Item 10c
ROOM DOCUMENT
38th Meeting of the Global Science Forum

SCOPING PAPER - DATA COMPETENCES AND SKILLS FOR OPEN SCIENCE

Secretariat Note

This scoping paper has been drafted by Stephen Brewer, with some editing and re-structuring by the Secretariat. It is not intended to be published but rather is an internal document to help GSF delegates consider some of the key issues relating to Data Skills for Open Science.

Following discussion of a preliminary version of this paper at GSF37, it was agreed that more consideration should be given to the diversity of different scientific fields and their potential needs in relation to data skills and competencies. Building on previous fieldwork for the EDISON project to develop a Data Science framework (Appendix A) a number of case studies from different domains have been included in the paper.

The paper suggests that there are opportunities for policy learning from a more systematic comparison of practices across different communities and countries. If GSF is to follow up on this then it will need the support of a number of countries, including the identification of suitable case studies and assistance in describing and analysing these.

1. Executive summary

1. The scientific process continues to evolve, and one of the biggest transformations is the dramatic increase in the use of data across all domains. As a result, many scientists are required to acquire (or access) a variety of new specialist skills to capture, handle and exploit this data. Different scientific domains are at different stages of adapting to this requirement. The needs in historically data-focused fields, such as particle physics and astronomy are different and arguably less acute than those in areas such as medical research or social sciences. With an increasing emphasis on inter- and trans-disciplinary research to address complex societal challenges, the skills to manage and analyse data from multiple scientific domains are becoming increasingly important.

2. A number of studies suggest that there is a large and growing shortage of suitably skilled data scientists worldwide and there is intense competition to academia from the commercial sector for the recruitment of such individuals. Within this broader socio-economic context, it is becoming apparent that all scientific domains will require data management and analysis skills and competences to be taught, or at least acquired, as part of traditional teaching pathways. Given the speed of development, lifelong learning is also likely to be necessary as careers progress. Some skills will be common across all disciplines, and some distinct. There are needs for data-skilling of scientists and skilled data scientists. There are particular needs in terms of data curation and long-term stewardship to ensure the quality, availability and usefulness of research data. In many fields research infrastructures and associated data services will have an important role to play, with important implications for their human resource policies.

3. This scoping study is specifically focused on public sector research . The broader context for this is the move towards Open Science and the rapid development of information and communication technologies that are affecting the conduct of research in all fields of science. 'Data competences and skills for open science' is adopted a a broad term that encompasses topics such as Big Data, Artificial Intelligence and other facets of Data Science. The scoping includes an analysis of the generic issues related to digital skills and what we mean by data science and data scientists and then briefly explores how these translate in different scientific domains.

4. The overall conclusion from this scoping is that this is a complex area with limited systematic analysis. There is undoubtedly an increased need for data skills in society as a whole and some of the necessary skill can be categorised and built into traditional teaching pathways. At the same time things are evolving very rapidly and flexibility in education curricula and training are essential. Within science, different domains are at different stages of preparedness for dealing with the challenges and exploiting the opportunities presented by big data, AI and Open Science. Many activities are taking place in different areas and different countries to address some of the perceived skills gaps. It is not clear whether these disparate, and largely uncoordinated, activities are sufficient and whether more strategic approaches may be necessary. However, without being prescriptive, there appear to be significant opportunities for policy learning from a more systematic comparison of practices across different communities.

Table of contents

2. Introduction

2.1. Objective

5. The scientific process continues to evolve, and one of the biggest transformations is the dramatic increase in the use of data across all domains. As a result, scientists are required to acquire a multitude of new specialist skills to capture, handle and exploit this data. Under the umbrella of data science skills, scientists are increasingly required to become or at least collaborate closely with data science professionals. However, exactly what specialist skills are required, and in what combination, and quantity is not consistently appreciated.

6. Whilst market forces play their role in motivating individuals and organisations to equip themselves with the skills to address these needs, there remains a science policy dimension to this situation. Universities and training organisations will seek to recruit appropriately and re-train where applicable but this, in itself, does not ensure a balance between supply and demand. The aim of this scoping paper is to identify and quantify the nature of this imbalance, and to identify good practices to address the unmet capacity needs for scientific research across OECD countries.

7. This study is specifically focused on public sector research. In terms of subject matter, data competences and skills for open science is taken as a broad term encompassing topics such as Big Data, Artificial Intelligence and other facets of Data Science.

2.2. Rationale

8. The Internet has changed many aspects of our lives over its relatively short lifetime. Indeed, it is timely to recall that the World Wide Web was created to aid the sharing of research publications amongst High Energy Physicists at CERN. Furthermore, it is this very research community that has led the way in conducting science that is dependent on harvesting and processing huge quantities of data. Other domains are now following suite, but processing different varieties of data, both structured and unstructured, disposable (due to repeatability) and irreplaceable (e.g. unique recordings of lost languages).

9. There is reportedly a large and growing shortage of suitably skilled data scientists worldwide, although this situation appears to be more severe in fields that have historically been less data intensive. At the same time, there is no simple accepted definition for the term data scientist so identifying candidates for posts is not straightforward. It is clear that data science skills will play a significant role in the future scientific workforce, so we need to know more about what skills might be needed and how a suitably skilled workforce will be created. There are both challenges and opportunities inherent in ensuring that scientists are suitably equipped with the skills and competences necessary for a digitally enabled and data-centric approach to research:

challenges in the sense that education and training for these roles will have to be delivered, and opportunities in the sense that acquiring these skills will enable scientists to deploy innovative new pathways to scientific discovery.

3. Background

10. Whilst the tsunami of data emerging from scientific research enables traditional science to be conducted better and faster, it also heralds the advent of data science as a recognised, and much sought after, field of expertise. The development and integration of appropriate data skills will in many ways determine the future of scientific research.

11. Before we examine the different, typically open, approaches to science adopted across domains in capturing and exploiting data, it is worth surveying the back drop of changing research technologies and the wider evolution of computing. All scientific domains are affected, from high energy physics and astronomy to the digital humanities, and personalised medicine. Therefore, all domains will require skills and competences to be taught or at least acquired as part of traditional teaching pathways as well as lifelong learning as careers progress. Some skills will be common across all disciplines, and some distinct. Who will lead these new and additional learning streams? Who will certify and accredit the courses and exams? What career pathways will open up as data scientists migrate between disciplines and organisations? Do we need a consolidation of research progression to align the evolution of data-centric science or should disciplines define their own interpretation of data science skills and competences? The knowledge base on data science continues to build but will it coalesce or fracture?

12. In addition to mapping the range of skills and competences needed by the science practitioners of the future, there is a further need to consider role of Human Resources (HR) management in relation to data professionals, teams, and portfolio careers. Individuals may switch roles, may adopt multiple roles, and will certainly need to be more adept at collaborative practices. HRM may need think about career paths beyond recruitment and departure, and annual reviews. Mobility between the public and private sector, for those with highly desirable skills, eg in Artificial Intelligence, may need to be considered strategically rather than simply left to market forces.

3.1. Data Driven Science

13. Big Data has had a significant impact on how we approach big challenges, not just in the world of business, but in science too. The innovative digital practices that became characterised as eScience have coalesced into a broad family of activities labelled under the umbrella of data science. The practices of eScience were notably encapsulated in the Microsoft Research eScience Group led by Jim Gray. Visionary thinking about data science activity was first published in a collection called The Fourth Paradigm: Data-Intensive Scientific Discovery[1]. Further insight into these new practices is provided in The Data Harvest: How sharing research data can yield knowledge, jobs and growth, published by the Research Data Alliance (RDA) Europe in 2014[2]. The Fourth Paradigm refers to an evolution of the scientific process. First, we had theory, hypothesis and logical reasoning. Second, we had observation and experiment as exemplified by Newton and Galileo. Thirdly, we had simulation of theory and modelling whereby digital simulations would be used to prove theories or models. Now we have data-driven

scientific discovery also known as data science. This is also sometimes referred to as e-Science, and is characterised by computing systems and information and communication technologies being combined with data to advance scientific discovery.

14. The advent of data-driven science also enables further innovation in areas such as artificial intelligence. With the adoption of machine learning scientists will be doing more than using data to support discovery, deep learning algorithms may be significantly directing the course of investigation. Already, Text and Data Mining (TDM) is playing a part in exploring unstructured information for knowledge discovery, seeking out new directions and opportunities.

3.2. Open Science

15. Open science is predicated on the idea that knowledge created from public research should be made available for the public good. The OECD has compared this to the 18th century concept of "commons" which is often engaged to reinforce the importance and historical precedent of societal sharing. However, in the digital age open access to research data can provide many opportunities to enrich the original shared knowledge base.

16. There is no single or simple definition of the term open science. It is an umbrella term capturing the global movement over recent years to make the outputs of publically funded research publically available. More comprehensively, OECD has described the term as: efforts by researchers, governments, research funding agencies or the scientific community itself to make the primary outputs of publicly funded research results – publications and the research data – publicly accessible in digital format with no or minimal restriction as a means for accelerating research; these efforts are in the interest of enhancing transparency and collaboration, and fostering innovation (OECD, 2015). The motivation towards this openness is more than simple altruism on the part of researchers, there is a broader desire for research to address societal challenges and, in doing so, harness the contribution that citizens can contribute to the process. Open science is often presented as a family of related "open" approaches including open access to academic publications (OA), open source software applications, and open data (Jomier, 2017).

17. Open access (OA) refers to the process of making research findings accessible, this can include everything from laboratory notebooks to final publications. Significant progress has been achieved since the Budapest Open Access Initiative declaration in 2002 which defined OA as follows:

> *"By 'open access' to this literature, we mean its free availability on the public internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself."* (Frankland & Ray, 2017).

18. The benefits of OA are seen as falling under three clear categories: promoting discovery, documenting social benefits (for others), and improving education (through the availability of the latest research findings). However, there is no single emergent viable economic model for OA; the relative merits of competing models have been compared and a single inclusive optimising solution that balances the predators and saviours, the suppliers and consumers, for the scholarly publishing market has not been found despite the obvious benefits of the Internet and other newer technologies (Frankland & Ray,

2017). In order to promote open access to publications a number of policy initiatives, mandates and incentives have been implemented across the world. This has been accompanied by public investment in enabling infrastructures, such as PubMed Central.

19. Open source software (OSS) is in many ways the forerunner of the open movement with a strong, in some ways politically motivated, community mobilised to create and curate reusable software. This approach grew steadily for many years in parallel with the separate world of commercial software application development and distribution. That said, viable economic models have been developed around OSS whereby people and businesses are able to develop and grow, developing the economic benefits of providing services around the implementation of otherwise free software.

20. Open data presents its own opportunities and challenges, with technology enabling greater storage and accessibility, but with the openness come challenges in relation to rights and responsibilities. Many of the complex issues relating to open data, are been addressed in efforts to implement the FAIR principles, where FAIR stands for Findable, Accessible, Interoperable and Reusable, in different scientific domains.

21. Whilst the main policy focus in most OECD countries has been on OA for publications there are a growing number of initiatives for sharing research data. Here too, there are challenges in terms of sustainable business models, and a lack of agreement on where responsibilities lie for the support, curation and management of research data. Librarians have embraced the opportunities and a number of the roles, but by no means wish to take on all roles in the research data lifecycle. Other specialist roles are emerging, but these actors have not yet been fully accepted in the research sector. There is a clear cost/value balance to be achieved across the roles, at its heart is the question of how to value academic support work, and how to incentivise academics to make their data available outside their own domain (Funamori, 2017).

3.3. Research infrastructure and data

22. Research Infrastructures(RIs) are one of the main sources of research data, the raw material for scientific endeavour. They can variously be data generators, data aggregators or data distributors. They can be single site RIs (either nationally owned with an internationally-based user group) or distributed RIs composed of geographically distributed facilities open to international user groups. RIs typically serve different scientific domains or sub-domains and vary enormously in terms of data types, data processing and access. However in many areas of science RIs are well positioned to play an important and leading role when it comes to consideration of data competencies and skills.

23. The Square Kilometre Array (SKA) for example is an international effort to build the world's largest telescope. The SKA exemplifies the concept of a data-generating infrastructure, with terabytes of data flowing out into sites collecting petabytes of data enabling multiple users to explore a range of science challenges from gravitational waves to dark energy.

24. Some of the main differentiators between RIs are in the heterogeneity of the data sources, and the approaches to data stewardship. The High Energy Physics community significantly focuses its efforts on data flowing from CERN; essentially four large underground detectors deliver the vast majority of the data for the Large Hadron Collider (LHC)'s collaborations of distributed physicists around the world. CERN plays a critical role in training this community and more broadly in promoting open science

25. Domains such as geo-informatics view the Earth as an integrated system through the synergy between science and informatics. Diverse social and technological approaches contribute to the biological and geological data captured by scientists around the world resulting in data recorded in notebooks, laptops and other locations. Conversely newer sources such as underwater sensors and satellites deliver large quantities of homogenous data, resulting in a complex bimodal data environment. Integrating these clearly different spheres of knowledge into a coherent whole requires advanced informatics techniques scientists with specific data analysis skills (Sinha, Thessen, & Barnes, 2013).

26. Dispersed and co-located teams are both used to conduct scientific enquiries. These present specific challenges in when it comes to handling data. Data, especially in quantity, must be managed in a rigorously consistent manner. Leadership and control, and communication procedures are needed to ensure that data management is kept in step with scientific processes.

3.4. Services and value chains and co-creation

27. Many labs now offer services to other labs. Science as a service can be seen as a creative business approach to market opportunities, or also as a manifestation of the Open Science phenomenon. Online services are provided in many fields through science gateways, data portals and platforms. Examples include of crystallography services, with remote access being provided to national experimental facilities at the UK science centre at Harwell near Oxford.

28. Open Labware[3] and Thingiverse[4] are two examples of an open source approach to lab equipment where 3-D designs are shared to enable the easy and legitimate re-creating of proven 3-D lab equipment. In these cases, the Open Source movement, which is a global phenomenon, effectivley brings people together in order to support science. The OS model can also work for commercial organisations as this enables improved customer relations and product feedback. There are also potential benefits for the developing world where open source approaches have been embraced, although capacity issues are perhaps still limiting their adoption.

29. There are many potential benefits of OS yet to be fully realised, these include more collaboration and greater democratization at both a national and international level. One model for looking at this potential is an open innovation triangle comprising: open science, co-creation of knowledge and open innovation (Ramírez-Montoya & García-Peñalvo, 2018). This triangle of open innovation which has been seen to deliver demonstrable success in a number of countries including Brasil, the USA, and the UK, can act as a robust model for others. New technologies and the Internet have been given much credit for enabling this transformation.

3.5. Contemplating the future

30. Data science and Artificial intelligence will have a transformative effect on all aspects of the world from finance to healthcare, as well as scientific research. The digitisation of business and industry, coupled with entertainment, leisure and education means that all of us are exposed to and dependent on digitalised and data-centric processes and services. The ramifications with regard to security, trust and technological competence are becoming clear to us all but are nowhere more evident than in research itself.

31. These transformations, which will in turn both enable and require learning and skills, are affecting not just services and processes but also jobs and lives. This means that education and training must change in many dimensions. There is an increasing complexity to this because the automation of digital services that we are all depending upon abstracts more and more activities away from us, meaning that it is increasingly harder to see beyond the layers and platforms. This matters for science. Whilst we may acquire a generation that is digitally and data science literate, they may not possess traditional scientific values, and they may not perceive how much ICT utility services are assisting them in their work. Their neural networks will be hard-wired differently.

4. Data Science and skills

4.1. Defining Data Science: a new and complex skills matrix

32. We start by considering the terminology. A good definition of data science is available from the NIST[1] Big Data Working Group: NIST Big Data Interoperability Framework: Volume 1, Definitions:

"Data science is the extraction of actionable knowledge directly from data through a process of discovery, or hypothesis formulation and hypothesis testing."[5]

33. The same report also has a useful definition of the term Data Scientist:

"A data scientist is a practitioner who has sufficient knowledge in the overlapping regimes of business needs, domain knowledge, analytical skills, and software and systems engineering to manage the end-to-end data processes in the data life cycle."

34. Data science is an overarching term that brings together an evolving collection of technologies and approaches including Big Data and Artificial Intelligence. It can be considered as a cross-disciplinary function as well as a subject in its own right.

35. Whilst there is moderate consensus on the terms data science and data scientist, there is no clear definition of the skills and competences inherent within these terms. The terminology and roles have grown steadily in number over recent years but this has not been paralleled by a harmonization of quantifiable characteristics. The EDISON Data Science Profile (appendix 1), when added to existing ICT frameworks, offers hope for a degree of formal systematic structuring, but this is not assured. The structuring of data science skills and competences matters as students will need to acquire confidence and capability, not simply awareness of various techniques.

36. The specific challenge for policy makers is to identify the nature and timeliness of the skills needed, and in particular the peaks and troughs over time. Whilst, both the commercial and the public-sector research community will have specific needs, the public sector will need to be flexible and adaptable in recruitment retention and promotion in order to manage less flexible budgets and more rigid employment regulations. The specific needs for different sub-skills will rise and fall in importance over time. Instilling a desire and aptitude for lifelong learning if practitioners want to remain in demand.

[1] National Institute of Standards and Technology (NIST)

4.1.1. Outcomes of the EDISON Study in relation to skills needs

37. The author of this report was a coordinator of the EC funded EDISON project to develop a Data Science Framework. This included fieldwork and interviews with scientists and educators. The main outcomes of this information collection exercise are summarised below (see also appendix A for more information on EDISON):

1. There is an appreciation of the need for data-related skills across the across all of science, but not every student can (or should) become a DS expert.

2. There is acceptance that a number of groups of skills are needed, and that all scientists, to varying degrees will need a familiarity if not competence in them at some point in their careers. So not all need to be mastered immediately, but early familiarity is important.

38. Skills and competences that are needed include:

- Data management in the form of curation skills will be increasingly critical as research data have to be catalogued and presented to validate the research
- Data integration including understanding of formalized semantics and ontologies, particularly for disciplines with more heterogeneous datasets that need to be combined.
- Software skills and programming best practices;
- Data visualization and mapping tools;
- Tools needed for team collaboration and effective dissemination;
- Collaboration needed across disciplinary as well as geographical boundaries;
- All of which in turn require good communication skills across the collaborative team including with the non-scientific community;

39. Surprisingly perhaps, a clear need was identified for a range of non-technical skills that are applicable across scientific domains. These are variously termed soft-skills, power skills, or 21th Century skills. These range from team oriented skills such as creativity, collaboration, and critical thinking, to skills such as planning, communication and ethics (See Appendix B for full list).

4.1.2. Quantifying the skills shortage

40. Barend Mons, Chair of the initial advisory group for the EC Open Science Cloud, has spoken about the need for 500,000 data stewards in the years ahead, but it is not clear exactly what roles such individuals will have within organisations[6]. The EDISON project and various other papers have identified a number of references to the shortage in data science practitioners both in the research and business environments. The 2011 McKinsey Global Institute report predicted: “By 2018, the United States alone could face a shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts with the know-how to use the analysis of big data to make effective decisions.” The deficit in the scientific research area is more specific with needs being dependent on different disciplines. For example, a recent paper on bioinformatics (Attwood, 2017) suggests that, “at least a minimum level of computational skills can help life and computational scientists to communicate and interact with one another more effectively.”

5. Case studies

41. On the basis of literature analysis and interviews with experts a number of use cases are briefly described below. These give an insight into what is happening in relation to data and skills in different domains. Whilst many of the challenges and opportunities are shared, the responses and solutions that are being adopted are often different.

5.1. Ocean Science

42. Ocean science is a good example of a scientific discipline that is going through such a rapid transformation with more data than ever now being captured. Researchers are making use of a variety of sensors, operating at higher fidelity and frequency, to inform our understanding of the global ocean. Furthermore, in order to take on global challenges such as climate change, animal migration, and sea/air interaction, open science data from other areas has been integrated in to researchers' own data. However, this transformation of practices as the community moves from data scarcity to data deluge, brings with it challenges highlight the need for skills development for the future. Skills are needed to overcome both technical and cultural barriers. (Wilson, Colborne, & Smit, 2017)

5.2. Environmental research

43. The proliferation of environmental data in recent years opens up exciting possibilities for new inter-disciplinary approaches to Environmental Science. The recent pace of technological change has been such that appropriate skills to accomplish data-intensive research are lacking among environmental scientists, and therefore training and mentorship are needed in computational skills and other related areas. New skills are needed to harness the potential of many new data flows from related sources. These include Government funded centres around the world. A key challenge is the heterogeneity of the sources and the fact that there are many pressing environmental challenges in the world today that need to be addressed.

44. Software Carpentry, Data Carpentry, and other informatics and computational training workshops are available to address the shortfall in training in the formal education pathways. One challenge reported with respect to this community is that of specialist technical support. Within the university context, disciplines such as Environmental Science might expect to receive help and support from such areas as mathematics, statistics, and computer-science. However, these units are focused on their own interests and not operating in a support capacity. (Hampton et al., 2017).

5.3. Satellite Remote Sensing

45. One Remote Sensing Group (RSG) that was interviewed for this scoping work has an international standing in satellite-related science and discovery. The group produces multi-year global data sets on ozone, methane, aerosol and clouds, as well as satellite retrievals of volcanic ash. The group has a sophisticated approach to data management but is looking to develop further data science skills and competences. One approach being considered is knowledge transfer through increased interactions over projects and other initiatives with students and others from their local University and elsewhere. The group have significant quantities of data that can be of interest to data scientists in other fields.

5.4. Structural Biology

46. The EU-funded Instruct project describes structural biology as harnessing methods to delve beneath the limits of fundamental cellular processes by bringing molecules into sharper focus in 3D. However, despite the familiarity with the necessary tools and techniques, there remains a need for more data stewards and data scientists to support these processes and also to help equip structural biologists with the necessary data skills for the future.

5.5. Bio-informatics

47. Bioinformatics combines elements of biology and computer science with a focus on molecular genetics and genomics. As a key component of life sciences, there is a recognised need for greater data science skills and also for data steward practitioners. Short courses have been identified as a solution to improve expertise and confidence in data analysis and interpretation and there is a need to cultivate a new cadre of trainer scientists. There is also a demand for bioinformatics to be woven into life science degrees to combat the recognised shortage of mathematical and computational foundations amongst students. Code scripting and data analysis need to be added to traditional laboratory skills, but there is a need to train the future trainers too (Attwood, Blackford, Brazas, Davies, & Schneider, 2017).

5.6. Digital humanities model

48. In many ways the humanities represents the long tail of data-centric approaches to research with a clear focus on unstructured data. However this group has coalesced into the digital humanities and therefore found ways to share and harmonise their approaches to exploiting data for research through conferences, workshops, and books. In many ways the wider scientific community could learn much from the digital humanities.

6. Analysis – to guide and frame options

49. Open Science is growing and evolving, skills and competences are emerging and research communities are adopting and adapting where possible. Some communities are very confident, and some are concerned that they lack the skills, competences and specialists needed to drove forward their domain. There have been a number of initiatives looking at the big picture, aiming to produce a framework or other mapping of the skills needed across the board, or at least across a particular field.

50. Data Science can be seen as an emergent profession, but in many ways is a role to be taken on as and when required in organisations and career pathways as the need arises. Although other frameworks exist, the EDISON Data Science Framework (presented in Appendix A) serves to present a useful mapping of the full landscape encompassing both technical and non-technical skills and reaching out also to connect with existing related frameworks. The EDISON Data Science Framework offers many practical solutions, some of which have been adopted by a number of course designers across Europe, or else used to review existing courses to identify upgrades and changes to their syllabus. Standardisation is a big challenge going forward as all parties and stakeholders seek to harmonise their differences in the future. (Demchenko et al., 2016)

51. There are 5 key areas that require attention when considering the needs, challenges and potential interventions for data science.

6.1. Skills and competences

52. The skills and competences needed for Open Science are many and diverse. Whilst to varying levels of granularity these have been captured through initiatives such as EDISON, there remain questions about how to measure and quantify them. A number of approaches are proposed and evaluated in the literature and also at large. These include certificates, digital badges and competitions.

53. For measuring or at least capturing skills acquisition certification remains the most traditional whether for degree modules or short courses. Digital badges, along the Mozilla model, represent a functional approach to micro-learning that is functional and practical both on the part of the busy data scientists and the lean training provider. A third option is the Kaggle approach of competitions to demonstrate superior competence.[7]

54. Furthermore, the ways in which people want to acquire skills changes as their careers progress and their spare time diminishes. Course delivery needs to reflect this.

55. Whilst there is a proliferation of ad hoc resources including workshops and online materials available to scientists in a variety of locations, there is no systematic coordination of these resources will increase their efficiency and discoverability across different domains and communities. Resolving this is a non-trivial task. Arguably this lack of integrated data-related education is leading to lack of "information-oriented

citizens" at large. Some would argue that a coordinated worldwide action plan is needed (Attwood et al., 2017).

6.2. Teaching and training

56. There remains a need, at least in the short and medium term, to continue to reduce the gap between supply and demand with universities and training organisations delivering the appropriately skilled workforce that employers require. However, this is not straightforward. In the context of academic research there are two significant obstacles. Firstly, the fact that what we know as data science brings together two frequently decoupled subject areas, computing and computer science, and mathematics and statistics. Bringing together a “working knowledge of probability” with the computational skills required to handle large data sets is a non-trivial challenge.

57. The second challenge is the fact that data science from a teaching (and career pathway) perspective is moving forward along two very different pathways, it is both a topic in its own right with masters and doctoral courses on offer, and, increasingly, undergraduate courses too in data science emerging, whilst at the same time data science is recognized an essential theme within individual science subject areas. This situation has ramifications for policy-makers and funders: should data science be supported as a science or subject in its own right, or as a sub-set of competences within traditional subject area teaching? Teaching a data skills set across the university has significant implications for how institutions are configured.

58. Many scientists agree with aims the and objectives of Open Science however a significant barrier to greater adoption is a deep seated fear of sharing key research data on the part of the research scientist. The reasons are not simple or straightforward, they range from concerns over sensitive data, the need for further quality assurance, and a sense of competition. However, reticence can also come from a lack of adequate research data management (RDM) skills. Questions still remain over where responsibility resides within the organisation for delivering training within the organisation. Ideally, professional skills will be acquired before students enter the workplace (Wilson et al., 2017).

6.3. Certification and accreditation

59. The EDISON experience of attempting to establish certification in the context of data competences and skills is that it is challenging. In Europe, the task has been partly taken up by the federated European e-Infrastructure, EGI.

60. One of the challenges of delivering a fit-for-purpose workforce in an accelerated timeframe is enabling a range of qualifications ranging from traditional, longer degrees to shorter, more flexible certificates. One option that has been gaining traction is digital or open badges[8]. Open badges originated in 2010 from work at the Mozilla and MacArthur foundations. The digital badges are a flexible and practical way to recognise learning in the workplace but also from formal education. An open specification exists to enable organisations to set up their own badge issuing platform, and the badges themselves enable employers to better match individuals to jobs. Today the initiative has spread to many big names in both not-for-profit, and for-profit technology organisations across the world with further support from many big names such as museums, universities and major ICT companies.

61. The Centre for Open Science has introduced some digital badges which others have taken up[9]. These are effectively a reflection on what makes a good scientist. There are three badges available here: open data, open material, and preregistered. Journals choose to award badges based on criteria and evaluation. Recipients are advised as to how to add these to their papers (van Elk, Rowatt, & Streib, 2018).

62. Digital badges have been identified a popular and practical form of "micro-accreditation" and can be displayed on social media and personal profiles etc. Digital badges can also be seen as a good means for student motivation (Seery et al., 2017)

6.4. Professional profiles and career trajectories

63. The EDISON Data Science Framework (Appendix A) defines a collection of professional profiles that relate to the research lifecycle and also the associated data-related skills required. In reality this picture is complex. Different tasks require different skills and individuals may require different skills, to different depths, at different times in their careers.

64. The Research Software Carpentry[10] initiative has gone some considerable way to addressing the need for skills in some of these components of the skills and competences needed for Open Science. Volunteer instructors have run hundreds of events over many countries. Lesson materials are freely reusable under the Creative Commons – Attribution license. Software Carpentry and Data Carpentry have now merged into a US non-profit organisation called The Carpentries.

65. From the EDISON project, we know that data science is a complex and challenging area. Data science can, and indeed should, be broken down into a suite of distinct but complimentary professional profiles. Whilst this is a normal approach in other organisational structures, for the academic research community it raises deep rooted challenges to the traditionally clear division between researchers and support roles. This in turn presents challenges for publication of papers in terms of credit for authors and contributors, not to mention the publication of data, and accreditation for the work on the data before and after publication.

6.5. Human resources, recruitment

66. Much of the discussion on data competences and skills for Open Science focuses on either the practitioners or the educators, but increasingly HR professionals will play a role in matching experts to roles. They should be able to support more dynamic career trajectories and more complex training needs. In and beyond academia a number of existing professional organisations are taking an interest in this area, these include CILIP (The Library and Information Association), the BCS (British Computing Society – which has a global perspective despite its name), and even specialist employment agencies like Eden Smith which takes a strong interest in this field. Professional associations have started to emerge in recent years such as the Data Science Association (DSA) (DSA, 2016) and the European Association for Data Science (EuADS) (EuADS, 2016)

7. Conclusions

67. Skills and competencies will continue to be an important consideration as the growth in the use of data to conduct and underpin science grows. This will be accentuated as interdisciplinary collaboration, innovative sub-divisions of traditional domains, and new demands on policy makers and strategists all conspire to necessitate new approaches to skills and competences. Data science, in the widest sense of the term, requires technical skills – both IT, and mathematical and statistical - and increasingly competences in unstructured data too, as well as communication and leadership skills.

68. There are 4 areas where policy experimentation is taking place in relation to data competencies and skills for open science and where coordination, cooperation and mutual learning across countries could be of considerable value.

7.1. New models for science education and training

69. The recent pace of technological change has been such that appropriate skills to accomplish data-intensive research are lacking among a significant number of scientists. Training and mentorship are needed in computational skills, and the spectrum of Data Science skills. Models have been proposed for closing the skills gap however the challenge is to develop a model that can align delivery of needed skills to the right people at the right time in their professional career paths in a manner that they can integrate into their busy lives.

70. There is a pressing need for traditional universities to respond to the need for future scientists to be equipped with not just traditional science skills, but also data competences and skills. This implies a matrix of cross-discipline modules in an evolving portfolio of data science approaches. With an additional requirement for support for lifelong learning, the traditional structures may need to adapt. Departmental boundaries will also need to be made more porous to support the needed collaborations and conversations that can deliver people, tools and resources able to tackle the world's grand challenges.

7.2. Coordination and monitoring

71. Determining where the gaps and depletions are in data competences and skills across the science sectors is more of an art than a science. Although there are many who cannot envisage any coordination of the proliferation in data-related modules and training courses, there is a strong case for providing a sturdy steer on skills and related matters. There must already be a significant duplication of module material, student feedback, and requests for innovation. It is entirely possible that education and training is the least open corner of the Open Science community.

7.3. The role of Research infrastructures and data services

72. Many research infrastructures already provide Science as a Service in the form of remote access to some of their services. More could be done to support others to develop these, and develop APIs so that more experimentation could be conducted remotely. The potential of RIs for providing data analysis tools and expertise to the broader research community and for training that community via workshops and on-line tools is underexploited in some fields. However this requires resources and needs to be approached strategically.

7.4. Sustainability and flexibility

73. If Open Science doesn't flounder, then the future is going to look quite different. All data, information, and knowledge, and more besides, will be out there on view (after a brief embargo period). This implies very different models for payment, and charging, essentially asynchronous. How will research be funded? How will academic excellence be measured? As data and other resources (such as code and 3-D Labware) are pulled together, amalgamated and repurposed, how will the building blocks be identified and credited, curated and reused?

8. About the author

Steve Brewer

I am currently engaged in doctoral research at the University of Southampton looking at data-driven digital transformation, comparing different economic sectors including various scientific domains. I am interested in the disruptive nature of data science for transformational models across the digital economy. With a background in computing and IT, my recent background over the last ten years or so is in engagement and dissemination for various EU and UK projects that have covered Data Science skills, and distributed and utility computing.

These include the EU-funded EDISON project, and the RCUK-funded IT as a Utility Network+ in the UK. For the ITaaU Network+ I helped create a programme of workshops and conferences that brought an interdisciplinary community of academics, policy-makers and business leaders together to share insights into the trend towards IT as a Service across many domains. The EDISON project was established to build the Data Science profession, this was achieved in part by reducing the gap between what teaching and training establishments deliver and the requirements that employers demand. Over the course of the two-year project I helped orchestrate a broad and diverse community passionate about the potential of data science to change careers, organisations and domains. We ran successful conferences in the UK, Spain and Poland and are now building on that legacy with a forthcoming data science conference in Edinburgh in May. Prior to this Steve was the Chief Community Officer at the Amsterdam-based EGI.eu which supported data services for Europe's Research Infrastructures. Steve has also played an active part in the International Research Data Alliance over a number of years.

I have also recently founded a start-up which is focused on supporting organisations, teams and leaders achieve successful data-driven transformation.

9. References

Attwood, T. K., Blackford, S., Brazas, M. D., Davies, A., & Schneider, M. V. (2017). A global perspective on evolving bioinformatics and data science training needs. *Briefings in Bioinformatics*, (September), 1–7. https://doi.org/10.1093/bib/bbx100

Baden, T., Chagas, A. M., Gage, G., Marzullo, T., Prieto-Godino, L. L., & Euler, T. (2015). Open Labware: 3-D Printing Your Own Lab Equipment. *PLoS Biology*, *13*(3), 1–12. https://doi.org/10.1371/journal.pbio.1002086

Demchenko, Y., Belloum, A., Los, W., Wiktorski, T., Manieri, A., Brocks, H., … Brewer, S. (2016). EDISON Data Science Framework: A Foundation for Building Data Science Profession For Research and Industry. *2016 8Th Ieee International Conference on Cloud Computing Technology and Science (Cloudcom 2016)*, (February 2017), 620–626. https://doi.org/10.1109/CloudCom.2016.104

Frankland, J., & Ray, M. A. (2017). Traditional versus Open Access Scholarly Journal Publishing: An Economic Perspective. *Journal of Scholarly Publishing*, *49*(1), 5–25. https://doi.org/10.3138/jsp.49.1.5

Funamori, M. (2017). Open Science and the Academy: A Theoretical Discussion. *2017 6th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*, (1), 109–115. https://doi.org/10.1109/IIAI-AAI.2017.19

Hampton, S. E., Jones, M. B., Wasser, L. A., Schildhauer, M. P., Supp, S. R., Brun, J., … Aukema, J. E. (2017). Skills and Knowledge for Data-Intensive Environmental Research. *BioScience*, *67*(6), 546–557. https://doi.org/10.1093/biosci/bix025

Jomier, J. (2017). Open science - towards reproducible Research. *Information Services and Use*, *37*(3), 361–367. https://doi.org/10.3233/ISU-170846

Meissner, D., Anastasiia, N., & Dmitry, P. (2016). The Meaning of Digitalization for Research Skills: Challenges for STI Policy. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2857133

OECD. (2015). Making open science a reality. *OECD Science, Technology and Industry Policy Papers*, (25), 112. https://doi.org/http://dx.doi.org/10.1787/5jrs2f963zs1-en

on Open Science. (n.d.).

Ramírez-Montoya, M. S., & García-Peñalvo, F.-J. (2018). Co-creation and open innovation: Systematic literature review. *Comunicar*, *26*(54), 9–18. https://doi.org/10.3916/C54-2018-01

Seery, M. K., Agustian, H. Y., Doidge, E. D., Kucharski, M. M., O'Connor, H. M., & Price, A. (2017). Developing laboratory skills by incorporating peer-review and digital badges. *Chem. Educ. Res. Pract.*, *18*(3), 403–419. https://doi.org/10.1039/C7RP00003K

Siddaway, A. (2014). What is a systematic literature review and how do I do one? *University of Stirling*, (Ii), 1–13. https://doi.org/10.1016/j.jmwh.2009.03.017

Sinha, A., Thessen, A., & Barnes, C. (2013). Geoinformatics: Toward an integrative view of Earth as a system. *Geological Society of America Special Papers*, *500*(April), 591–604. https://doi.org/10.1130/2013.2500(19)

van Elk, M., Rowatt, W., & Streib, H. (2018). Good Dog, Bad Dog: Introducing Open Science Badges. *The International Journal for the Psychology of Religion*, *28*(1), 1–2. https://doi.org/10.1080/10508619.2018.1402589

Vicente-Saez, R., & Martinez-Fuentes, C. (2018). Open Science now: A systematic literature review for an integrated definition. *Journal of Business Research*, (December 2017). https://doi.org/10.1016/j.jbusres.2017.12.043

Wilson, L., Colborne, A., & Smit, M. (2017). Preparing Data Managers to Support Open Ocean Science : Required Competencies , Assessed Gaps , and the Role of Experiential Learning, 3902–3911.

10. Appendix A: EDISON Data Science Framework (EDSF)

10.1. Introduction

The EDISON project brought together results from field work, focus and expert group meetings, and surveys, which were analysed and transformed into a functional toolkit in the form of a framework. Just as this study is focused on data science practice in the scientific domain, the EDISON project primarily focused on the research infrastructure domain. The project was funded within the Directorate for Digital Excellence and Science Infrastructure in the eInfrastructure and Science Cloud Unit[2] which defines research infrastructures (RIs) as “facilities, resources and services used by the science community to conduct research and foster innovation.”

10.2. Background: The EU Horizon 2020-funded EDISON project[3]

The EDISON project delivered a framework that can be used as a flexible toolkit to understand and evaluate skills and competences and design teaching and training for data science. It can also assist with recruitment and career management of data science practitioners. The EDISON Data Science Framework (EDSF) for skills and competences is freely available and will be updated and reviewed over time.[4]

The EDISON project ran from 1 September 2015 to 31 August 2017 and provides a basis for establishing data scientist as a profession. The aim was that this would be achieved by aligning industry needs with available career paths, and supporting academies in reviewing their curricula with respect to expected profiles, required expertise and professional certification. The longer-term goal was a significant increase in the number and quality of data scientists graduating from universities and training institutions in Europe.

10.3. Overview: Data Science Framework (EDSF) overview

It is worth looking at the EDISON Data Science Framework (EDSF) in terms of what it is and how it was created, and how it can be used for data science curricula design and skills management. The two-year project had seven partners from academia, business (large and small, including teaching and training), plus a European research e-infrastructure that itself supports traditional research infrastructure. This core team was augmented with groups of expert advisors and pioneering users.

[2] https://ec.europa.eu/research/infrastructures/index_en.cfm?pg=about

[3] The EDISON project received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No. 675419.

[4] http://edison-project.eu/

The process of developing the framework began with a model of the professional labour market capturing the gap between what is being taught and the needs and requirements of employers.

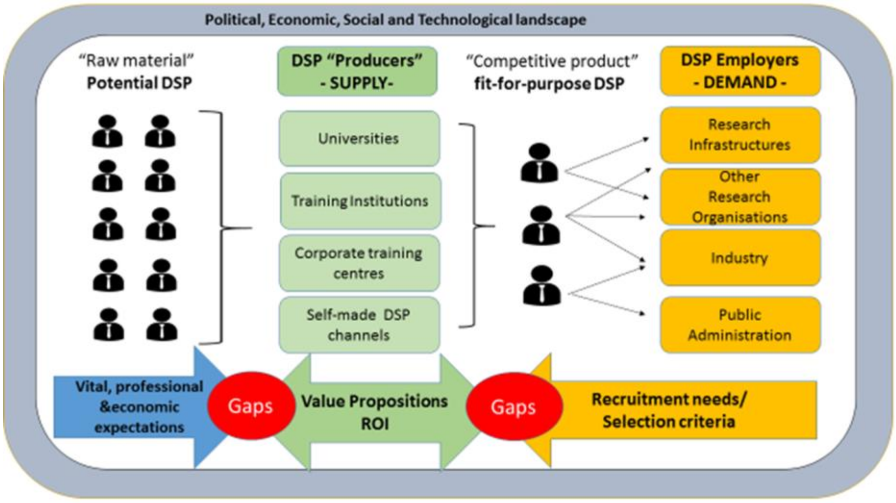

Figure 1 - The Data Scientist Professional labour market modelling

10.4. EDISON project fieldwork

Desk research was conducted and figures were obtained for quantitative evidence that confirmed the expected growth in demand for data science skills across the board and also the anticipated shortfall in individuals possessing these talents.[11],[12],[13] Furthermore, fieldwork also captured the qualitative characteristics of the impact of ICT and data-related skills shortage. The impacts included problems with hiring and retention. There is also evidence of open data science and data steward positions staying unfilled. Companies and research organisations also want experienced data science workers so there is no time to acquire necessary experience. There is also a Millennials factor, do we understand the differences of the millennials workforce. The emergent challenge is: how to obtain, train in shorter period and sustain new digital (ICT and Data-related) skills in organisations and laboratories.

The European Open Science Cloud (EOSC[14]) initiative represents a landmark opportunity and demand case for data science. For this and other data-intensive initiatives a suitably equipped workforce will be required. This is an ongoing and growing situation demanding a sustained supply of data and ICT skills. A High-Level Expert Group report on EOSC has already raised questions regarding a critical need for core data experts. Challenges like these raise the question of balancing education against training. Training is a short-term solution lacking depth, education forms a basis for sustainable skills development but takes time and has a greater cost. Given the rate of change of technology, education has to have a focus on instilling lifelong learning skills that prepare for a dynamic and changeable future. Otherwise too much of the curricula is outdated on completion of the course.

10.5. EDISON Data Science Framework (EDSF)

The EDISON Data Science Framework (EDSF) comprises an integrated collection of products for data science skills management and tailored education. These have been developed and designed to be compliant with the ESCO Competence Framework for ICT, e-CFv3.0, a key EU standard on competences and professional occupations. The EDISON Data Science Framework has been designed to serve a number of specific functions: course design; skills development and career management for core data experts and related data handling professions; capacity building and data science team design; academic programmes and professional training courses (self-) assessment and design. In addition, the project convened an EU-wide network of champion universities pioneering data science academic programmes, and consortium members engaged with relevant activities and groups in the global Research Data Alliance. There was also interaction with international professional organisations such as the Institute of Electrical and Electronics Engineers (IEEE[5]), the Association for Computing Machinery (ACM[6]), and the Business-Higher Education Forum (BHEF[7]).

The key components of the EDISON Data Science Framework (EDSF) are as follows:

- Data Science Competence Framework (CF-DS)
- Data Science Body of Knowledge (DS-BoK)
- Data Science Model Curriculum (MC-DS)
- Data Science Professional profiles (DSP)
- Data Science Taxonomies and Scientific Disciplines Classification
- EDISON Online Education Environment (EOEE)

The methodology behind the design of the EDSF included a job market study, coupled with a survey of existing practices in academic, research and industry. Review and feedback on this research from the Expert Liaison Group (ELG), expert community members, domain experts and others.

[5] https://www.ieee.org/

[6] http://www.acm.org

[7] http://www.bhef.com

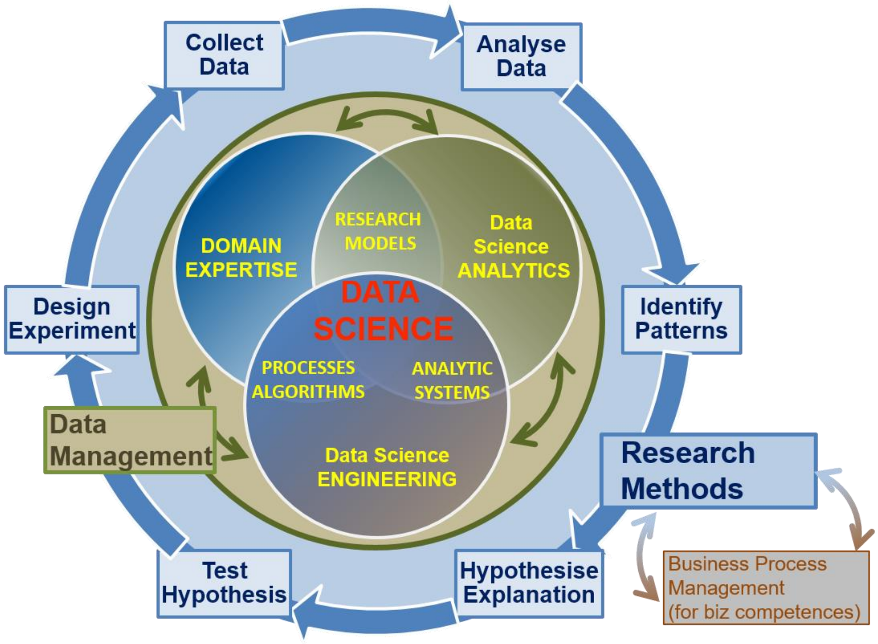

Figure 2 - Data Science Competences Groups

The above diagram captures the high-level model for the main competence groups within data science. These are: data science analytics, data science engineering, data management, scientific methods, and domain knowledge and expertise. These are augmented by the following activities: scientific methods, design of experiments, collection of data, analysis of data, pattern identification, hypothesis explanation, and hypothesis testing.

The Framework also maps these high-level competence groups into sub-categories of skills and experiences:

Group 1: Skills/experience: related to competences

- Data Analytics and Machine Learning
- Data Science Engineering (hardware and software) skills
- Data Management/Curation (including both general and scientific data management)
- Scientific/Research Methods
- Application/subject domain related (research or business)
- Mathematics & Statistics

Group 2: Big Data (Data Science) tools and platforms: related to skills

- Big Data Analytics platforms
- Mathematics & Statistics applications & tools
- Databases (SQL and NoSQL)
- Data Management and Curation platform
- Data and applications visualisation
- *Cloud based platforms and tools*

Group 3: Programming and programming languages and IDE: related to skills

- General and specialized development platforms for data analysis and statistics

Group 4: Soft skills/power skills or 21st Century Skills: orthogonal/cross-cutting skills

- Personal, inter-personal communication, team work, professional network

This last section, Group 4, proved both interesting and challenging, both in terms of the name and its existence in a framework defining data science. Whatever they are called, these valuable cross-cutting skills go towards creating, what is sometimes referred to as a T-shaped profile: a person who combines deep specialist skills in a particular discipline with a broad portfolio of common collaborative skills. Some participants in discussions questioned the need for an extra dimension in terms of 21st century skills, asserting that these were more for entrepreneurs and startups. However, we have seen job advertisements for data scientists specifically requiring such skills. Collating research between European, US and Asia Pacific projects and other research papers on the topic identified the following 21st Century Skills: critical thinking, communication, collaboration, creativity and attitude, planning and organizing, business fundamentals, customer focus, working with tools and technology, dynamic (self-) re-skilling, professional networking, ethics. Whilst realistically not every data scientist will have all of these skills, the list forms a valuable dimension within the framework that can help define recruitment, team configuration and course design.

The term competence in the context of digital technology and data science practices is used to quantify an ability to do something. This implies knowing about something, and also knowing how to actually do that something. Simplistically, competence = knowledge + skill. Where knowledge is acquired through education, and skills are acquired through experience. As has traditionally been the case in professional fields such as medicine and engineering, this requires a structured and systematic approach.

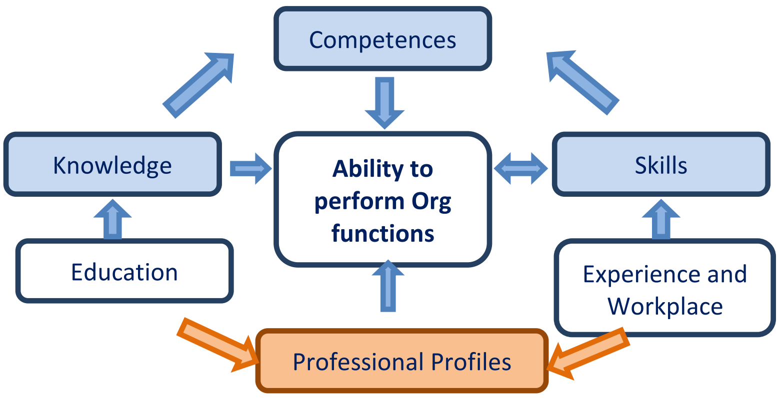

Figure 3 - Competence is a demonstrated ability to apply knowledge, skills and attitudes for achieving observable results

With these details in place, the EDSF then becomes useful as an adaptable tool for matching candidates to roles. A prototype application was created that made use of gamification and visualisation to compare individual competences formally identified for a particular job to those captured for a particular candidate. The aim was to enable particular training courses to be followed to remedy the identified skills deficiencies.

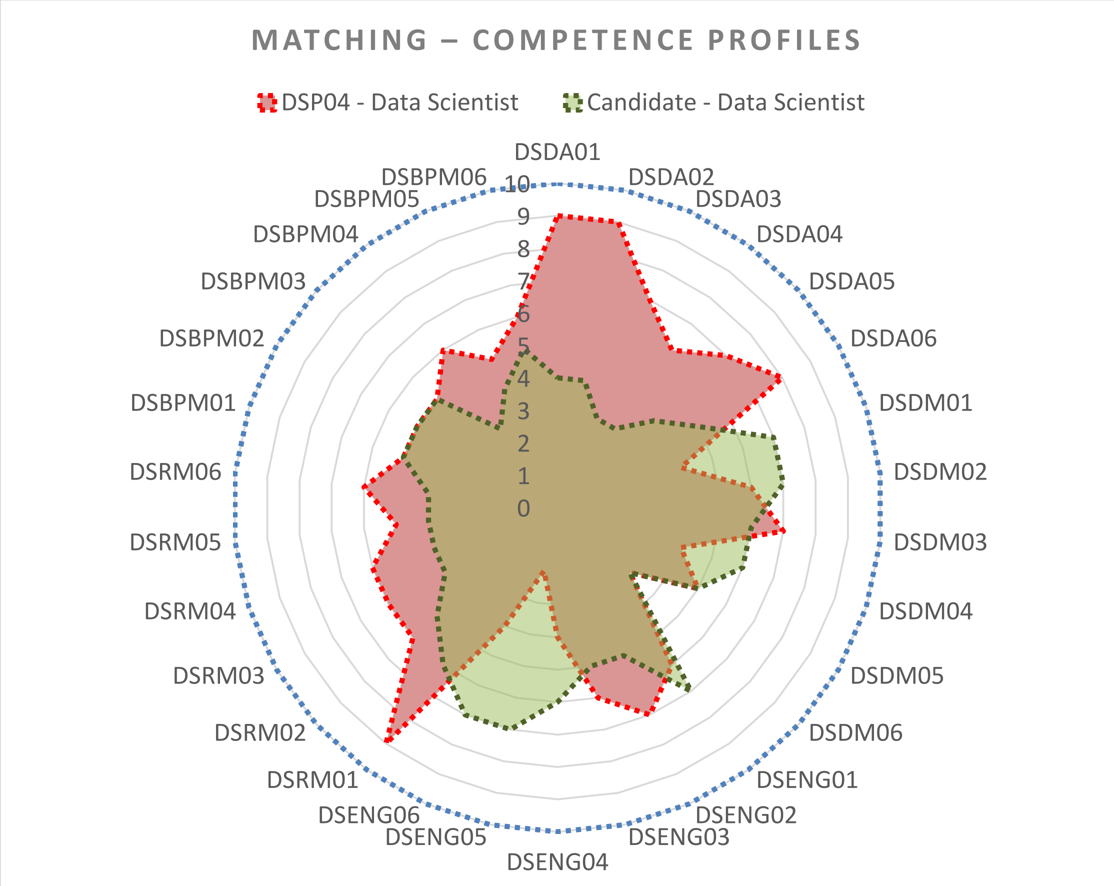

Figure 4 - Individual Competences Benchmarking

In the example given the red polygon indicates the competences required for the chosen professional profile, data scientist (general), and the grey polygon indicates the candidates' competences profile. As a result, a structured plan can be drawn up for future training and skills acquisition that addresses the differences.

The EDISON project has now been completed, however the legacy remains in terms of the EDSF, the community portal and a strong community of champion practitioners and expert advisors. It is anticipated that future releases of the EDISON Data Science Framework will follow.

11. Appendix B: 21st Century Skills (DARE & BHEF & EDISON)

1. **Critical Thinking:** Demonstrating the ability to apply critical thinking skills to solve problems and make effective decisions
2. **Communication:** Understanding and communicating ideas
3. **Collaboration:** Working with other, appreciation of multicultural differences
4. **Creativity and Attitude:** Deliver high quality work and focus on final result, initiative, intellectual risk
5. **Planning & Organizing:** Planning and prioritizing work to manage time effectively and accomplish assigned tasks
6. **Business Fundamentals:** Having fundamental knowledge of the organization and the industry
7. **Customer Focus:** Actively look for ways to identify market demands and meet customer or client needs
8. **Working with Tools & Technology:** Selecting, using, and maintaining tools and technology to facilitate work activity
9. **Dynamic (self-) re-skilling:** Continuously monitor individual knowledge and skills as shared responsibility between employer and employee, ability to adopt to changes
10. **Professional networking:** Involvement and contribution to professional network activities
11. **Ethics:** Adhere to high ethical and professional norms, responsible use of power data driven technologies, avoid and disregard un-ethical use of technologies and biased data collection and presentation

12. Endnotes

[1] By Jim Gray, Microsoft, 2009. Edited by Tony Hey, Kristin Tolle, et al. http://research.microsoft.com/en-us/collaboration/fourthparadigm/
2 https://rd-alliance.org/data-harvest-report-sharing-data-knowledge-jobs-and-growth.html

[3] https://open-labware.net/ Open Labware

[4] https://www.thingiverse.com Thingiverse

[5] NIST Big Data Interoperability Framework: Volume 1, Definitions: http://dx.doi.org/10.6028/NIST.SP.1500-1

[6] http://e-irg.eu/news-blog/-/blogs/we-need-500-000-respected-data-stewards-to-operate-the-european-open-science-cloud

[7] https://www.kaggle.com/ The Home of Data Science & Machine Learning

[8] https://openbadges.org/ Mozilla Open Badges

[9] https://cos.io/our-services/open-science-badges/ - Centre for Open Science

[10] https://software-carpentry.org/

[11] Final Report on European Data Market Study by IDC (Feb 2017)
–The EU data market in 2016 estimated EUR 60 Bln (growth 9.5% from EUR 54.3 Bln in 2015)]

[12] PwC and BHEF report "Investing in America's data science and analytics talent: The case for action" (April 2017)
–http://www.bhef.com/publications/investing-americas-data-science-and-analytics-talent

[13] Burning Glass Technology, IBM, and BHEF report "The Quant Crunch: How the demand for Data Science Skills is disrupting the job Market" (April 2017)
–https://public.dhe.ibm.com/common/ssi/ecm/im/en/iml14576usen/IML14576USEN.PDF
[14] https://eoscpilot.eu - European Open Science Cloud – currently at the pilot project stage.