Now is the best time to be a Data Scientist or a Data Steward in Europe

Why the European Union plans to train over half a million data scientists and data stewards

Ioana Spanache, PhD · Follow

Published in Towards Data Science · 9 min read · Oct 9, 2020

171

Photo by Pixabay on Pexels.com

Content

Introduction

In 2016, a group of EU high-level experts were estimating that Europe will need for making Open Science possible **over half a million "Core Data Experts"** within a decade.

In their view, these "Core Data Experts" will support its 1.7 million researchers and 70 million professionals in Science and Technology throughout the entire research lifecycle, ensure good data management, help in data capturing (formats, metadata, standards, provenance, publishing), as same as in data analysis.

Two years later, in 2018, the idea is reinforced, and the "Core Data Experts" are identified as being **Data Scientists and Data Stewards** (Turning FAIR into Reality Final Report and Action Plan) who, in addition to technical expertise, will also be required to have domain knowledge in the research and innovation fields they will be working.

In the context of research, the two skillsets are defined as follows:

- **Data Science —** *"the ability to handle, process and analyze data to draw insights from it"*; And the required skills are: computer science, software development, statistics, data visualization, machine learning, as same as computational infrastructures.
- **Data Stewardship —** *"a set of skills to ensure data are properly managed, shared and preserved, both throughout the research lifecycle and for long-term preservation"*. Some examples of mentioned skills are: information management, data cleaning, data management.

Given these identified needs, the Report calls for new formal or less formal educational and training programs that can help prepare the large cohort of data scientists and data stewards required to make FAIR data possible.

What is also interesting here is that, the Report also calls for researchers to become data-savvy, so that they can better make use of available data and technologies.

What is FAIR data?

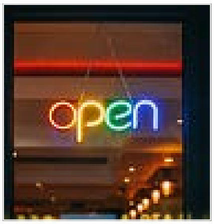
From a philosophical perspective, the principles of FAIR data, as same as the entire Open Science movement are related to the idea that science, as same as research data can be considered a global public good. Especially when they are funded by public money.

More about Open Data as a public good, as same as information about sources of free open data can be found in my previous article:

Data Science for Social Good: Best Sources for Free Open Data

Types, benefits and where to find them

towardsdatascience.com



Every year, the amount of data that is being generated, in science and beyond, grows exponentially. This leads to great opportunities in how we use that data in science, in making business decisions or in building evidencebased policies. But there are also big challenges when it comes to capitalizing on the produced data, such as the quality of data and its long term preservation.

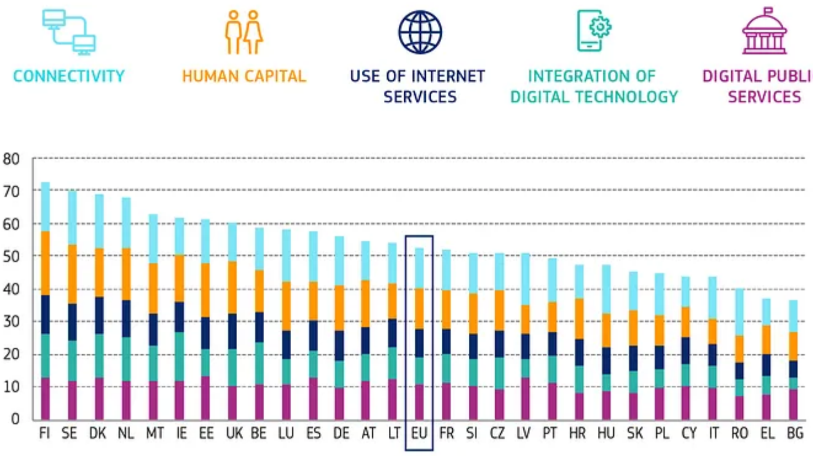
In this logic, we must be able to turn data into knowledge that afterwards results in action.

So, what does FAIR stand for? (FAIR data principles)

- **Findable** — Both humans and computers should be able to easily find data and its metadata. For this, metadata should be machine readable, allowing for the automatic discovery of datasets and services.
- **Accessible** — data should come with clear information regarding how it can be accessed, even if that involves authentication or authorisation. Plus, the corresponding metadata should be accessible even when the data are no longer available.
- **Interoperable** — Data needs to be interoperable with applications (such as API) or workflows for analysis, storage and processing.
- **Reusable** — This principle refers to the main goal of FAIR, which is to optimise and facilitate the reuse of data. For this, metadata and data need to be well-described, to include information about usage license, as same as provenance.

FAIR data is not equal with open data. For example, as can be noticed from the Accessibility principle, sometimes it can involve authorisation or authentication.

According to Simon Hudson, Chair of the EC Expert Group on FAIR data, FAIR will not be possible without “considerable and wide-reaching enhancement of skills for data science and data stewardship.” And Europe is not really a best practice example when it comes to digital skills, with an average of 42% of its population lacking in this regard.



Digital Economy and Society Index 2020

What is expected of a Data Scientist in science and research?

Together with Data Stewards, Data Scientists will be required to support researchers throughout the research lifecycle, and will be embedded within research projects at institutional level or in specialized services per domain.

In many cases, Data science positions are filled by people who already have a research background or who were trained as information professionals. The former is especially important, as much of the work of Data Scientist in this field is discipline specific, and requires a deep understanding of both curation and research.

After looking at some job descriptions of Data Scientist, offered by research performing organizations, I have noticed that, beside the usual requirements related to machine learning, statistics, Python, data visualization, sometimes high performance computing or cloud computing, some of the future employees are also expected to be knowledgeable in research discipline-related software tools, to write content for scientific journal articles and blogs, or to attend and present during national and international conferences. In terms of education, Data Scientists in this area need to have masters or PhDs in fields such as astronomy, physics, statistics or computers science (varying according to the research discipline).

One potential drawback for those who come from Computer Science or other related disciplines, and not from one of the specific research fields, is the requirement to be at least a First Stage Researcher (R1). In those cases, it might be preferable to first become a researcher in that field, and then add Data Science skills. However, this is not the case in all circumstances.

On the flip side, an advantage of working as a Data Scientist in research performing organizations can be the fact that it might involve contributing to social impact. For example, some jobs in this field require applying Data Science to social-tech topics such as pandemics, fake news, citizen participation or renewable energy.

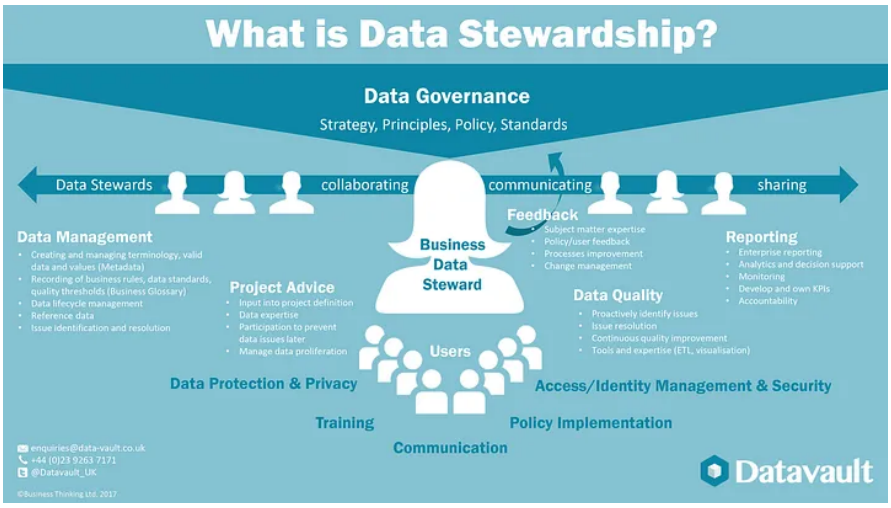
Data Science for Social Good

Going beyond what type of movies we want to see, to what type of world we want to live in. Resources, examples and…

towardsdatascience.com

What does a Data Steward do?

In the business sector, Data Stewardship is associated with improving data quality, dealing with sensitive and confidential data, data cleaning, defining policies and monitoring systems, defining standards, adding metadata, and the entire process of data management over its lifecycle, from curation, and until it becomes obsolete.



What is Data Stewardship?

Data Governance

Strategy, Principles, Policy, Standards

Data Stewards collaborating communicating sharing

Business Data Steward

Feedback

- Subject matter expertise
- Policy/user feedback
- Processes improvement
- Change management

Data Management

- Creating and managing terminology, valid data and values (Metadata)
- Recording of business rules, data standards, quality thresholds (Business Glossary)
- Data lifecycle management
- Reference data
- Issue identification and resolution

Project Advice

- Input into project definition
- Data expertise
- Participation to prevent data issues later
- Manage data proliferation

Reporting

- Enterprise reporting
- Analytics and decision support
- Monitoring
- Develop and own KPIs
- Accountability

Data Quality

- Proactively identify issues
- Issue resolution
- Continuous quality improvement
- Tools and expertise (ETL, visualisation)

Users

Data Protection & Privacy

Access/Identity Management & Security

Training

Policy Implementation

Communication

enquiries@data-vault.co.uk
+44 (0)23 9263 7171
@Datavault_UK

©Business Thinking Ltd. 2017

Datavault

Infographic by Datavault

Data Steward
"I ensure data usage complies with our policies."
Data Scientist
"I want to cleanse the data and find appropriate insights."
Data Engineer
"I need to understand requirements & build data platforms to fit business needs."
Marketer
"I want to use the right data for the right marketing needs."

Data Governance roles at Adobe

At Adobe, Data Stewards are responsible especially for interpreting regulations, any contractual restrictions, and policies, and applying them to the data. Therefore, among their responsibilities are: creating data policies and applying them to datasets; reviewing data, datasets, and data samples to apply and manage metadata usage labelling, and other.

As we can see from the Adobe example and other, in certain settings, Data Stewardship is associated more with data policies and dealing with sensitive and confidential data. And we will see next how, when it comes to the European setting of Open Science, Data Stewardship revolves especially around implementing the FAIR principles.

. . .

In the context of research, Data Stewards can be responsible for activities such as data cleaning, for eliminating inconsistencies, organizing and structuring the data, dealing with metadata, ensuring reusability, access, and long-term preservation of data (even when technology changes), plus other data management operations. Depending on their level of experience and expertise, in some cases, Data Stewards might also be required to deal with defining standards, best practices and interoperability frameworks.

However, regardless of sector, public or private, Data Stewardship is more than data management, as it also includes data collection or capturing, processing, long-term preservation, and its reuse.

In a presentation regarding **Data Stewardship at the Delft University of Technology**, is mentioned that Data Stewards should be able to answer to all data related question in a specific research domain. Some questions might be related to data storage, data recovery and back-up, how to handle confidential and sensitive data, sharing data in the context of patenting, offer help in drawing Data Management Plans and keeping track of data, ensuring long-term preservation of data.

Moreover, at TU Delft, data steward job specifications include as a core requirement knowledge of the research area in which the Data Steward is expected to work in.

Short intro to Data Stewardship at the TU Delft: Shalini Kurapati
Compartir
Data Stewardship at TU Delft
Supporting disciplinary research data needs
TUDelft
MÁS VÍDEOS
0:00 / 3:45 • Data Stewardship at TU Delft Supporting disciplin...
YouTube

https://youtu.be/WmG3ItcIaSE

Some examples of job related responsibilities of Data Stewards (extracted from EURAXESS) are: support researchers and managers throughout the research lifecycle, development of research data management workflows and best practices, Data Management Planning of research projects, research data curation, develop and run trainings and workshops for researchers, data management, ensure data validity, protection and security, ensure compliance with international standards regarding formats, metadata, monitor data management, maintain glossaries, implement FAIR data principles, development of domain-specific vocabularies, ontologies and metadata schemas, participation in research activities and elaborating publications.

How the two roles compare and intersect with one another?

Average annual salary in the U.S. in 2020:

- Data Steward — $67,982 (Payscale.com)
- Data Scientist — $96,101 (Payscale.com)

Popular skills (in that order, on Payscale.com):

- Data Steward: data analysis, data management, data quality, Microsoft Excel, SQL
- Data Scientist: machine learning, Python, data analysis, statistical analysis, R

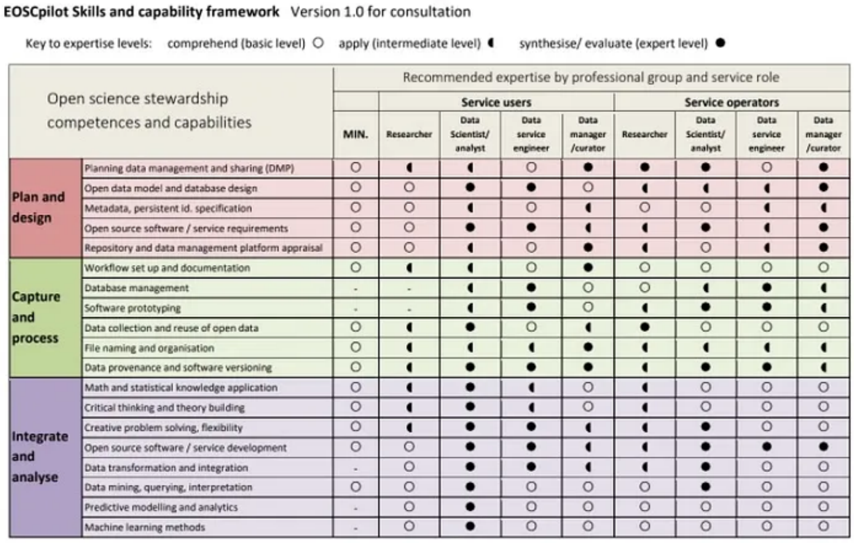
No. of job listings on EURAXESS, a European platform dedicated to researchers:

- **Data Steward: only 3**, in the Netherlands, Germany and Luxembourg
- **Data Scientist: 488**, but the number is not entirely accurate as the list also includes other types of jobs, such as UI/UX designer, Software Developer

When it comes to job responsibilities in research performing organizations, the two overlap to a certain degree. Data Scientist roles also include tasks related to data management or dealing with FAIR data principles.

The authors of the Turning FAIR into Reality report consider that data science and data stewardship can often be combined in the same individual, but that it is preferable to drive greater specialization in the two areas. In some cases, this might depend on the available budget and on how data-oriented is the research organization. For many such organizations it might be difficult to have two separate positions in this regard.

Below, is an example of how a pilot project related to the European Open Science Cloud described in detail the skills and capability framework when it comes to data stewardship. In this case, Data Science as part of the larger concept of data stewardship. For the entire framework, please see the report.



EOSCpilot Skills and capability framework Version 1.0 for consultation

Key to expertise levels: comprehend (basic level) ○ apply (intermediate level) ◖ synthesise/ evaluate (expert level) ●

Open science stewardship competences and capabilities		Recommended expertise by professional group and service role								
			Service users				Service operators			
		MIN.	Researcher	Data Scientist/ analyst	Data service engineer	Data manager /curator	Researcher	Data Scientist/ analyst	Data service engineer	Data manager /curator
Plan and design	Planning data management and sharing (DMP)	○	◖	◖	○	●	●	●	○	●
	Open data model and database design	○	○	●	●	○	◖	◖	◖	●
	Metadata, persistent id. specification	○	○	◖	○	◖	○	○	◖	◖
	Open source software / service requirements	○	○	●	●	◖	◖	●	◖	●
	Repository and data management platform appraisal	○	○	◖	○	●	◖	○	◖	●
Capture and process	Workflow set up and documentation	○	◖	◖	○	●	○	○	○	○
	Database management	-	-	◖	●	○	○	◖	●	◖
	Software prototyping	-	-	◖	●	○	◖	●	●	◖
	Data collection and reuse of open data	○	◖	●	○	◖	●	○	○	○
	File naming and organisation	○	◖	◖	◖	●	◖	◖	◖	◖
	Data provenance and software versioning	○	◖	●	●	●	◖	●	●	◖
Integrate and analyse	Math and statistical knowledge application	○	◖	●	◖	○	◖	○	○	○
	Critical thinking and theory building	○	◖	●	◖	○	◖	○	○	○
	Creative problem solving, flexibility	○	◖	●	●	◖	◖	●	○	○
	Open source software / service development	○	○	●	●	◖	◖	●	●	●
	Data transformation and integration	-	○	●	●	◖	◖	●	○	○
	Data mining, querying, interpretation	○	○	●	○	○	○	●	○	○
	Predictive modelling and analytics	-	○	●	○	○	○	○	○	○
	Machine learning methods	-	○	●	○	○	○	○	○	○

EOSC Pilot Skills and Capability Framework

Other European initiatives for defining Data Science and Data Stewardship can be consulted on the following pages:

- Edison project — acceleration and creation of the Data Science profession
- DigCurV — a curriculum framework for digital curation

Conclusions

Given that, for now, there are not many job openings available for Data Stewards, it seems that it might not be the best career to pursue. However, the low number of openings might be also because the role is sometimes integrated within that of Data Scientist, as the latter also involves dealing with research data management, and the FAIR principles for data (at least, in Europe).

When the two roles are separate, Data Stewards can be extremely important for the reuse of data, even if it is related to the private or public sector. Data has greater value and can yield better results and impact when it can be reused and built upon. In order to do their jobs, Data Scientists need access to good, reliable data, and, in time, Data Stewards can help in that regard.

Data Stewardship can be a solution to the g*arbage in, garbage out* (GIGO) problem in Data Science. If you don't have good quality data to start with, it doesn't matter how good a Data Scientist is. As a friend of mine once said, we need to **start by how and what type of data is collected, and then move to processing and drawing insights** from it.

Either way, the demand for data experts in Europe is high and expected to increase even more, as EU institutions wish to catch up to the United States and China in the innovation and tech race. And this cannot be possible without embracing new technologies and stepping up its game when it comes to human resources.

If you wish to try your hand on open research datasets, check the following repository:

Zenodo — Research. Shared.

August 12, 2020 (v1.0) Dataset Open Access Bardi, Alessia; Kuchma, Iryna; Bobrov, Evgeny; Truccolo, Ivana; Monteiro…

zenodo.org