.

An evaluation of this resource will take place over the coming months, and will form the focus of a future paper.

Although this tool was the primary motivation in the review of plagiarism material and in the development of the taxonomy, the resulting categorization provides a general resource which may be useful both in providing a framework for discusión of plagiarism in the computing disciplines, and also as a basis for other online resources.

## 5 ACKNOWLEDGEMENTS

The authors thank the Higher Education Academy Subject Centre for Information and Computer Sciences for funding for this work.

## References

[1] Broughton, V. 2004. Essential Classification, London, Facet Publishing, 257-283.

[2] Carroll, J. 2007. A Handbook for Deterring Plagiarism in Higher Education (2nd edition), Oxford Centre for Staff and Learning Development.

[3] Cosma, G., and Joy, M.S. 2008, Towards a Definition on source code Plagiarism, IEEE Transactions on Education 51, 2, 195-200.

[4] Decoo, W. 2002. Crisis on Campus: Confronting Academic Misconduct. MIT Press.

[5] Devlin, M. 2004. Policy, preparation, prevention and punishment: One faculty's holistic approach to minimizing plagiarism. Educational Integrity Conference, Adelaide, Australia, 2004. Online: http://www.deakin.edu.au/dro/view/DU:30006770

[6] JISC 2004. Proceedings of the 1st International Plagiarism Conference. Online: jiscpas.ac.uk/conference2004.

[7] JISC 2006. Proceedings of the 2nd International Plagiarism Conference. Online: jiscpas.ac.uk/conference2006.

[8] JISC 2008. Proceedings of the 3rd International Plagiarism Conference. Online: jiscpas.ac.uk/conference2008.

[9] Lancaster, T., and Clarke, R. 2007. Assessing Contract Cheating through Auction Sites – a Computing Perspective, In Proceedings of the 8th Annual Conference of the Higher Education Academy Subject Centre for Information and Computing Sciences (Southampton, UK, August 28-30, 2007), 91-95.

[10] Roberts, T.S. 2008. Student Plagiarism in an Online World: Problems and Solutions. IGI Global.

[11] Carroll, J. and Appleton, J. 2001. Plagiarism: A good practice guide. Online: www.jisc.ac.uk.

[12] Culwin, F., MacLeod, A., and Lancaster, T. 2001. Source code plagiarism in U.K H.E computing schools, issues, attitudes and tools. Technical Report SBU-CISM-01-02, South Bank University, London, September 2001.

[13] Dey, S., and Sobhan, A. 2006. Impact of unethical practices of plagiarism on learning, teaching and research in higher education: Some combating strategies. In 7th International Conference on Information Technology Based Higher Education and Training ITHET '06, pages 388–393.

[14] Hannabuss, S. 2001. Contested texts: issues of plagiarism. Library Management, 22(6/7):311–318.

[15] Jenkins, T., and Helmore, S. 2006. Coursework for cash: the threat from on-line plagiarism. In Proceedings of the 7th Annual Conference of the Higher Education Academy Network for Information and Computer Sciences, pages 121–126, Dublin, Ireland, 29-31, August 2006.

[16] Joy, M., Cosma, G., Yau, J.Y-K., and Sinclair, J. 2009. Source Code Plagiarism – a Student Perspective. Submitted for publication.

[17] Kasprzak, J., and Nixon, M. 2004. Cheating in cyberspace: Maintaining quality in online education. Association for the Advancement of Computing In Education, 12(1):85–99, 2.

[18] Lambe, P. 2007. Organizing knowledge: Taxonomies, Knowledge and Organizational Effectiveness. Chandos Publishing.

[19] Martin, B. 1994. Plagiarism: a misplaced emphasis. Journal of Information Ethics, 3(2):36–47.

[20] Nadelson, S. 2007. Academic misconduct by university students: Faculty perceptions and responses. Plagiary, 2(2):1–10.

[21] Scanlon, P., and Neumann, D. 2002. Internet plagiarism among college students. Journal of College Student Development, 43(3):374–85.