



NPL's User Data Analysis

Executive Summary

Why is the User Data important? It is without any doubt a known fact that an organisation relies upon quality data to develop strategies and make informed decisions. Analysis and Evaluation Team within NPL supports the business by providing data and analysis to assess and improve performance. This includes monitoring the growth of our impact to achieve our **Key Performance Targets to double our impact.**

One of the core datasets that lay the foundation for understanding the organisation and its network of complex activities is the USER DATA. We understand the organisation at a much aggregate level that is the inputs and outputs. To understand the internal workings, data prepared at a granular level is key. The User data utilised in its full glory is and will further be a way of shining light into the black box. It will allow the organisation to use a quantitative method to detangle the complexities and help make intervention and management easy. Previously, a qualitative method explored by Callum Dias; an economist within the A&E Team analysed the network of activities using the Case studies. It is indeed a very sophisticated way of understanding the nature and impact drivers of each sector. This process requires much manual effort in terms of the data preparation and due to the nature of the methodology opted, can only be effectively conducted every 5 years. User data analysis on the other hand operationalises the Case study analysis and can be repeated every year. The process of updating the user data, is time consuming, but the analysis itself can be easily replicated. Moreover, it will help broaden the understanding of the nature of each sector and allow NPL to set realistic targets specific to each sector.

In a nutshell, the User Data is a powerful tool which constructs the base and connects the analysis A&E team aims to conduct, to reach its desired targets. In time this will also pave way in to assessing the impact of trailblazing projects like Measurement for Recovery, Analysis for Innovators and the National Timing Centre.

The note gives a clear understanding of the components and the processes involved in the preparation of the user data. On average Advanced Manufacturing and Energy & Environment are two most active sectors across most modes of engagement. The activity level of Life Sciences & Health has been more or less constant over the years. Digital & Quantum Technology shows variation across the modes. It is a relatively younger sector, has received significant grants in the recent years and therefore expected to grow more across the modes in the coming years.

1 INTRODUCTION TO THE USER DATA

NPL offers a wide range of products and services, a combination of which are paid for services and others are free. There are six main sets of data that make up the User database, namely:

1. Invoicing
It gives information on NPL users that pay for a service. Could be measurement related, R&D, consultancy, or Training etc. It gives us the yearly invoices that come to NPL which includes specification of the project, and which NPL group and sector it relates to.
2. Collaborators
It consists of 5 types of collaborators. That is a list of organisations that collaborated with NPL in co authorising papers, working on grant funded projects, sharing staff and interlab standard comparisons. It also includes instances where NPL has provided consultancy services to other organisations. It is yearly and gives information on customers who collaborated with NPL via different routes and the group and sector the work related to.
3. Classroom Training
NPL offers a wide range of technical courses with an emphasis on practical hands on learning, an apprenticeship programme to develop skills across a wide range of sectors and bespoke training tailored to the organisation's needs. It consists of a yearly summary of the training and details of those that attended it.
4. E-Learning
NPL's eLearning platform offers convenient flexible training to help accelerate professional development. This provides a yearly summary of the learning course along with the details of the attendees.
5. Events
NPL hosts series of events all year round touching various areas of science. There are both free and paid for events. It gives a yearly summary of the events and the organisation/people who attended it etc.
6. Downloads
The downloads platform collected the user information for people making downloads on the website. This was discontinued sometime between 2017-18. So, there aren't any recent records for this type of data set. Hence, not shown here.

2 DATA DESCRIPTION

2.1 SOURCE

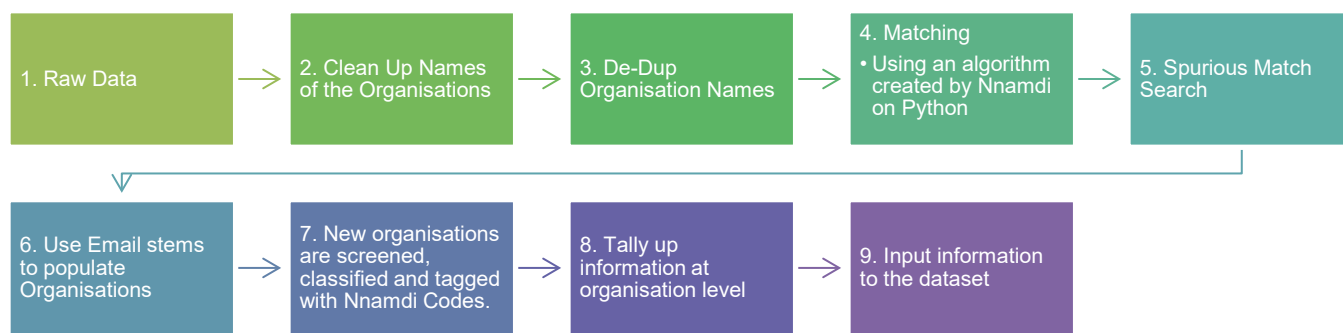
Each set of data is obtained from a different source within NPL. The table below summarises the information.

Type of Data	Sources	Timeline
Invoicing	Order Management System from Oracle	2001-Present
Collaborators	NMS indicators	2007-Present
Classroom Training	Salesforce	2012-Present
E-Learning	Intelli-board Knowledge Management System	2014-Present
Events	Eventbrite	2014-Present
Downloads	Library system	Discontinued

2.2 DATA CONSTRUCTION

Constructing the data in a form that is readable and suitable to conduct analysis on, is a big challenge. Nnamdi Nwaigbo; a former economist in the Analysis and Evaluation Team constructed a method to match the company names. The trouble was that company names differed across the various sources and it was impossible to track them. So, he invented codes for each organisation known as the Nnamdi Codes. Each year the new data sets are matched against a reference file which contains unique codes for the organisations and a company house id where applicable. It is kept up to date and contains a list of all the organisations which have interacted with NPL starting from 2001 (also linked with the Fame database). It also contains a tag for the organisations' type (Public or Private) and nature of business (Company, Hospital, Government Body, PSRE etc.). For instance, there are approximately 10,000 organisations of which 8,000 are private sector firms, 300 are Government Bodies and Councils, 500 are Academic and Research Institutes, 200 are Hospitals and the rest correspond to some additional categories.

The diagram below shows the process of data construction.



This does not include the troubleshooting that occurs between step 4 and 5 which involves manually checking if the match made by python algorithm is correct or not. This is the most time-consuming step.

Each year the new data once received and matched, is added to the existing database. This process takes good 3 to 4 months because the compilation is dependent on receiving data from various sources (one of which is the collaborators data from the NMS indicators). To understand the sectorial activity, the entries within each data set were tagged a relevant sector. This process required both manual and automated effort. It is repeated across all 5 datasets. A key has also been made which will make sector tagging process simple for the next year's data.

2.3 CAVEATS

With all research and data analysis, there are always some caveats associated, that should be brought into attention. Despite being meticulous in the process of ensuring data quality, there will always be some level of uncertainty attached.

The data is compiled site wise where it can be. Doing this is easy with OMS data as it has the address and can easily be linked up with a unique company house number. But with other data sets, it is often a judgment call. Sometimes looking at the email stem helps. But where it is not so clear, the code of the parent company is taken. Organisations which cannot be found using a website address and are not registered with company house are marked as unidentifiable.

During this process, there is a chance of duplicating an organisation by giving it two unique

Nnamdi Codes. One of the reasons; not being able to search the company house id in the reference file as that one contains a space (unintentional error). In addition, it is not evident from the name if an organisation is public or private and requires manually searching each organisation and making a judgment where necessary.

To minimise the errors involved, a quality assurance exercise was conducted in which both automated formulae and manual processes were used. It was found that approximately 60 codes had a duplicate which implies that 0.6% of the data had to be corrected for. In addition to the duplication of organisations, a few had an incorrect tagging for the characteristics (type and nature) of the organisation (rough approximation 5%). Codes used for some organisations were not consistent. For example, a hospital's code should start with an 'H', but some started with a 'C' (for company). Approximately 400 codes were inconsistent. To summarise, inconsistent codes do not cause any disruption to the analysis conducted but duplicate codes and incorrect tagging can have an impact on the analysis. If only private firms are being looked at and some firms have been incorrectly tagged as public then those will be left out from the analysis. Nevertheless, the quality check exercise mentioned previously, at the very least resolves 98% of the issues. A&E will continue to conduct these quality exercises for the robustness of the data.

Furthermore, a mismatch between the raw and processed collaborators data has been identified for 2017. As pointed out by Nnamdi, "The data curation from 2007 - 2017 differs from the method used in 2018". The issue is that Oracle reports only capture grant funded projects in which we directly get revenue. It turns out there are a whole group of other projects that we are a partner in that don't involve a direct payment. When we got access to Dimensions, I had an easy method of obtaining these other projects. In addition, what I found when I started looking into things more closely was that there were extra projects from the IUK dataset that didn't appear in our Oracle records. Same for EMPIR projects (Euramet). Hence, I started using those as separate sources as well.

A&E will aim to correct the mismatch before the 2022 round of updating datasets (for 2021).

3 ANALYSIS

For this part we look at each sector tagged dataset one by one from 2017-2020. NPL has four sectors Advanced Manufacturing, Digital & Quantum Technology, Energy & Environment and Life Sciences & Health. One thing which will be common across all the data sets, is the way we deal with sector tagging for entries which fall in more than one sector, is classified as generic or as unknown. There were three approaches

1. Keep them all separate and combine Generic with Unknown
2. Keep the four sectors separate but put the multiple sectors in to the Generic/Unknown category
3. Divide the multi sector values in to two and add them in to their respective sector. You end up with five categories 4 sectors and Generic/Unknown.

The third approach seems to be the most reasonable approach and gives you neat picture of the activity across the sectors. Nevertheless, the excel file contains the results for all three approaches.

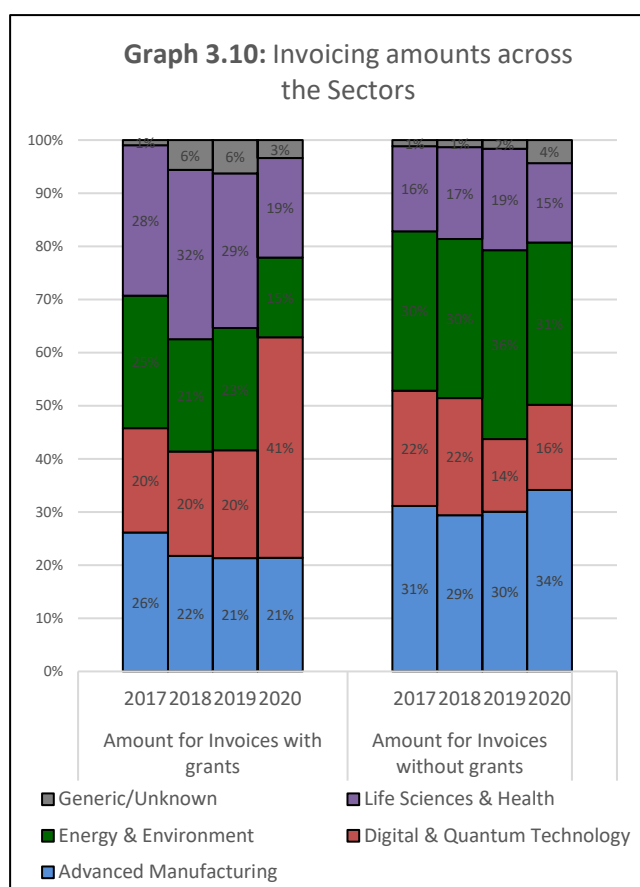
3.1 INVOICING DATA

It required a series of steps before it could be analysed. A 'broad Revenue type' column was added to the data which is an indication of the type of paid service. This gives 6 focused categories namely **"R&D/instruments"**, **"Grants"**, **"Measurement Services"**, **"Knowledge Transfer"**, **"None-NMS"** & **"None-other"**. The entries which related to NMS funding were tagged as None-NMS and recharges or capital grants were classified as a 'None-Other'. Because these two categories do not connect with actual services that NPL provides, they were taken out of the data.

A couple of issues were encountered for instance there was some tweaking in the original data specifically for BEIS entries for the NTC programme and so those had to be updated. Some entries which were classified to Central Business group had a tag of time and frequency. They were then tagged to Digital & Quantum Technology sector. Another issue arose when amount in GBP for DQT spiked in 2020 which seemed out of the ordinary. After quite a thorough investigation and interaction with the finance team, it was discovered that Invoice line amount is different to General Ledger Accounted amount. A transaction ID may be split in to more than one invoices, and the GL amount will correspond to each invoice for that particular transaction. Invoice line amount will be the total for that transaction. The example below explains this.

Transaction ID	Invoice Line Amount	GL Accounted Amount GBP
133241	24333	24323
133241	24333	10

The graphs below summarise the level of activity across the five sectors. It is further broken down by the revenue type. Graph 3.10 shows two scenarios one that takes in to account the grants and the other without the grants. If we look at the first scenario, from 2017-19 most of the income seems to be coming from LSH then a somewhat equal distribution among the other three sectors. DQT experiences a spike in 2020 which cannot be seen in the scenario 2. This implies that DQT received a significant amount of grants in 2020. This was then reconfirmed with the Finance Department. The funding to DQT has nearly tripled to what was received in the previous years because UK government has set priorities to digitalisation. The table below further shares light on the amount of grants received within DQT.

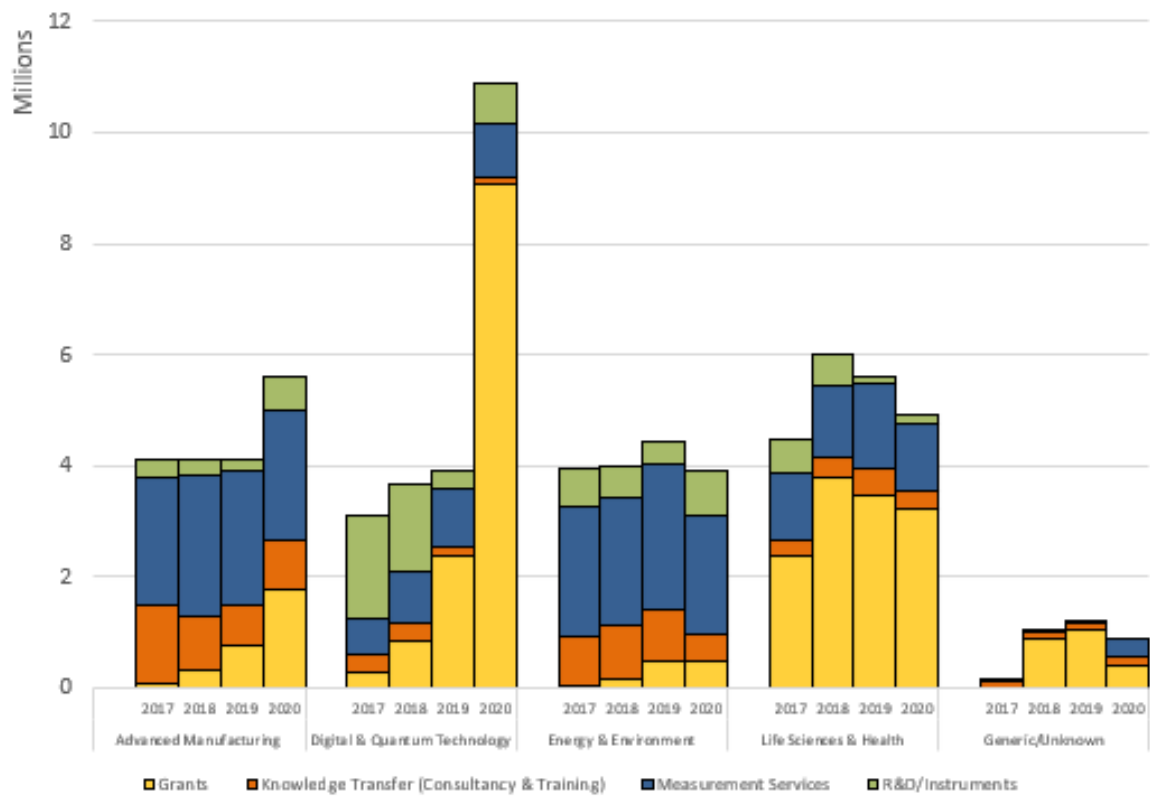


	Amount from grants in £m across the five years				
Types of Grants	2017	2018	2019	2020	Total
EMPIR	.00	.00	.00		.00
H2020 (not Marie Curie)		.00	.00		.00
ISCF			.47	4.42	4.89
I-UK	.26	.82	1.13	.56	2.95
M4R				.05	.05
NTC			.33	3.11	3.45
SPF			.40	.93	1.34
Grand Total	.26	.82	2.34	9.08	12.68

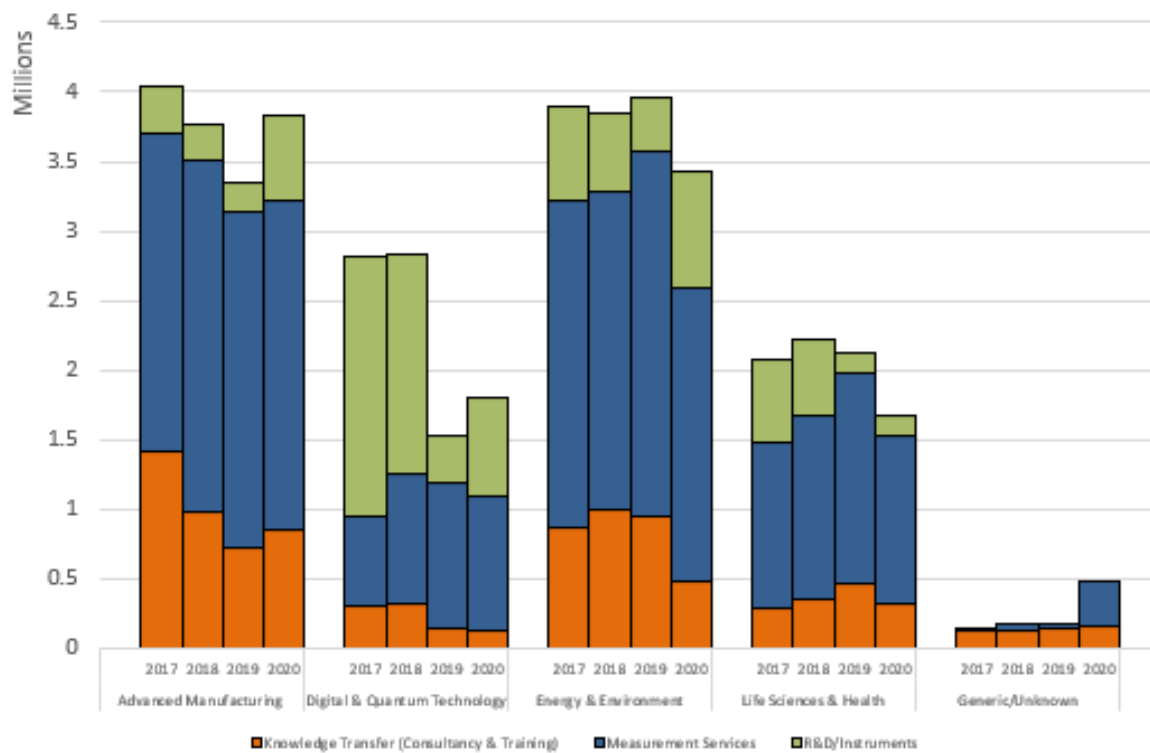
Without the grants, a more realistic picture of the income that comes from products and services can be seen. More income comes from AM and E&E than from DQT and LSH. Even during the pandemic, the flow of services from AM continued to increase. The other sectors experienced a decrease in activity of 2% to 5% (from 2019-2020).

Graphs 3.11 and 3.12 are a more detailed version of Graph 3.10. Graph 3.12 excludes grants and only focuses on services that NPL provides. AM, EE and LSH are concentrated to providing measurement services. DQT is heavy on R&D and instrumentation, at least for 2017 and 2018. Thereafter it experiences a decline which could be a result of reduced Instrumentation contracts by NPL.

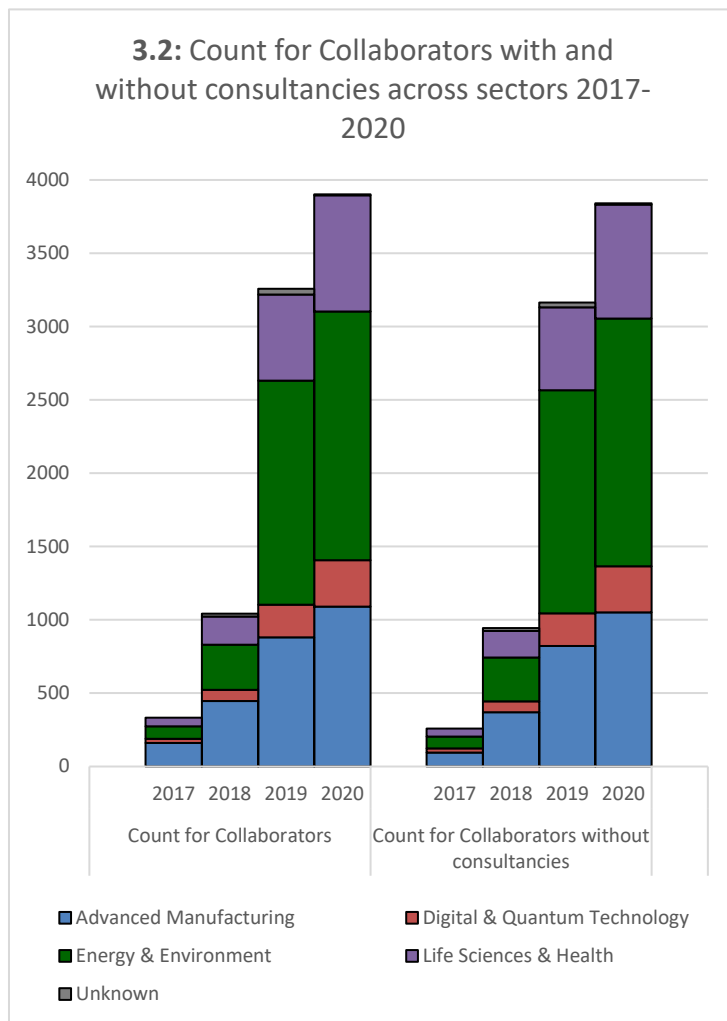
3.11 Invoicing amount across sectors



3.12 Invoicing amount across the sectors without grants



3.2 COLLABORATORS DATA



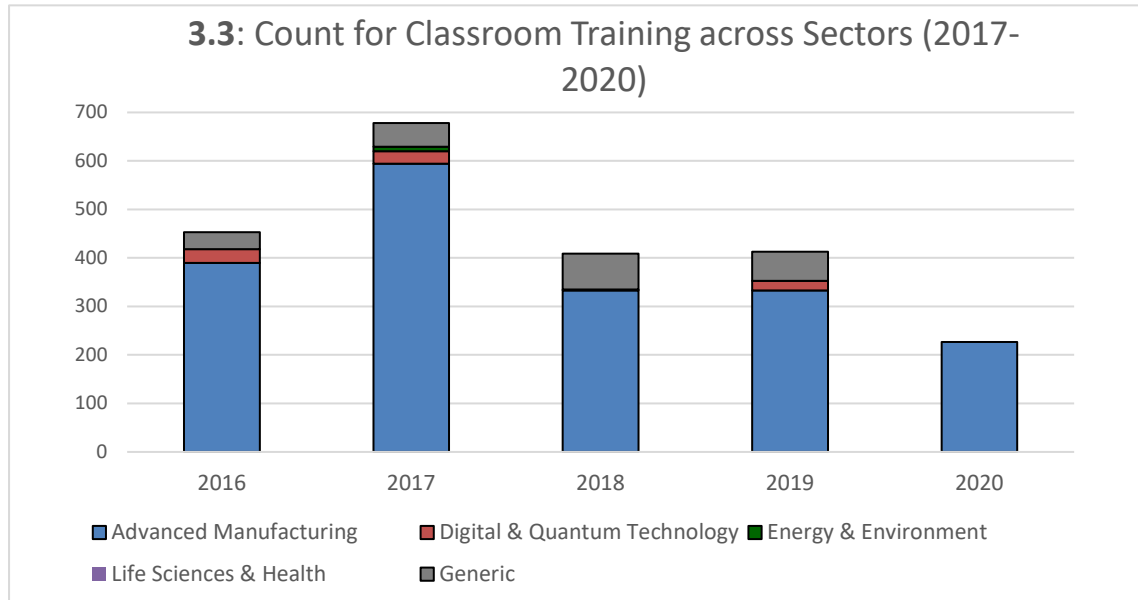
As touched upon before, the data for collaborative activities is comprised of various activities like papers co-authored, grant funded projects, consultancies, shared staff and interlay. Including all the activities within one spreadsheet has led to a much-complicated dataset.

Several complexities have risen as a result. A&E team will be working to build a dataset that is much easy to handle in the upcoming year. To comply with a tight and appropriate definition of collaboration, A4I and M4R projects which were previously marked as consultancies were taken out and put in grant funded projects. To monitor the activity within each sector, two scenarios are taken, similar to invoicing data. One where we include the consultancy and the other where we take it out.

The trend of activity is not different in the two scenarios. A 4% difference can be seen between the two in totals. Collaborative activities surged from 2018. Clearly, the A4I and especially M4R project have kept the collaboration up and running during COVID-19. The jump is mostly visible for EE and LSH. This can be explained through UK's top priority on Net Zero by 2050 and levelling up the nation's health.

3.3 CLASSROOM TRAINING DATA

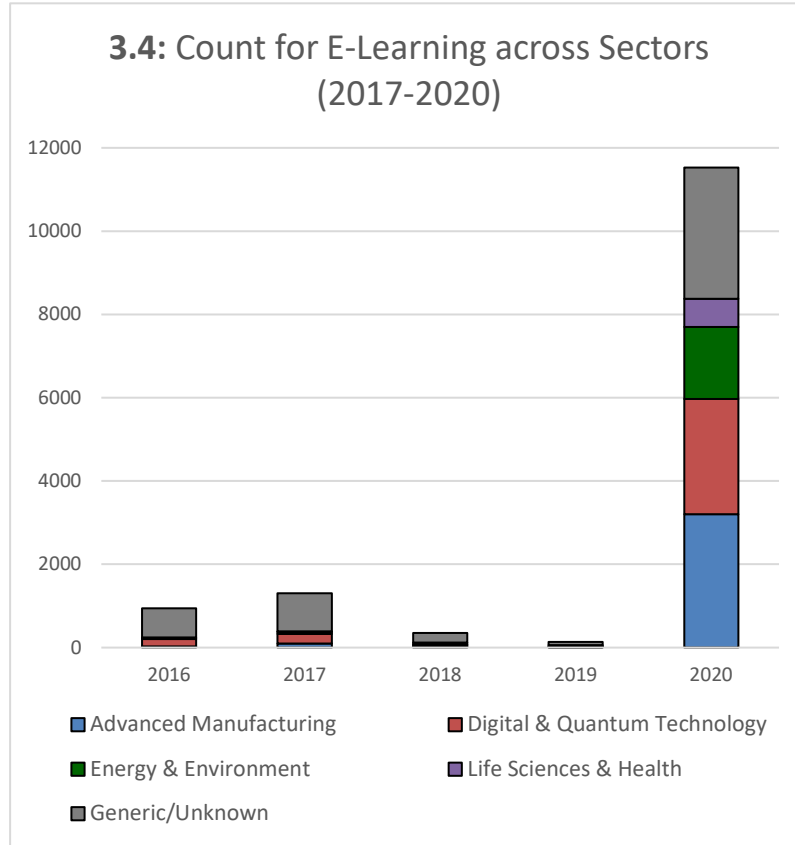
It is very evident from the data for face to face training that it is completely directed towards Advanced Manufacturing. During the pandemic this form of training has declined substantially which is not surprising due to the inevitable restrictions imposed.



3.4

E-LEARNING DATA

There has been a huge uptake in e-learning courses in 2020 compared to previous years.

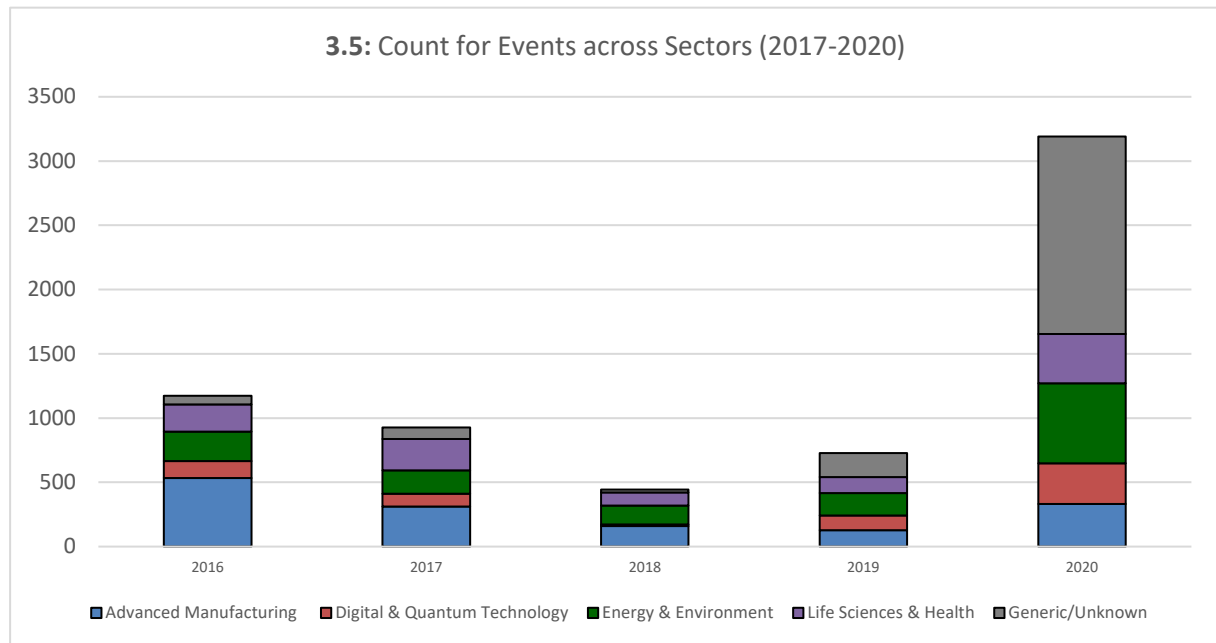


This supports the fact that during the pandemic, people opted for e-learning courses and possibly had more time and flexibility to do a particular course. In addition, NPL made quite a few paid e-learning courses free of cost which again supports this. The e-learning courses are approximately equally divided across AM, DQT and the Generic category. The Generic category includes a combination of courses which are too general to be tagged to a specific sector and unknown courses, which is a small number.

3.5 EVENTS DATA

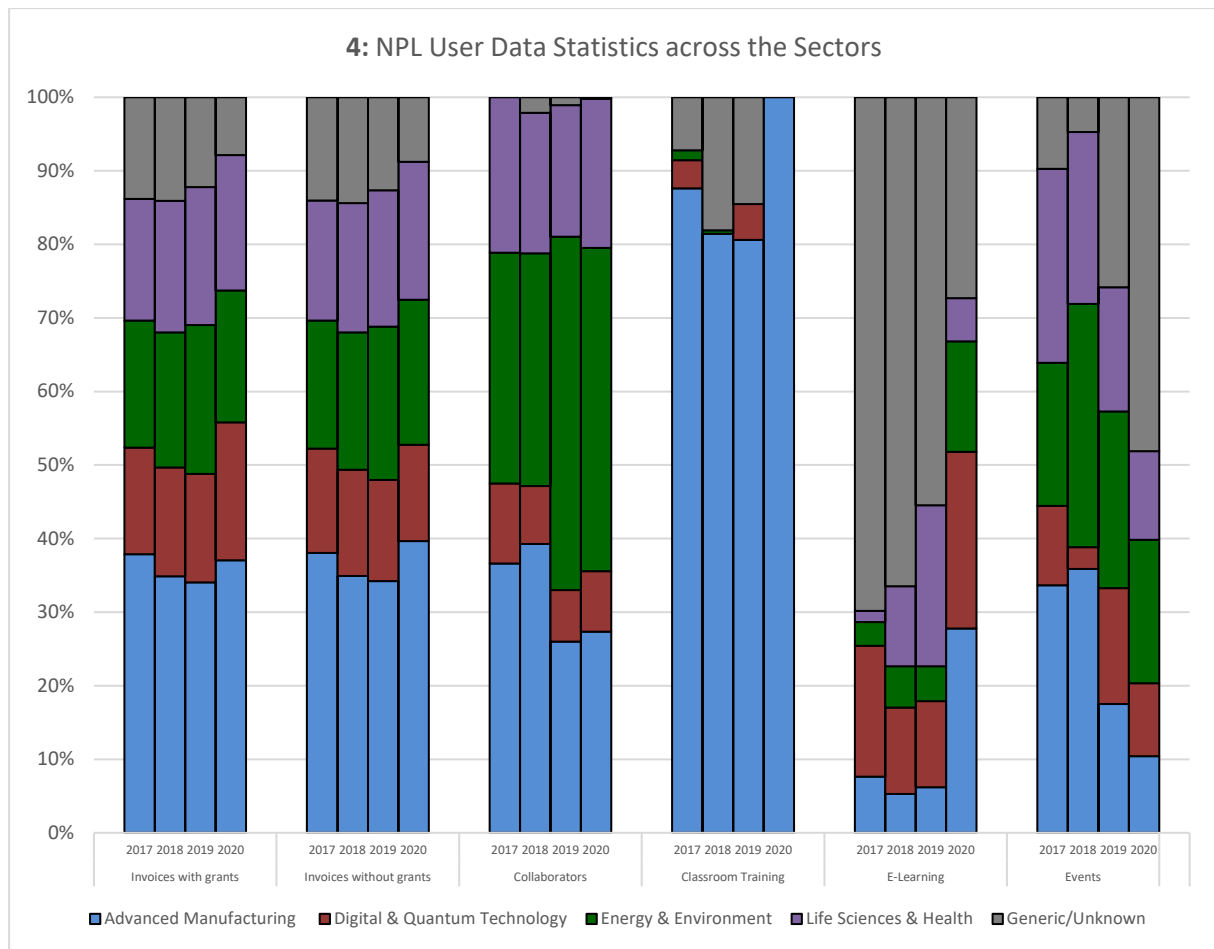
The stats for events attended look similar to e-learning courses attended in 2020. There is increased activity among all the categories in 2020 as compared to the previous years. Generic category being at the top for this.

Attending events virtually eliminates the geographical proximity factor, which explains the increased participation.



4 CONCLUSION

To summarise the findings, the graph below shows the activity within each sector across the various modes of interaction by looking at the number of invoices received over the period of 4 years. AM connects to the majority of invoices received by NPL. Moreover, if we look at the amount of income (deflated), AM and EE are the most active (Graph 3.10). Collaborative instances of interaction (excluding consultancies) are more heavily distributed between AM and EE. Classroom Trainings (face to face trainings) in its entirety belongs to AM. E-learning courses are more spread out across the sectors. The courses seem to have moved from being very generic to much sector focused topics. The events are also variably distributed among the sectors.



The note attempts to give insight on the engagement activity across NPL associated with the five sectors. This can be converted into a time series dashboard where previous and more recent years can be added on. The methodology applied here is simple and can be replicated for additional years. In addition, the dashboard can be made more insightful, by adding the type of organisation and type of business (Private or Public). A more sophisticated way would be to create reports on a Power BI Dashboard which would automatically update each year, once new data is added in.

Zahrah Qureshi | zahrah.qureshi@npl.co.uk

National Physical Laboratory
Hampton Road, Teddington, Middlesex, United Kingdom, TW11 0LW

This work was funded by the UK Government's Department for Business, Energy and Industrial Strategy (BEIS) through the UK's National Measurement System programmes.

Extracts from this report may be reproduced provided the source is acknowledged and the extract is not taken out of context.