Institution: The Open University

Unit of Assessment: B11 Computer Science and Informatics

Title of case study: Supporting Professional Development and Improving the Quality and Efficiency of Editorial Processes at Springer Nature

Period when the underpinning research was undertaken: 2011-2020

Details of staff conducting the underpinning research from the submitting unit:

Name(s):	Role(s) (e.g. job title):	Period(s) employed by submitting HEI:
Prof Enrico Motta	Professor of Knowledge Technologies	1984-present
Dr Francesco Osborne	Research Fellow	2013-present
Dr Angelo Salatino	Research Associate	2017-present

Period when the claimed impact occurred: 2016-2020

Is this case study continued from a case study submitted in 2014? N

1. Summary of the impact

Since 2016, Springer Nature, one of the world's foremost academic publishing companies, has used a software solution developed through **Motta**, **Osborne** and **Salatino's** research on scholarly knowledge mining to significantly improve the efficiency and quality of its editorial processes.

Using the researchers' Smart Topic Miner (STM) software to automate the process of annotating conference proceedings for its Computer Science book series, Springer Nature has reduced its associated editorial costs by 75%. STM has also enabled its editorial assistants' professional development and improved the quality of its processes, leading directly to an additional 12 million downloads of its approximately 3,200 STM-edited Computer Science books. This corresponds to a 53% positive growth differential with respect to the average growth trend for non-STM-edited Computer Science books.

2. Underpinning research

Led by **Motta**, The Open University's Scholarly Knowledge Modelling, Mining and Sense-Making research group (SKM^3) develops innovative techniques and tools to generate value from scholarly data. Its technologies provide academic publishing companies and other organizations with insights on the research dynamic in a particular scientific field or venue, geographical area, or organization. They also facilitate the effective curation of scholarly resources. Since 2011, the group has attracted funding for about GBP750,000, including direct sponsorships from the two leading international academic publishing companies, Elsevier and Springer Nature.

A significant output of SKM^3's research programme is the *Klink* algorithm **[O1, O2]**. Developed by **Motta** and **Osborne**, this ontology learning technique applies an innovative amalgamation of statistics with heuristic and knowledge-based reasoning, to analyse extensive collections of research publications and automatically generate a comprehensive and granular taxonomy of the research areas in the scientific field relevant to the given corpus, e.g., Computer Science. Klink is distinct in its approach to automatic taxonomy generation, as it considers the conditional probability of the candidate topics and their similarity in a vector space, as well as their diachronic relationships. It is also the first method in the scholarly domain to combine multiple sources of evidence, including both large-scale repositories of publication metadata from sources such as Scopus and Microsoft Academic Graph, as well as general-purpose knowledge sources available online, such as DBpedia. As demonstrated rigorously, this hybrid approach is key to optimizing the quality of the output **[O1, O2]**.

Motta, **Osborne** and **Salatino** have used the Klink algorithm to generate the *Computer Science Ontology* (CSO) **[O3]**. This large-scale ontology defines the most complete model of Computer Science research areas currently available, covering well over 14,000 topics in Computer Science. In particular, CSO's classification is an order of magnitude bigger than the most widely used alternative, the Association for Computing Machinery's (ACM) Computing Classification System. Moreover, in contrast with ACM's classification, which was produced manually over many years, the Klink-generated taxonomy requires no human intervention. Hence, it is possible to automatically update the taxonomy regularly, ensuring an accurate reflection of the most recent trends in research. Formal evaluations of CSO have shown a high degree of accuracy **[O1, O2]** and the availability of such a comprehensive and fine-grained characterization of the space of research areas has allowed the SKM^3 team to develop a number of innovative techniques, able to capture accurately detailed elements of the research dynamic in the Computer Science field **[O3]**.

To take full advantage of the granularity and accuracy provided by CSO in the context of automatic classification of research papers, **Motta**, **Osborne**, and **Salatino** developed the *CSO Classifier* **[O4]**. This new unsupervised approach takes the metadata associated with a research paper, such as its title, abstract, and keywords, as an input, and outputs the relevant research topics drawn from CSO. When evaluated against a gold-standard sample of manually annotated articles, the researchers found that the CSO classifier demonstrated a significant improvement in annotation accuracy in comparison with alternative methods **[O4]**.

The Open University researchers also employed CSO as the critical domain model underpinning the Smart Topic Miner (STM), a solution they specifically developed for the global academic publishing company, Springer Nature **[O5]**. This software uses the CSO Classifier (https://pypi.org/project/cso-classifier/) to annotate individual papers in Computer Science conference proceedings automatically. It then enhances an intelligent set-covering algorithm, which makes use of the topology provided by CSO, with a number of domain heuristics, to determine the best set of topics describing the overall scientific contribution of each volume. STM also provides a highly interactive interface, which allows users to investigate its rationale for proposing classifications. STM's output forms part of the publication metadata, which the publisher then uses for classifying proceedings in digital and physical libraries. Crucially, the goal of STM is not only to improve the efficiency of the metadata generation process but also metadata quality, to facilitate the discoverability of the Springer Nature's annotated conference proceeding volumes, both on their online portal, SpringerLink, as well as other digital libraries and third-party sites.

3. References to the research

O1. **Osborne, F**., and **Motta, E**. (2012) Mining Semantic Relations between Research Areas. In: Cudré-Mauroux P. et al. (eds) The Semantic Web – International Semantic Web Conference 2012, Boston, MA. Lecture Notes in Computer Science, vol 7649, pp. 410–426. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-35176-1_26

O2. **Osborne, F**., and **Motta, E**. (2015) Klink-2: Integrating Multiple Web Sources to Generate Semantic Topic Networks. In: Arenas M. et al. (eds) The Semantic Web - International Semantic Web Conference 2015, Bethlehem, Pennsylvania. Lecture Notes in Computer Science, vol 9366, pp. 408–424. Springer, Cham. https://doi.org/10.1007/978-3-319-25007-6_24

O3. **Salatino, A.A**., Thanapalasingam, T., Mannocci, A., Birukou, A., **Osborne, F**., and **Motta, E.** (2020) The Computer Science Ontology: A Comprehensive Automatically-Generated Taxonomy of Research Areas. Data Intelligence, 2(3), pp. 379-416. https://doi.org/10.1162/dint_a_00055

O4. **Salatino, A.A**., **Osborne, F**., Thanapalasingam, T., and **Motta, E**. (2019) The CSO Classifier: Ontology-Driven Detection of Research Topics in Scholarly Articles. In: Doucet A. et al. (eds), Digital Libraries for Open Knowledge. TPDL 2019. Lecture Notes in Computer Science, vol 11799. Springer, Cham. https://doi.org/10.1007/978-3-030-30760-8_26 Shortlisted for Best Paper Award.

O5. **Salatino, A.A**., **Osborne, F**., Birukou, A., and **Motta, E**. (2019) Improving Editorial Workflow and Metadata Quality at Springer Nature. In: Ghidini C. et al. (eds) The Semantic Web – International Semantic Web Conference 2019, Auckland, New Zealand. Lecture Notes in Computer Science, vol 11779. Springer, Cham. https://doi.org/10.1007/978-3-030-30796-7_31

4. Details of the impact

Generating scholarly metadata is a complex task in the academic editorial workflow. The proceedings of a particular conference may easily contain hundreds of contributions, where each of these can be relevant to several areas of research at varying levels of granularity. As a result, the number of topics covered by the proceedings can be very high, which means selecting a subset of topics to best describe the entire set of contributions, when compiled in book volume, is very challenging. Hence, this complex and time-consuming task has traditionally been tackled by highly experienced Senior Editors at Springer Nature. However, their limited numbers have resulted in bottlenecks, delays and a costly editorial process. In addition, there were also significant quality issues, as it was impossible for these editors both to have the required in-depth knowledge to classify every contribution accurately, as well as being able to manage the complexity of the task of selecting topics at proceedings level, without intelligent computational support.

In October 2014, Springer Nature asked **Motta**, **Osborne** and **Salatino** to help them find an efficient and robust solution to this problem, based on the researchers' work on Klink **[O1, O2]** and the Computer Science Ontology (CSO) **[O3]**. Their solution, the Smart Topic Miner (STM), has been in routine use at Springer Nature since 2016 and has i) resulted in an overall 75% cost reduction for this annotation task, ii) enabled the professional development of editorial assistants, and iii) dramatically improved the overall quality of the process, as evidenced by the additional 12M downloads recorded for the ~3200 Computer Science proceedings annotated by STM. Hence, STM has generated a significant impact on *Productivity* and *Enhanced Performance*, in terms of *efficiency*, *cost reduction*, and *quality*, and also on the *Practitioners* themselves, in terms of *professional development*. The beneficiaries include *Springer Nature*, who is benefitting from improved productivity, as well as the individual *members of the Editorial Assistants Team*, whose professional development has been supported by the deployment of STM.

Impact on Productivity: Improving efficiency and reducing costs

Computer Science Conference Proceedings constitute about 70% of all books Springer Nature publishes in Computer Science, which is their top-ranked field overall in terms of book downloads, and around 8% of its entire book output. While Springer Nature would not provide exact figures, we understand that the market value of this sector (Computer Science Proceedings) is in the order of millions of Euros. Using the researchers' Smart Topic Miner (STM) software to automate the process of annotating these conference proceedings, the academic publishing company has reduced its related costs by 75% **[C1]**. This dramatic cost reduction is the result of two benefits brought by STM. First, the software has halved the time required to annotate individual proceedings, compared to human annotation. Second, STM has made it possible for relatively junior Editorial Assistants, rather than the members of the Computer Science Editorial Team, to undertake the task. As the ratio of assistants to Senior Editors is five to one, STM has not only reduced costs but also effectively eliminated the delays and bottlenecks which arose from relying on a smaller senior team.

Impact on Productivity: Improving quality

STM has also significantly improved the quality of the metadata generation process for all Computer Science Proceedings at Springer Nature, including LNCS, LNAI, IFIP-AICT and others, for a total of about 800 volumes per year. As a result, both individual users and search engines can now more effectively identify the resources they need on the company's online portal, SpringerLink. In particular, there have been around 12 million more downloads in the past 4 years for STM-annotated Computer Science volumes, in comparison to those Computer Science volumes that do not make use of STM **[C1]**. More precisely, while downloads of Computer Science books at Springer Nature have been growing steadily across the board for a number of years, since 2016 the rate of growth for Computer Science Proceedings has accelerated dramatically, in comparison to other Computer Science volumes, a trend that can only be explained with the improved metadata quality brought about by the introduction of STM in the editorial process. Specifically, the growth differential between Computer Science Proceedings and other Computer Science books published by Springer Nature is 53%. The key reason here is that, in contrast with human editors, STM relies on a highly accurate classification of individual papers, provided by the

CSO Classifier **[O4]**. Building on this initial set of metadata, it then exploits CSO's topology **[O3]** and additional information, such as the popularity of different topics in the proceedings, to apply a heuristic set covering method that identifies those topics which best describe the set of conference proceedings compiled in book volume form. Crucially, STM guarantees that quality is scalable by avoiding historical classification errors resulting from different editors performing such a complex task manually. Writing in October 2020, the former Editorial Director, LNCS and Computer Science Proceedings at Springer Nature (currently, Vice-President Journals, Russia), highlighted the impact of STM on the company's productivity and quality in the following terms: "*In conclusion, I can state without hesitation that our collaboration with the team at The Open University has been extremely successful, allowing us to deploy a state-of-the-art AI solution within our editorial processes, which has addressed a key business need. I am very much looking forward to continuing our collaboration*".

Enhancing Editorial Assistants' professional development

STM has had a significant impact on the professional development of Springer Nature's team of 18 Editorial Assistants by empowering them to undertake a critical task that, until the introduction of this technology, was reserved to Senior Editors. This change has enabled these relatively junior professionals to become directly involved with the more scholarly aspects of the editorial production process, and more familiar with the large space of Computer Science topics. In the aforementioned letter, the former Editorial Director also describes how the introduction of STM has facilitated a "*mental shift*" in Editorial Assistants, allowing them to gain the confidence necessary to tackle more complex tasks **[C1]**.

In addition, in a July 2020 letter to **Motta**, an Editorial Assistant at Springer Nature points out that, before STM was adopted, "*the role of editorial assistants in the Computer Science Department primarily focused on performing a number of administrative and clerical tasks, such as managing the flow of manuscripts, acting as point of contact across different departments, and supporting editors and authors with their queries*" **[C2]**. Thanks to the introduction of STM in the editorial workflow, the letter goes on to explain, Editorial Assistants can now not only undertake more scholarly tasks but "*they have also gained a much better understanding both of the editorial processes regarding content classification and the computer science domain*" **[C2]**.

This learning also means that several new opportunities within Springer Nature are now open to Editorial Assistants, such as progressing into the role of Associate Editor or other positions that require experience in content classification. Springer Nature's senior leaders have welcomed these beneficial side-effects, in particular because the professional development of Editorial Assistants is a key plank of their internal Human Resources strategy **[C1, C3]**.

5. Sources to corroborate the impact

C1. Letter from former Editorial Director, LNCS and Computer Science Proceedings, Springer (now, Vice-President Journals, Russia), highlighting the impact on productivity and professional development resulting from the deployment of STM at Springer Nature. 27 October 2020.

C2. Letter from Editorial Assistant at Springer, highlighting the impact of STM on her professional development and that of her colleagues in the Editorial Assistant team at Springer. 20 July 2020.

C3. Letter from Director Human Resources Heidelberg, explaining the value of STM for the professional development of Editorial Assistants, and how this technology is supporting the company's HR strategy. 3 December 2020.