Customisable Data Science Educational Environment: From Competences Management and Curriculum Design to Virtual Labs On-Demand

Yuri Demchenko, Adam Belloum, Cees de Laat
University of Amsterdam, The Netherlands
{A.S.Z.Belloum, y.demchenko,
C.T.A.M.deLaat}@uva.nl
Charles Loomis
SixSq Sàrl, Switzerland
cal@sixsq.com

Tomasz Wiktorski
University of Stavanger, Norway
tomasz.wiktorski@uis.no
Erwin Spekschoor
Amsterdam Business School
erwin.spekschoor@planet.nl

Abstract— Data Science is an emerging field of science, which requires a multi-disciplinary approach and is based on the Big Data and data intensive technologies that both provide a basis for effective use of the data driven research and economy models. Modern data driven research and industry require new types of specialists that are capable to support all stages of the data lifecycle from data production and input to data processing and actionable results delivery, visualisation and reporting, which can be jointly defined as the Data Science professions family. The education and training of Data Scientists currently lacks a commonly accepted, harmonized instructional model that reflects all multi-disciplinary knowledge and competences that are required from the Data Science practitioners in modern, data driven research and the digital economy. The educational model and approach should also solve different aspects of the future professionals that includes both theoretical knowledge and practical skills that must be supported by corresponding education infrastructure and educational labs environment. In modern conditions with the fast technology change and strong skills demand, the Data Science education and training should be customizable and delivered in multiple form, also providing sufficient data labs facilities for practical training. This paper discussed both aspects: building customizable Data Science curriculum for different types of learners and proposing a hybrid model for virtual labs that can combine local university facility and use cloud based Big Data and Data analytics facilities and services on demand. The proposed approach is based on using the EDISON Data Science Framework (EDSF) developed in the EU funded Project EDISON and CYCLONE cloud automation systems being developed in another EU funded project CYCLONE.

Keywords—Data Science, Data Scientist Professional, Big Data, EDISON Data Science Framework (EDSF), Data Science Competences Framework, Data Science Body of Knowledge, Data Science Model Curriculum, FAIR principles in Open Education, Virtual Data Labs (VDLabs), CYCLONE Cloud Automation platform, SlipStream.

I. INTRODUCTION

Industry and science are evolving into a new data driven and data-enabled technological world/space that is characterized by fast development, high level of technologies convergence and increased role of knowledge, skills and human factors to enable continuous and sustainable science and technology development. Such type of economy requires new type of data driven and Data Science and Analytics enabled competences and workplace skills.

Sustainable development of the modern data driven economy requires re-thinking and re-design of both traditional educational models and existing courses reflecting multi-disciplinary nature of Data Science and its application domains. However, at present time most of the existing university curricula and training programs are built based on available courses and cover limited set of competences and knowledge areas that are related to multiple Data Science and general data management professional profiles and organisational roles required by research and industry. In conditions of continuous technology development and shortened technology change cycle, Data Science education requires effective combination of theoretical, practical and workplace skills. Importance of effective use of existing data analytics and data management platforms and tools and corresponding hands on experience is growing and their elements need be generically incorporated into modern curriculum design.

The EDISON Data Science Framework (EDSF) [1], which is the products of the EDSION Project, provides a basis for building effective educational environment combining educational or training components and practical hands on experience with virtual and data labs. The educational Data Science labs and project development environment can benefit from using clouds and available data analytics and data handling applications and services that can be made and their e available on demand for specific time

2330-2186/17 $31.00 © 2017 IEEE
DOI 10.1109/CloudCom.2017.59

periods when the education of training take place. This paper proposes to use CYCLONE framework and SlipStream cloud automation tools for designing educational Data Science labs that can be provisioned on demand as used together with the campus facilities to create hybrid virtual educational environment.

The paper refers to the previous authors works that researched new approaches to building effective curricula in Cloud Computing, Big Data and Data Science [6, 7. 8, 9] and discussed cloud automation platforms [10] that all can be used for creating modern Data Science Education Environment.

The paper is organized as follows. Section II looks into available studies indicating demand for Data Science related specialists and describes the challenges and specifics in professional education and training of the data scientists. Section III describes the proposed EDSF and its components. Sections IV describes the EDSF data model and its use for defining customised curricula and required practical training. Section V provides brief overview of the cloud based Data Science and Analytics services and platforms and section VI provides information about SlipStream cloud automation platform and its functionality for building virtual labs. Section VII provides summary and suggestions for future work.

II. Demand for Data Scientists and Challenges

Growing demand for Data Science and Analytics enabled and general data driven professions is confirmed by multiple European and global market studies.

The IDG report 2017 [11] provided deep analysis of the European data market and growing demand for data workers, the value of the data market, the number of data user enterprises, the number of data companies and their revenues, and the overall value of the impact of the data economy on EU GDP. The EU data market is estimated as EUR 60 Bln with growth to EUR 106 Bln in 2020. With the total number of data workers to grow 6.1 mln (2016) 10.4 million in 2020 the data worker skill gap is estimated as 769,000 or 9.8% (2020). Addressing this demand and gap is becoming critical for European economy and challenge for universities. The report stresses that not satisfied demand in data workers with lead to under-performing economy, industry, research and loss of competitiveness.

Business Higher Education Forum (BHEF) has published two important reports in cooperation with PriceWaterhouseCoopers (PwC), IBM and Burning Glass Technologies (BGT) [12, 13] that studied Data Science and Analytics (DSA) job market in US and identified a number of actions to be addressed by business, higher education, government and professional organisations to address increased demand and growing gap in demand and supply of skilled DSA workforce capable to effectively work in modern data driven economy.

The authors' experience of developing a pilot project for re-/up-skilling employees of one of the Dutch governmental organisations confirmed trend that organisations, in a way to become data driven and agile, will intend to make the existing organisational roles DSA enabled and require corresponding DSA training in a customizable and flexible form.

An effective professional education needs to provide a foundation for future continuous professional self-development and mastering new emerging technologies, that can provide a basis for the life-long learning model adoption. Wide use of available online resources and platforms for so demanded Data Science and other digital and data skills will facilitate adoption of FAIR principles in the future Open Education to become Findable, Accessible, Interoperable, and Re-usable that were initially proposed for Open Data [14]. This will contribute to the future shift in technology development paradigm from infrastructure and data to skills and insight. The universities can contribute to building FAIR life-long educational space that can serve both organisational and individual needs of students and learners, including support for widely apprised citizen scientists.

Such educational environment should beneficially use the same or similar data analytics tools and platforms as used by companies. The solution may be in using cloud based resources that can eventually instantiate workplace environment used by companies. Example to this is wide use of such cloud based online services as Office 365 with their data analytics services, online Google services GoogleSheets, GoogleAnalytics, IBM Docs services. Using cloud based data analytics services and platforms will also make Data Science education and training more practically effective as companies increasingly using such tools and platforms in their practical daily work.

III. EDISON Data Science Framework (EDSF)

Designing future effective Data Science educational environment will require developing and widely accepted a general framework for Data Science education, curriculum design and competences management that can be based on the proposed EDISON Data Science Framework (EDSF) that is a core product of the EDISON Project. EDSF provides a basis for the definition of the Data Science profession and other components related to Data Science education, training, organisational roles definition and skills management, as well as professional certification and career transferability.

Figure 1 below illustrates the main EDSF components and their inter-relations:

- CF-DS – Data Science Competence Framework [2]
- DS-BoK – Data Science Body of Knowledge [3]
- MC-DS – Data Science Model Curriculum [4]
- DSPP - Data Science Professional profiles and occupations taxonomy [5]
- Data Science Taxonomy and Scientific Disciplines Classification

The proposed framework provides the basis for other components of the Data Science professional environment such as

- Data Science Education Environment (DSEE) intended to be cloud based and customizable and based on standards

- Education and Training Directory connected to Marketplace and Virtual Data Labs
- Data Science Community Portal (CP) that provides information and community support services. It also provides gateway to DSEE, Marketplace and Virtual Data Labs. CP is intended to include tools for individual competences benchmarking and personalized educational path building

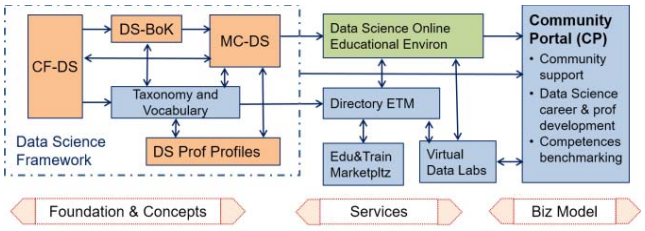

Figure 1 EDISON Data Science Framework components and Data Science Educational environment.

The CF-DS provides the overall basis for the whole framework. The CF-DS includes the core competences required for the successful work of a Data Scientist in different work environments in industry and in research and through the whole career path. The CF-DS is defined using the same approach as e-CFv3.0 [15] (competences defined as abilities supported by knowledge and skills with applied proficiency levels) but have competence structured according to the major identified functional groups (as explained below).

The following core CF-DS competence and skills groups have been identified (refer to CF-DS specification [2] for details):

- Data Science Analytics (including Statistical Analysis, Machine Learning, Data Mining, Business Analytics, others) (DSDA)
- Data Science Engineering (including Software and Applications Engineering, Data Warehousing, Big Data Infrastructure and Tools) (DSENG)
- Data Management and Governance (including data stewardship, curation, and preservation) (DSDM)
- Research Methods and Project Methods (DSRMP)
- Domain Knowledge and Expertise (Subject/Scientific domain related)

Data Science competences must be supported by knowledge that are defined primarily by education and training and skills that are defined by work experience correspondingly. The CF-DS defines two types of skills:

- Skills Type A which are related to the professional experience and major competences, and
- Skills Type B that are related to wide range of practical computational skills including using programming languages, development environment and cloud based platforms (refer to CF-DS [2] for full definition of the identified knowledge and skills groups).

The DS-BoK defines the Knowledge Areas (KA) for building Data Science curricula that are required to support identified Data Science competences. DS-BoK is organised by Knowledge Area Groups (KAG) that correspond to the CF-DS competence groups. DS-BoK is based on ACM/IEEE Classification Computer Science (CCS2012) [16], incorporates best practices in defining domain specific BoK's and provides reference to existing related BoK's. It also includes proposed new KA to incorporate new technologies and scientific subjects required for consistent Data Science education and training.

The MC-DS [4] is built based on DS-BoK and linked to CF-DS where Learning Outcomes are defined based on CF-DS competences (specifically skills type A), and Learning Units are mapped to Knowledge Units in DS-BoK. Three mastery (or proficiency) levels are defined for each Learning Outcome to allow for flexible curricula development and profiling for different Data Science professional profiles. Practical curriculum should be supported by corresponding educational environment for hands on labs and educational projects development.

The formal DS-BoK and MC-DS definition will create a basis for Data Science educational and training programmes compatibility and consequently Data Science related competences and skills transferability.

IV. EDSF Data Model and Curriculum Design

Currently EDSF toolkit contains a number of datasets representing different components of the EDSF and mapping between them. Future EDSF development will formally define the ontologies related to the EDSF components and related dictionaries.

Figure 2 illustrates the relation between different data sets and ontologies comprising EDSF. The CF-DS is structured along four dimensions (similarly to European e-Competence Framework e-CFv3.0 [15]) that include (1) competence groups, (2) individual competences definition, (3) proficiency levels, and (4) corresponding knowledge and skills. In this context, each individual competence includes a set of required knowledge topics and a set of skills type A and skills type B. Such CF-DS structure allows for competence based curriculum design where competences can be defined based on the professional profile (see DSPP [5] for mapping between professional profiles and competences) or target leaners group when designing a full curriculum, or based on competence benchmarking for tailored training to address identified competences and knowledge gaps.

When a set of required competences is defined together with the required ranking or proficiency level, the set of required knowledge topics can be extracted and ordered according to proficiency level and relevance (or benchmark score) for further mapping to DS-BoK Knowledge Areas and Knowledge Units. The set of KAs and KUs defined for a specific competence set define the structure of the curriculum

that further can be mapped to Model Curriculum Learning Units defined as individual courses and KAG related courses groups.

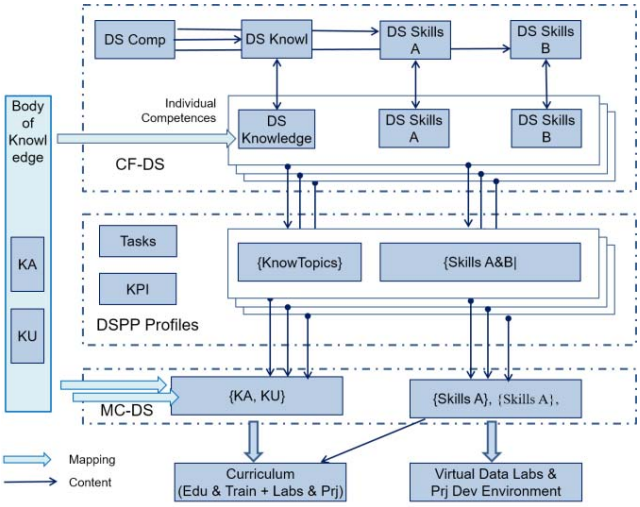

Figure 2. EDSF Data Model and customised curriculum design for target professional group(s)

At the same time, required proficiency level is scored for each KA and KU, which will define a mastery levels and corresponding learning outcome for the targeted education or training. The following mastery levels are defined (using workplace terminology that can be easy mapped to mastery levels defined in MC-DS):

A - Awareness
1) Understand Terminology
2) Understand Principles
3) Apply principles
4) Understand Methods

U - Use/Application
5) Apply basics
6) Supervised use
7) Unsupervised Use

P - Professional/Expert
8) Development of applications using wide range of technologies
9) Supervise project development, team of professionals,

where borderline mastery levels 4 and 7 actually belong to both higher level and lower level groups.

Set of Skills type A will define Learning Outcomes and Skills type B will provide advice on the required hands on training and practical project development environment and platform. As an example, the Data Scientist curriculum should include the following elements to achieve necessary skills Type B:

- Python (or R) and corresponding data analytics libraries
- NoSQL and SQL Databases (Hbase, MongoDB, Cassandra, Redis, Accumulo, MS SQL, My SQL, PostgreSQL, etc.)
- Big Data Analytics platforms (Hadoop, Spark, Data Lakes, others)
- Real time and streaming analytics systems (Flume, Kafka, Storm)
- Kaggle competition, resources and community platform
- Visualisation software (D3.js, Processing, Tableau, Julia, Raphael, etc.)
- API management and web scrapping
- Git versioning system as a general platform for software development
- Development Frameworks: Python, Java or C/C++, AJAX (Asynchronous Javascript and XML), D3.js (Data-Driven Documents), jQuery, others
- Cloud based Big Data and data analytics platforms and services, including large scale storage systems.

V. Cloud based DSEE and Virtual Data Labs

Using cloud resources to build effective and up-to-date professional Data Science education environment is inevitable with current fast technology development and required computational performance that can be requested on-demand.

Major Cloud Service Providers (CSP) provide wide range of data analytics and business analytics services and platforms that can be equally used by big, small and medium companies and individuals on the pay-per-use basis. In addition to a possibility to use same resources for education and training purposes, the major CSPs provide also designated education and self-training resources that are in many cases supported also educational grants for students and teachers.

The following cloud based resources from the major cloud providers can be used to build hybrid DSEE and VDLabs (in addition to regular compute and storage resources):

- Microsoft Azure Data Lakes Analytics, Power BI, HDInsight Hadoop as a Service, others
- AWS Elastic MapReduce (EMR), QuickSight, Kinesis and wide collection of open datasets
- IBM Data Science Experience, Data Labs, Watson Analytics.

Important component of Data Science education is educational datasets that often need to be provided with their specific applications. While many educational datasets are available from mentioned above cloud platforms, from community run Kaggle (https://www.kaggle.com/) and UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/index.php), use of cloud based VDLabs allows to instantiate the whole experimental setup or environment together with used data sets in case of specific domain focused education or training.

VI. SlipStream: Cloud Application Management Platform

The VDLab model allows using practical application instantiation from research or production environment and

using them for education or training purposes. In many cases it will also allow guidance support from the organisation that donated such VDLab environment.

The authors are exploring the possibility of using CYCLONE cloud automation framework to design the VDLabs and manage their complete lifecycle when used on-demands with the periodically running education or training courses. Interaction and interfacing with the potentially multicloud resources and cloud providers is enabled via using SlipStream, an open source cloud application management platform [17].[1] Through its plugin architecture, SlipStream supports most major cloud service providers and the primary open source cloud distributions. By exposing a uniform interface that hides differences between cloud providers, SlipStream facilitates application portability across the supported cloud infrastructures.

To take advantage of cloud portability, developers define "recipes" that transform available "base" virtual machines into the components that they need for their application. By reusing these base virtual machines, developers can ensure uniform behavior of their application components across clouds without having to deal with the time-consuming and error-prone transformation of virtual machine images. Developers bundle the defined components into complete cloud applications using SlipStream facilities for passing information between components and for coordinating the configuration of services.

Once a cloud application has been defined, the operator can deploy the application in "one click", providing values for any defined parameters and choosing the cloud infrastructure to use. With SlipStream, operators may choose to deploy the components of an application in multiple clouds, for example, to provide geographic redundancy or to minimize latencies for clients. To respond to changes in load, operators may adjust the resources allocated to a running application by scaling the application horizontally (changing the number of virtual machines) or vertically (changing the resources of a virtual machine).

SlipStream combines its deployment engine with an "App Store" for sharing application definitions with other users and a "Service Catalog" for finding appropriate cloud service offers, providing a complete engineering PaaS supporting DevOps processes [18]. All of the features are available through its web interface or RESTful API.

A. Functionality used for VDLab/applications deployment

The definition of an application component actually consists of a series of recipes that are executed at various stages in the lifecycle of the application. The main recipes, in order, are:

- **Pre-install**: Used principally to configure and initialize the operating system's package management.
- **Install packages**: A list of packages to be installed on the machine. SlipStream supports the package managers for the RedHat and Debian families of OS.
- **Post-install**: Can be used for any software installation that cannot be handled through the package manager.
- **Deployment**: Used for service configuration and initialization. This script can take advantage of SlipStream's "parameter database" to pass information between components and to synchronize the configuration of the components.
- **Reporting**: Collects files (typically log files) that should be collected at the end of the deployment and made available through SlipStream.

There are also a number of recipes that can be defined to support horizontal and vertical scaling that are not used in the defined here use cases. The applications are defined using SlipStream's web interface, the bioinformatics portal then triggers the deployment of these applications using the SlipStream RESTful API.

B. Example recipes for bacterial genomics Data Lab

The section provides example of an application for the bacterial genomics analysis consisted of a compute cluster based on Sun Grid Engine with an NFS file system exported from the master node of the cluster to all of the slave nodes. The master node definition was combined into a single "deployment" script that performed the following actions:

1) Initialize the yum package manager.
2) Install bind utilities.
3) Allow SSH access to the master from the slaves.
4) Collect IP addresses for batch system.
5) Configure batch system admin user.
6) Export NFS file systems to slaves.
7) Configure batch system.
8) Indicate that cluster is ready for use.

The deployment script extensively uses the parameter database that SlipStream maintains for each application to correctly the configure the master and slaves within the cluster. A common pattern is the following:

```
ss-display "Exporting SGE_ROOT_DIR..."
echo -ne "$SGE_ROOT_DIR\t" > $EXPORTS_FILE
for ((i=1; i<=`ss-get
        Bacterial_Genomics_Slave:multiplicity`;
i++ ));
do
     node_host=`ss-get
        Bacterial_Genomics_Slave.$i:hostname`
     echo -ne $node_host >> $EXPORTS_FILE
     echo    -ne    "(rw,sync,no_root_squash)    ">>
$EXPORTS_FILE
done
echo "\n" >> $EXPORTS_FILE # last for a newline
exportfs -av
```

[1] Community Edition of SlipStream, is available under the Apache 2.0 license (https://github.com/slipstream)

VII. Conclusion and Further Developments

EDSF provides a common semantic basis for interoperability of all forms of the Data Science curriculum definition and education or training delivery, as well as knowledge assessment based on fully enumerated definition of EDSF components and individual units. Besides defining academic components of the effective and consistent curriculum, EDSF provides also advice on the required Data Science Education Environment to facilitate fast practical knowledge and skills acquisition by students and learners.

The paper suggests using widely available and rich cloud based resources such as data analytics platforms and services, data sets and general large scale storage solutions, practical knowledge of which will significantly shorten the future graduates' integration into their workspace environment. The authors are in the process of the practical implementation of the hybrid solutions for Virtual Data Labs that can combine local campus resources and cloud based resources both provider based and openly available. The paper also motivates adoption of FAIR principles for the future Open Education that should address new needs of the emerging data driven digital economy.

The EDSF and the proposed in this paper its further integration with the Data Science Education Environment will facilitate education and training for highly demanded Data Science and Analytics competences and skills.

Acknowledgment

The research leading to these results has received funding from the Horizon2020 projects CYCLONE (funded by the European Commission under grant number 644925) and EDISON (funded under grant n. 675419.

References

[1] EDISON Data Science Framework (EDSF). Available at http://edison-project.eu/edison/edison-data-science-framework-edsf

[2] Data Science Competence Framework. Available at http://edison-project.eu/data-science-competence-framework-cf-ds

[3] Data Science Body of Knowledge. Available at http://edison-project.eu/data-science-body-knowledge-ds-bok

[4] Data Science Model Curriculum. Available at http://edison-project.eu/data-science-model-curriculum-mc-ds

[5] Data Science Professional Profiles. Available at http://edison-project.eu/data-science-professional-profiles

[6] Demchenko, Yuri, David Bernstein, Adam Belloum, Ana Oprescu, Tomasz W. Wlodarczyk, Cees de Laat, New Instructional Models for Building Effective Curricula on Cloud Computing Technologies and Engineering. Proc. Workshop "Requirements Engineering for Cloud Computing (RECC)", in conjunction with The 5th IEEE International Conference and Workshops on Cloud Computing Technology and Science (CloudCom2013), 2-5 December 2013, Bristol, UK.

[7] Demchenko, Yuri, Emanuel Gruengard, Sander Klous, Instructional Model for Building effective Big Data Curricula for Online and Campus Education. 1st IEEE STC CC and RDA Workshop on Curricula and Teaching Methods in Cloud Computing, Big Data, and Data Science, in Proc.The 6th IEEE International Conference and Workshops on Cloud Computing Technology and Science (CloudCom2014), 15-18 December 2014, Singapore.

[8] Manieri, Andrea 2015, et al, Data Science Professional uncovered: How the EDISON Project will contribute to a widely accepted profile for Data Scientists, Proc.The 7th IEEE International Conference and Workshops on Cloud Computing Technology and Science (CloudCom2015), 30 November - 3 December 2015, Vancouver, Canada

[9] Yuri Demchenko, Adam Belloum, Wouter Los, Tomasz Wiktorski, Andrea Manieri, Steve Brewer, Holger Brocks, Jana Becker, Dominic Heutelbeck, Matthias Hemmje, EDISON Data Science Framework: A Foundation for Building Data Science Profession For Research and Industry, 3rd IEEE STC CC and RDA Workshop on Curricula and Teaching Methods in Cloud Computing, Big Data, and Data Science (DTW2016), in Proc.The 8th IEEE International Conference and Workshops on Cloud Computing Technology and Science (CloudCom2016), 12-15 December 2016, Luxembourg.

[10] Yuri Demchenko, Miroslav Zivkovic, Cees de Laat, José Ignacio Aznar Baranda, Christophe Blanchet, Mohamed Bedri, Jean-François Gibrat, Oleg Lodygensky, Mathias Slawick, Ilke Zilci, Rob Branchat, Charles Loomis, CYCLONE: A Platform for Data Intensive Scientific Applications in Heterogeneous Multi-cloud/Multi-provider Environment, Fifth IEEE International Workshop on Cloud Computing Interclouds, Multiclouds, Federations, and Interoperability (Intercloud 2016), In Proc. IEEE International Conference on Cloud Engineering (IC2E), April 4 - 8, 2016, Berlin, Germany

[11] Final results of the European Data Market study measuring the size and trends of the EU data economy, EC-IDC, March 2017 [online] https://ec.europa.eu/digital-single-market/en/news/final-results-european-data-market-study-measuring-size-and-trends-eu-data-economy

[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] Barend Mons, et al, The FAIR Guiding Principles for scientific data management and stewardship [online] https://www.nature.com/articles/sdata201618

[15] e-CF3.0, 2016 European e-Competence Framework 3.0. A common European Framework for ICT Professionals in all industry sectors. CWA 16234:2014 Part 1. Available at http://ecompetences.eu/wp-content/uploads/2014/02/European-e-Competence-Framework-3.0_CEN_CWA_16234-1_2014.pdf

[16] CCS, 2012 The 2012 ACM Computing Classification System. Available at http://www.acm.org/about/class/class/2012

[17] SlipStream Cloud Automation [online] http://sixsq.com/products/slipstream/

[18] Davis, Jennifer; Daniels, Katherine (2015). Effective DevOps. O'Reilly. ISBN 978-1-4919-2630-7.