OPEN PEER REVIEW

Advancing the international data science workforce through shared training and education [version 1; peer review: 1 approved, 1 approved with reservations]

John Van Horn1,2, Sumiko Abe1,2, José Luis Ambite2,3, Teresa K. Atwood4, Niall Beard4, Louisa Bellis5,6, Avnish Bhattrai1,2, Alex Bui7, Gully Burns2,3, Lily Fierro2,3, Jonathan Gordon2,3, Jeffrey Grethe8, Jeana Kamdar1,2, Xiaoyu Lei1,2, Kristina Lerman2,3, Annette McGrath9,10, Nicola Mulder11, Caroline O’Driscoll1,2, Crystal Stewart1,2, Sonika Tyagi12

1USC Mark and Mary Stevens Neuroimaging and Informatics Institute, University of Southern California, Los Angeles, CA, 90033, USA
2BD2K Training Coordinating Center, University of Southern California, Los Angeles, CA, 90033, USA
3Information Sciences Institute, University of Southern California, Los Angeles, CA, 90292, USA
4Department of Bioinformatics, University of Manchester, Manchester, M13 9PL, UK
5CRUK Cambridge Centre Dept of Oncology, University of Cambridge, Cambridge, CB2 0XZ, UK
6European Bioinformatics Institute (EBI), Wellcome Genome Campus, European Molecular Biology Laboratory (EMBL), Hinxton, CB10 1SD, UK
7BD2K Centers Coordinating Center, University of California Los Angeles, Los Angeles, CA, 90095, USA
8BIOCADDIE, Center for Research in Biological Systems, University of California San Diego, La Jolla, CA, 92093, USA
9EMBL-ABR: Monash Node, Monash University, Melbourne, VIC, 3010, Australia
10Life Science Informatics DATA61, CSIRO, Dutton Park, QLD, 4102, Australia
11Human Heredity and Health in Africa (H3Africa) Initiative, Integrative Biomedical Sciences Faculty of Health Sciences, University of Cape Town, Observatory 7925, Cape Town, South Africa
12School of Biological Sciences and Monash eResearch Centre, Monash University, Clayton, VIC, 3800, Australia

Abstract
The increasing richness and diversity of biomedical data types creates major organizational and analytical impediments to rapid translational impact in the context of training and education. As biomedical data-sets increase in size, variety and complexity, they challenge conventional methods for sharing, managing and analyzing those data. In May 2017, we convened a two-day meeting between the BD2K Training Coordinating Center (TCC), ELIXIR Training/TeSS, GOBLET, H3ABioNet, EMBL-ABR, bioCADDIE and the CSIRO, in Huntington Beach, California, to compare and contrast our respective activities, and how these might be leveraged for wider impact on an international scale. Discussions focused on the role of i) training for biomedical data science; ii) the need to promote core competencies, and the ii) development of career paths. These led to specific conversations about i) the values of standardizing and sharing data science training resources; ii) challenges in encouraging adoption of

Open Peer Review

Reviewer Status

Invited Reviewers

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>report</td>
</tr>
<tr>
<td>2</td>
<td>report</td>
</tr>
</tbody>
</table>

version 1
published 04 Mar 2019

1 Ariel Rokem, University of Washington, Seattle, USA
2 David N. Kennedy, University of Massachusetts Medical School (UMMS),
training material standards; iii) strategies and best practices for the personalization and customization of learning experiences; iv) processes of identifying stakeholders and determining how they should be accommodated; and v) discussions of joint partnerships to lead the world on data science training in ways that benefit all stakeholders. Generally, international cooperation was viewed as essential for accommodating the widest possible participation in the modern bioscience enterprise, providing skills in a truly “FAIR” manner, addressing the importance of data science understanding worldwide. Several recommendations for the exchange of educational frameworks are made, along with potential sources for support, and plans for further cooperative efforts are presented.

**Keywords**
biomedicine, data science, training, education, personalized, standards, collaboration, global partnerships

This article is included in the Bioinformatics Education and Training Collection collection.

Corresponding author: John Van Horn (jvanhorn@usc.edu)

**Author roles:** Van Horn J: Conceptualization, Funding Acquisition, Project Administration, Writing – Original Draft Preparation, Writing – Review & Editing; Abe S: Writing – Original Draft Preparation, Writing – Review & Editing; Ambite JL: Conceptualization, Writing – Original Draft Preparation, Writing – Review & Editing; Attwood TK: Conceptualization, Writing – Original Draft Preparation, Writing – Review & Editing; Beard N: Conceptualization, Writing – Original Draft Preparation, Writing – Review & Editing; Bellis L: Conceptualization, Writing – Original Draft Preparation, Writing – Review & Editing; Bhattrai A: Data Curation, Writing – Original Draft Preparation, Writing – Review & Editing; Bul A: Writing – Original Draft Preparation, Writing – Review & Editing; Burns G: Conceptualization, Methodology, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; Gordon J: Data Curation, Writing – Original Draft Preparation, Writing – Review & Editing; Grethe J: Writing – Original Draft Preparation, Writing – Review & Editing; Kamdar J: Conceptualization, Project Administration, Writing – Original Draft Preparation, Writing – Review & Editing; Lei X: Data Curation, Writing – Original Draft Preparation, Writing – Review & Editing; Lerman K: Writing – Original Draft Preparation, Writing – Review & Editing; McGrath A: Writing – Original Draft Preparation, Writing – Review & Editing; Mulder N: Writing – Original Draft Preparation, Writing – Review & Editing; O’Driscoll C: Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; Stewart C: Conceptualization, Project Administration, Writing – Original Draft Preparation, Writing – Review & Editing; Tyagi S: Writing – Original Draft Preparation, Writing – Review & Editing

**Competing interests:** No competing interests were disclosed.

**Grant information:** This work was supported in part by National Institutes of Health (NIH) grant U24 ES026465 02, part of the NIH Big Data to Knowledge (BD2K) initiative. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Copyright:** © 2019 Van Horn J et al. This is an open access article distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. Data associated with the article are available under the terms of the Creative Commons Zero “No rights reserved” data waiver (CC0 1.0 Public domain dedication).

**How to cite this article:** Van Horn J, Abe S, Ambite JL et al. Advancing the international data science workforce through shared training and education [version 1; peer review: 1 approved, 1 approved with reservations] F1000Research 2019, 8:251 (https://doi.org/10.12688/f1000research.18357.1)

**First published:** 04 Mar 2019, 8:251 (https://doi.org/10.12688/f1000research.18357.1)
1. Introduction
The ability to reap the wealth of information contained in large-scale biomedical data holds the promise to further our understanding of human health and disease. Biomedical big data come from many sources, from massive stand-alone data-sets generated by large collaborations to small data-sets produced by individual investigators. The value of all these data can be amplified through aggregation and integration. However, the diversity of such data types creates major organizational and analytical impediments to rapid translational impact. As biomedical data-sets increase in size, variety and complexity, they challenge conventional methods for sharing, managing and analyzing those data. Furthermore, researchers’ abilities to capitalize on approaches based on biomedical data science are limited by poor data accessibility and interoperability, the lack of appropriate tools, and insufficient training (Brazas et al., 2017).

In response to the opportunities and challenges presented by the era of “big data” in biological and biomedical research, the National Institutes of Health (NIH) launched the Big Data to Knowledge (BD2K) initiative in 2015 (Van Horn, 2016). Earlier, in 2006, the European Strategy Forum for Research Infrastructures (ESFRI) published its first European Roadmap for Research Infrastructures. As part of this initiative, a 5-year pan-European preparatory-phase project was launched in 2007 to create a European Life-Science Infrastructure for Biological Information (ELIXIR); ELIXIR entered its operational phase in 2013, tasked to construct a world-class, globally-positioned European infrastructure for the management and integration of life-sciences information. Recognizing the need for training to support the development and use of the bioinformatics tools and databases disseminated by such infrastructures, the Global Organisation for Bioinformatics Learning, Education and Training (GOBLET) was established in 2012, to cultivate the global bioinformatics trainer community, set standards, and provide high-quality training resources; as part of its mission to professionalize training, GOBLET worked closely with ELIXIR to develop a joint training strategy in 2015 (Attwood et al., 2015). The H3Africa initiative, funded by the National Institutes of Health and Wellcome Trust, was established in 2012, when the first projects were launched to investigate the genetic basis of diseases in Africa. H3ABioNet was established as a pan-African bioinformatics network for H3Africa to build capacity for genomics research on the continent. Similarly, a range of programs in Australia, such as EMBL-ABR (Schneider et al., 2017) and BPA/CSIRO training network, were also established. Collectively, these programs seek to nurture the training elements of the digital research enterprise within biomedicine, to facilitate “discovery science”, to support the identification of new knowledge, and to capitalize on community engagement.

In May 2017, during a two-day meeting between the BD2K Training Coordinating Center (TCC), ELIXIR Training/TeSS, GOBLET, H3ABioNet, EMBL-ABR, bioCADDIE and the CSIRO, convened in Huntington Beach, California, to describe our respective activities and how these might be leveraged for wider impact on a global scale. Subsequent interactions between participants at international events associated with each group have reinforced a focus on the following questions:

- What is the role of training for biomedical data science, promoting core competencies and the development of career paths?
- Is there value in standardizing and sharing data science training resources?
- What challenges exist in encouraging adoption of training material standards?
- How should resources be personalized or customized to the individual, and what ways would work best?
- Who are the stakeholders in this enterprise, and how should they be accommodated?
- How can we jointly partner to lead the world on data science training in ways that are seen as positive from the point of view of the stakeholders?

In what follows, we summarize the themes of our discussions, each of our individual efforts, and driving questions important for our interactions. We propose several key processes to advance international best practices in training standards and programs.

2. Training in biomedical data science
Education and training are fundamental components of many major data science initiatives worldwide. For example, the BD2K has funded training activities that include a variety of grant mechanisms: data science educational courses and Open-Educational-Resource (OER) grants, data science training programs, young researcher career development awards, and training components of major BD2K research centers. This arguably represents the largest commitment ever undertaken by the NIH toward data science education. Their intention to achieve a new, forward-thinking goal of broader data science for biomedical will require: 1) bringing together individuals involved in training and career-development programs to network and share experiences; 2) engaging individuals developing OERs in meaningful dialogue to ensure that the resources they develop will be freely available and easy to access by the broader biomedical community; and 3) developing an online educational resource index for the creation of personalized learning paths through linked educational resources.

With this in mind, the BD2K TCC has been established (Bui et al., 2017) to pursue the following: 1) coordinate activities across the BD2K Training Consortium to enable the exchange of ideas and best practices in data science training, both within BD2K and in the broader biomedical research community; 2) facilitate the discovery, access and citation of educational resources through the development of a living educational resource discovery index (ERuDIce) (Ambite et al., 2017; Van Horn et al., 2018); 3) personalize the discovery of biomedical data science educational resources; 4) facilitate outreach and engagement with the data science training community to identify and
hold relevant workshops; 5) support biomedical research training collaborations through short-term rotations into biomedical data science labs; and 6) evaluate and summarize these supported activities. In addition to sponsoring and coordinating meetings, workshops and rotations in biomedical data science, the TCC seeks to develop technology and a user interface for the BD2K’s Training Consortium to communicate activity, and, for the broader biomedical community, to provide personalized access to educational resources.

ELIXIR unites Europe’s leading life-science organizations in managing and safeguarding the increasing volumes of data being generated by publicly-funded research. It coordinates, integrates and sustains bioinformatics resources across its member states, and enables users in academia and industry to access services that are vital for their research. To inform researchers about associated bioinformatics training courses and materials, and where in Europe (and the rest of the world) they can be found, ELIXIR has developed a Training eSupport System (TeSS; Beard et al., 2016; Larcombe et al., 2017). ELIXIR also develops and deploys training (both face-to-face and via e-learning approaches) to researchers, developers, infrastructure operators, and to trainers themselves (Morgan et al., 2017); the training focuses on topics identified as knowledge or skill gaps, whether in specific ELIXIR Use Cases or within the general ELIXIR community. This training program aims to both transform scientists into effective users of ELIXIR’s data, tools, standards and compute infrastructure, and to increase the number of capable data science trainers.

GOBLET (Attwood et al., 2015) was established as a Dutch Foundation during ELIXIR’s preparatory phase. It was created as an umbrella organization (incorporating international and national networks and societies, research and academic groups, etc.) to unite, inspire and equip bioinformatics trainers worldwide – ultimately, to serve as a professional body for bioinformatics training. To this end, GOBLET aims to provide a global, sustainable support infrastructure for the international community of bioinformatics¹ trainers (which includes teachers in high-schools) and trainees. Part of the Foundation’s work has involved developing a portal (Corpas et al., 2015) for sharing training materials and publications (including guidelines and best practice documents (Vía et al., 2013)), and resources to help train trainers. The portal is now one of the core content providers for ELIXIR’s TeSS. Alongside ELIXIR and other organizations, GOBLET also monitors bioinformatics skill gaps and training needs, and plays a role in advocacy for the introduction of bioinformatics training into core life-science curricula (Attwood et al., 2017; Brazas et al., 2017). GOBLET has a particular interest in developing standards, both to help drive up the quality of training materials and courses, and to facilitate their discovery. In this latter context, GOBLET works closely with ELIXIR and BioSchemas to harmonize metadata standards and avoid duplication of effort.

¹Bioinformatics is used here in its broadest sense to encompass aspects of computational biology, biocomputing, biocuration and data science.

The vision of H3Africa has been to create and support a pancontinental network of laboratories that will be equipped to apply leading-edge research to the study of the complex interplay between environmental and genetic factors that determine disease susceptibility and drug responses in African populations. Data generated from their efforts should inform strategies to tackle health inequity and ultimately lead to health benefits in Africa. To achieve this, the following issues are being addressed: 1) ensuring access to, and education on, relevant genomic technologies for African scientists; 2) facilitating integration between genomic and clinical studies; 3) facilitating training at all levels, and, in particular, training research leaders; and 4) establishing necessary research infrastructure and providing training in its use. H3ABioNet, the bioinformatics network for H3Africa, has led an extensive and diverse training program to build skills in bioinformatics and genomic data analysis. H3ABioNet uses a variety of training approaches, including face-to-face and distance-based courses to teach a variety of audiences.

EMBL Australia Bioinformatics Resource (EMBL-ABR) is a developing national research infrastructure, providing bioinformatics resources and support to life science and biomedical researchers in Australia. As a collaboration with the European Bioinformatics Institute (EMBL-EBI), it was established to maximize Australia’s bioinformatics capability. This close partnership is made possible by Australia’s associate membership of EMBL.

EMBL-ABR aims to: 1) increase Australia’s capacity to collect, integrate, analyze, share and archive the large heterogeneous data-sets that are now part of modern life science research; 2) contribute to the development of, and provide training in, data, tools and platforms to enable Australia’s life science researchers to undertake research in the age of big data; 3) showcase Australian research and data-sets at an international level; 4) enable engagement in international programs that create, deploy and develop best practice approaches to data management, software tools and methods, computational platforms and bioinformatics services. EMBL-ABR is a truly national resource that consists of a Hub and eleven nodes, which are organized around six key areas: Data, Tools, Compute, Standards, Training and Platforms, mapped to their respective expertise in terms of bioscience domains. EMBL-ABR is a member of GOBLET and has contributed to the development of Bioschemas, and works closely with the EBI and other ELIXIR Nodes to exchange and share resources (e.g., using the TeSS portal and Bio.tools utilities). EMBL-ABR uses these international connections to 1) provide Hub workshops on crucial topics not covered elsewhere in national bioinformatics training, 2) bring expertise to Australia through visits by international experts and improved access to training resources, and 3) help coordinate and disseminate node end-user training activities.

The bioCADDIE team has worked to develop a Data Discovery Index (DDI) prototype which, like the TCC’s ERuDite, indexes primary research data that are stored elsewhere. The DDI seeks to play an important role in promoting data integration through
the adoption of content standards and alignment to common data elements and high-level schema. bioCADDIE intends to provide the means to test the utility of these standards against its DDI, thus serving as an incubator for spurring the types of quality metrics that are currently being developed around article metrics, including citation analysis and other metrics of resource utilization.

Finally, the Commonwealth Scientific and Industrial Research Organization (CSIRO) is the federal government agency for scientific research in Australia. Its chief role is to benefit the community by improving the economic and social performance of industry. Through a diverse set of flagship programs, the CSIRO works with leading organizations around the world, and, with CSIRO Publishing, issues journals with the latest research by leading scientists. Data science education and training play a central role in many CSIRO activities and, since 2012, has forged a very successful training partnership with BioPlatforms Australia (BPA) to deliver national bioinformatics training in the life sciences.

By joining forces in creating and adopting training metadata standards, these distinct initiatives would help reach critical mass to expedite standards adoption, and would avoid the risk of having “yet another, but not quite compatible” standard. One such beneficiary of these resources, and the standards that underpin them, is H3Africa, who are ideally positioned to reach out to a broad spectrum of international scientists requiring advanced training in the most recent scientific methods and approaches.

We seek to adopt a collaborative approach to enhance mechanisms for creating a tight coupling between life-science/biomedical training resources and their underlying tools and data-sets, for scientific audiences throughout the US, Europe, Australia and Africa. This work will permit the deployment and use of data science methods in biomedical applications to keep up with the overwhelming pace of methodological development in data science.

Building and sharing virtual social environments for collaboration enables massive-scale analysis of primary bioscience data. Exploration of the means for interoperability between such systems will be an important point of discussions. A collaborative approach will enhance training workflow support and promote development of a “training workflow bazaar” – a social environment for personalized access to publicly available biomedical and bioinformatics educational resources.

Finally, we seek to utilize a synergistic approach across these efforts for defining the basic collaborative building blocks and using them in our respective training platforms. This will allow us to reduce duplication of efforts, to produce compatible, harmonized standards, and to prevent loss of time in reinventing code, thus alleviating the risk of having to build bridges later on and make (likely imperfect) mappings between incompatible infrastructures.

3. International interactions

Bringing together experts from training consortia across four continents (North America, Europe, Australia and Africa) to discuss, share and plan for enhanced collaboration, the exchange of knowledge, and the development of further joint activities, has many advantages. These alliances may result in a set of guiding principles and shared standards that will form a solid basis for interoperability and exchange of large-scale data science training across these major international efforts. Consequently, such interactions of the TCC with these international training experts will be of direct benefit to the overall BD2K community, as well as investigators seeking a broad basis for training in large-scale biomedicine. These will be of particular importance to training and education for the BD2K community, as well as the larger biomedical data science community.

4. Assessing the FAIRness of ELIXIR TeSS and BD2K TCC ERuDItc

To facilitate and encourage the free exchange of learning resources between training organizations, we assessed whether ELIXIR’s and the TCC’s respective international training web portals complied with the FAIR principles. For data to be classed as being ‘FAIR’, it must comply with the guiding principles of being: Findable, Accessible, Interoperable and Reusable. We specify our aligned efforts with each principle (as noted on The FAIR Data principles) below. In the sections that follow we assess TeSS, the web portal of ELIXIR, and ERuDItc, the index that powers the TCC web portal.

To be Findable:
- F1. (meta)data are assigned a globally unique and eternally persistent identifier.
  - Both TeSS and ERuDItc assign URIs to resource pages, as well as collect and store external links and references.
- F2. data are described with rich metadata.
  - TeSS and ERuDItc use Schema.org schemas for core resource metadata (which are converging). As TeSS uses the EDAM ontology and ERuDItc uses the DSEO ontology to assign concept/tags to learning resources, it was necessary to begin to map the ontology terms to each other.
- F3. (meta)data are registered or indexed in a searchable resource.
  - TeSS and ERuDItc provide search engines for accessing resource metadata.
- F4. metadata specify the data identifier.
  - TeSS and ERuDItc link to the original learning resource URI.

To be Accessible:
- A1. (meta)data are retrievable by their identifier using a standardized communications protocol.
Both portals are being developed to supplement existing training networks. These communities include experienced trainers who are familiar with projects, other trainers and resources in a wide variety of disciplines. We leverage this expertise to identify and evaluate new high-quality resources.

A. **Expert domain knowledge** – Both portals are being developed to supplement existing training networks. These communities include experienced trainers who are familiar with projects, other trainers and resources in a wide variety of disciplines. We leverage this expertise to identify and evaluate new high-quality resources.

B. **Provider-supplied resources** – As the user-bases and reputations of the portals grow within their target communities, trainers wishing to raise the exposure of their work can notify a portal that they want to be included; or, if the option is available, register their resources through a self-service interface. To encourage trainers to register their materials with the portals, the portals should ensure there are clear benefits to the trainers. One suggestion is to display usage metrics, such as the number of searches and clicks, and the amount of positive feedback a resource receives. Trainers can then use this information not only to demonstrate their impact to funders and show leadership within their research communities, but also to keep track of usage and popularity of the material.

C. **Open-web search** – We build upon the capabilities of existing search engines, such as Google, to cast a wide net for relevant training materials, which may be subsequently filtered to ensure resource quality allowing for refined searches. For example, the BD2K TCC executes about 100 queries for specific data science concepts on YouTube to identify relevant videos. These searches produce tens of thousands of videos, which are filtered by applying machine-learning classifiers trained to select those of high quality. A complementary approach is to use Google’s structured data API to identify web resources containing specific Schema.org classes and properties indicative of learning resources. The metadata of such resources can be obtained and added to the index.²

5. ELIXIR and TCC’s shared challenges and approaches to enriching experiences on learning web portals

In the following sections, we describe areas of development for TeSS and the TCC Portal to support online learning communities.

5.1 Identifying resources for indexing

Training portals, such as ELIXIR TeSS and the BD2K TCC Portal, powered by ERuDItc, aim to provide access to high-quality training resources that are available in a wide variety of online sources. Identifying such resources is a challenging problem, given that they are widely distributed and continuously evolving. We are following a three-pronged approach to resource identification:

- **A. Expert domain knowledge** – Both portals are being developed to supplement existing training networks. These communities include experienced trainers who are familiar with projects, other trainers and resources in a wide variety of disciplines. We leverage this expertise to identify and evaluate new high-quality resources.

- **B. Provider-supplied resources** – As the user-bases and reputations of the portals grow within their target communities, trainers wishing to raise the exposure of their work can notify a portal that they want to be included; or, if the option is available, register their resources through a self-service interface. To encourage trainers to register their materials with the portals, the portals should ensure there are clear benefits to the trainers. One suggestion is to display usage metrics, such as the number of searches and clicks, and the amount of positive feedback a resource receives. Trainers can then use this information not only to demonstrate their impact to funders and show leadership within their research communities, but also to keep track of usage and popularity of the material.

- **C. Open-web search** – We build upon the capabilities of existing search engines, such as Google, to cast a wide net for relevant training materials, which may be subsequently filtered to ensure resource quality allowing for refined searches. For example, the BD2K TCC executes about 100 queries for specific data science concepts on YouTube to identify relevant videos. These searches produce tens of thousands of videos, which are filtered by applying machine-learning classifiers trained to select those of high quality. A complementary approach is to use Google’s structured data API to identify web resources containing specific Schema.org classes and properties indicative of learning resources. The metadata of such resources can be obtained and added to the index.²

²A number of existing tools facilitate parsing Schema.org metadata from web pages: https://developers.google.com/custom-search/docs/tutorial/creatingse
5.2 Describing resources

Having rich, structured metadata annotations benefits multiple audiences that interact with resource aggregators, such as ELIXIR’s TeSS and the BD2K TCC’s ERuDite. First, with richer metadata, learners have more information to use when evaluating whether or not a resource suits their needs. Secondly, search engines can more easily find and index resources, thereby increasing discoverability by learners, resource creation organizations, and any resource aggregator. Finally, providing structured metadata facilitates data interchange between resource indexes. However, to maximize these benefits, common metadata standards are desirable.

In this workshop, we focused on aligning the metadata standards used by TeSS and ERuDite. TeSS (Beard et al., 2016) collects training events and materials related to the life sciences, with a special interest in bioinformatics, whilst ERuDite focuses on materials related to data science (Ambite et al., 2017). Both portals are interested in creating a common representation of online training materials, the organizations that create and host them, and the people who create and teach them.

Currently, Bioschemas.org has defined specifications on how to represent life science training materials, people, organizations and events, based on the CreativeWork, Person, Organization and Event classes of the prevalent and popular Schema.org standard. Given ELIXIR’s membership in Bioschemas.org, in order to facilitate data exchange between TeSS, ERuDite and other training sites and indexes, ERuDite has aligned its learning resource descriptions to a common standard built from Schema.org. TeSS and ERuDite have determined a core set of fields for our respective usage of the CreativeWork, Person and Organization classes, and we have documented how we use the properties of each class. These are the core properties and classes in ERuDite, with their respective namespaces. Properties are also inherited from superclasses:

- **schema:Thing** : dcterms:identifier, schema:name, schema:alternateName, schema:description, schema:url, schema:sameAs, schema:image
- **schema:Person** (schema:Thing) : schema:givenName, schema:additionalName, schema:familyName, schema:jobTitle, schema:memberOf, erudite:teaches
- **schema:Organization** (schema:Thing) : schema:location, schema:logo, schema:member, schema:memberOf, erudite:provides
- **schema:Course** (schema:CreativeWork) : schema:coursePrerequisites, erudite:syllabus, schema:startDate, schema:endDate, schema:instructor, lom:typicalLearningTime, xcri:contactHours, schema:grantsCredential

ERuDite is aligned with Bioschemas.org and Schema.org. ERuDite includes three additional properties not available in Schema.org.

TeSS and ERuDite provide concept tags to every material, but the two indexes use different tagging ontologies. TeSS uses EDAM (Ison, et al., 2013), which is designed to suit the specific coverage of life science, healthcare and bioinformatics events and materials. ERuDite uses the custom Data Science Education Ontology (DSEO), which was designed using both top-down and bottom-up approaches (Ambite et al., 2017). In a top-down fashion, the DSEO was first manually designed as an initial ontology based on expert knowledge. In the bottom-up approach, the ERuDite team automatically computed topic models from materials’ text to identify additional concepts. Finally, the team curated and organized all these concepts. We are creating a mapping between DSEO and EDAM, which will allow us to see which materials we are interested in and then tag the materials with the terms in our respective ontologies. While this tagging ontology mapping will emerge from the TeSS and ERuDite collaboration, it also opens up collaborations with any other institutions that use EDAM, such as GOBLET, H3ABioNet and EMBL-ABR, who also provide a rich set of training content suitable for indexing.

5.3 Helping users find resources

Currently, TeSS and ERuDite use similar approaches to enable resource discovery. Both portals have simple, intuitive search interfaces that allow free text searching, and also provide resource tags that facilitate search filtering. Although search interfaces handle the majority of resource-discovery needs, both TeSS and ERuDite hope to provide other meaningful ways to explore resources.

First, both portals are actively working on initiatives that allow learners to create personalized resource collections. While these collections are formed to serve the specific learning needs of their creators, they also serve the needs of learners who are interested in similar topics and learning paths. In addition to these learner-created collections, both TeSS and the TCC portal aim to introduce expert-level collections, which will contain recommended resources that learners can use as templates for their own learning. Given the growth and rapid evolution of the learning domains of both ERuDite and TeSS, having expert validated collections will help learners find a starting place for their learning journey through a topic or concept.

The BD2K TCC is also interested in providing interactive resource browsing interfaces. Currently, the TCC provides resource summarizations that guide learners to resources based on topics and providers of interest. In the future, the TCC aims to provide a meaningful resource-to-resource browsing interface where learners can see and select groups of similar resources. This visualization paradigm can then be shared with TeSS, if a browsing visualization is requested by its learning community.

Lastly, both the TCC and TeSS are interested in creating personalized recommendations for learners. Both sites are currently constrained by the size of their user bases, but as more user data are available, both portals will be able to provide resource-level suggestions that can be used on the portal directly or
through community outreach methods (e.g., email campaigns or lists).

5.4 Reusing training resources

Ultimately, one clear way to assess the success of the TeSS and TCC web portals is the volume of resource reuse and re-indexing by other organizations. However, this relies heavily on the relevance and quality of the resources presented on TeSS’s and TCC’s respective websites. Both organizations hope that quality and relevance assessment can be accomplished by two groups: online learners and portal developers.

Resource quality evaluations from portal developers would need to consist of procedures that could semi-automatically check for resource status (e.g., active link or dead link), and duplication and accuracy of collected metadata. Using appropriate standards and schemas, the version history of training resources, and their author and contributor lists, can also be tracked in a transparent fashion, giving the content developers due credit and visibility for their materials. Resource quality and relevance assessments from online learners will come from user input, which could be explicit (e.g., the frequency of a resource in learners’ collections), or explicit (e.g., ratings and surveys). Both implicit and explicit feedback from learners will be essential for improving user experiences and resource reuse on both portals.

Once the reputation of the high quality of resources in the respective portals has been established, others will be encouraged to reuse and share them. Ideally, reuse will be accomplished by extracting our metadata markups, which will then spread the learning resource standards we have specified, allowing greater visibility and community use of our representations.

6. Personalized learning

Through years of efforts, GOBLET, TeSS and ERuDite (GTE) have built data science ontologies – hierarchical concept schemas that define knowledge networks in bioinformatics. Additionally, GTE have collected a large number of publicly accessible educational resources associated with the concepts in these ontologies. Using these resources, GTE are developing online platforms that enable people from varied professional backgrounds and levels of proficiency to efficiently and conveniently gain access to data science training resources.

Personalized training programs can recommend resources that are better aligned with learners’ interests and strengths, and may increase their motivation. Therefore, we are interested in the creation of personalized learning environments that enable users to actively customize their education using a wide variety of online resources. Throughout this process, GTE would assist users in optimizing their learning experiences by providing individualized collections of resources that properly fit their goals.

This personalization requires the creation of a systematic process to generate a learner model that identifies competency levels and learning objectives, which can then be updated with progress towards those goals. Subsequently, based on the learner model, an automatic personalization process would plan an optimal route through concepts in the ontology to identify collections of resources appropriate for users’ personal skills and educational goals. The system should also be able to adjust and update its recommendations based on user feedback, to better serve users in accomplishing data science-related training goals.

6.1 Learner profiles, learner modality and recommendations

The more information a user provides to the platform, the more personalized their learning experience can be. Initially, users may be asked to fill out a learner profile survey, which should ask them to rate their current level of expertise with various concepts, their interest in learning more about those concepts or achieving broader training goals. Users who would prefer to explore resources for themselves rather than take a survey can still receive personalized experiences. They can incrementally complete a learner profile through explicit or implicit indicators: e.g., rating resources they are familiar with, or simply browsing resources related to particular topics. As learners use the platform, personalization is most obvious through resource recommendations: when looking at a resource, they may be recommended to follow it with another that continues an appropriate progression for their goals. They may also be presented with an entirely automatically generated learning plan, where the topics and the resources reflect their expertise and interests. In the following sections, we explore learner modeling and recommendations in more depth.

Mastery Rubric and Bloom’s Taxonomy. The Bloom Taxonomy describes six levels of cognitive skills, providing a description over a coded continuum of expertise that would be useful for resource developers. The levels are: 1) Remember/Reiterate, 2) Understand/Summarize, 3) Apply/Illustrate, 4) Analyze/Predict, 5) Evaluate/Synthesize, and (6) Create/Apply/Judge; these increase in cognitive complexity with each level (Bloom 1956). Mastery Rubric (MR) development provides a powerful methodology for developing curricula and assessments designed to teach key Knowledge, Skills and Abilities (KSAs) for specific disciplines (Tractenberg et al., 2010; Tractenberg et al., 2017). The MR formulation involves outlining how KSAs align with performance levels to provide a trajectory for students as they advance in a subject. MRs have been developed for Clinical Research (Tractenberg et al., 2017), Ethical Reasoning (Tractenberg et al., 2015), and Statistical Literacy and Evidence-Based Medicine (Tractenberg et al., 2016).

6.2 Modeling learners

6.2a High-level structure. There are many ways to model learners, but we will focus on the aspects that are important to enabling the automated personalization of learning experiences. The aspects we have identified are a) the learner, b) the concept-skill combinations comprising the domain of interest, and c) the available learning resources. We will describe these in this section.

6.2b A four-dimensional learner model. Perhaps the key question when considering how we might develop user-specific
capabilities for an individual within a learning portal is to think carefully about how we model that individual. We propose modeling users with a matrix that links subject domains and skills to users’ expertise and interests. This is shown diagrammatically below.

The example in the figure depicts the situation of a learner who wants to find out about the field of genetics but does not know where to start. It shows that the same learner is also an R programmer with existing expertise, but wants to learn Python to fulfill the requirements of a project. The system should recommend courses that meet these interests, whilst leveraging elements in which the learner already has expertise.

The key aspect of this underlying model is that the table shown above would need a large number of rows that codify the various different domains and skills possessed by a trained individual. Clearly, how we go about modeling each of these dimensions is crucial to the utility or effectiveness of this approach. We begin here to outline a first pass over the process of characterizing these dimensions, based on the various ontologies and conceptual schema being developed within the GTE community to date.

6.2c Domains. To model people’s learning goals and their progress toward them, it is important to have a model of the domain of interest. Domain models take various forms, ranging from ontologies that domain experts and knowledge engineers construct by hand at great cost, to methods that are less reliable but fast and scalable, such as automatic topic modeling using Latent Dirichlet location (LDA; Blei et al., 2003).

We do not prescribe a method for creating a domain model or a precise form, but there are several requirements for supporting personalized learning. The domain model must provide a list of discrete, named concepts. These allow the user model to represent the state of the learner’s knowledge about particular aspects of the domain, and those aspects that they have expressed interest in learning more about, either explicitly or implicitly: e.g., by browsing resources related to a particular concept. The model of the domain must also connect the identified concepts to the available resources with weighted links, giving the degree to which each resource pertains to a particular concept. Concepts can be connected to one another by various relations, including prerequisite, similarity and part-of. These relations are used to recommend concepts and their associated learning resources, described in Section 4.

6.2d Skills and abilities. Following the methodology defined in the development of Mastery Rubrics, we encode KSAs as a key element. To our knowledge, the development of taxonomies or ontologies for skills and abilities in biomedical data science is still in its infancy, but existing resources provide informal representations that support this notion. For example, Welch et al. (2016) provide a preliminary encoding of “core competencies” that broadly outline various general skill-sets required for bioinformatics.

Going beyond these general categories into more detail may require a finer-grained representation of specific skills and abilities. Potential starting points are provided by the EDAM ontology. This resource provides a taxonomy of operations that could act as a proxy in the specific area of bioinformatics-based data analysis. Similarly, the Data Science Education Ontology (DSEO) provides a taxonomy of data science processes that could be elaborated into a more complete description of skills and abilities that we could use here.

Beyond these resources, a starting set of high-level skill categories could include: simple comprehension, concept mastery, basic procedural execution, procedural troubleshooting, and procedural implementation. It may then be necessary to further specialize these abilities within domains, programming languages, data science processes, etc.

6.2e Knowledge/expertise. Here, we follow the encoding provided by the Bloom Taxonomy/Mastery Rubric methodology (Bloom, 1956; Tractenberg et al., 2010), which specifies expertise as an ordinal scale with the following categories (in ascending order of expertise): a) beginning, b) novice, c) competent, and d) proficient. Subsequent incarnations of the Mastery Rubric specify a trajectory from a) Novice, to b) Beginner, to c) Apprentice, to d) Journeyman.

6.2f Interest. Learners may have only a general idea of what they are interested in (e.g., “data science” or “bioinformatics”), or they may have specific interests (e.g., “using R to analyze genomic data”). Part of a functional learner model is to record which concepts – at different levels of granularity – a learner would like to learn more about. Wherever possible, this interest should be assigned to the particular skill: e.g., conceptual vs. procedural. Our model of learners’ interests is a mixture of what they tell us (e.g., by defining their goal), and what we infer (e.g., from identifying themes in learning resources they access).

6.2g Leveraging existing resources and implementation. ERuDiTe (Ambite et al., 2017; Van Horn et al., 2018) consists of an index of both manually and automatically identified data-science resources – predominantly online courses and videos – which are tagged with relevant concepts from DSEO. Ongoing work seeks to identify natural sequences of resources based on prerequisites, building on the architecture to do so from the TechKnAcq project (Gordon et al., 2016).

The EDAM ontology (Ison et al., 2013) provides an extensive representation of four branches: a) Data (with Data Identifiers) denoting types of information that may be processed computationally; b) Operations that may be applied to those data; c) the many Formats in which the data may be expressed; and d) the

---

6.2c Domains. To model people’s learning goals and their progress toward them, it is important to have a model of the domain of interest. Domain models take various forms, ranging from ontologies that domain experts and knowledge engineers construct by hand at great cost, to methods that are less reliable but fast and scalable, such as automatic topic modeling using Latent Dirichlet location (LDA; Blei et al., 2003).

We do not prescribe a method for creating a domain model or a precise form, but there are several requirements for supporting personalized learning. The domain model must provide a list of discrete, named concepts. These allow the user model to represent the state of the learner’s knowledge about particular aspects of the domain, and those aspects that they have expressed interest in learning more about, either explicitly or implicitly: e.g., by browsing resources related to a particular concept. The model of the domain must also connect the identified concepts to the available resources with weighted links, giving the degree to which each resource pertains to a particular concept. Concepts can be connected to one another by various relations, including prerequisite, similarity and part-of. These relations are used to recommend concepts and their associated learning resources, described in Section 4.

6.2d Skills and abilities. Following the methodology defined in the development of Mastery Rubrics, we encode KSAs as a key element. To our knowledge, the development of taxonomies or ontologies for skills and abilities in biomedical data science is still in its infancy, but existing resources provide informal representations that support this notion. For example, Welch et al. (2016) provide a preliminary encoding of “core competencies” that broadly outline various general skill-sets required for bioinformatics.

Going beyond these general categories into more detail may require a finer-grained representation of specific skills and abilities. Potential starting points are provided by the EDAM ontology. This resource provides a taxonomy of operations that could act as a proxy in the specific area of bioinformatics-based data analysis. Similarly, the Data Science Education Ontology (DSEO) provides a taxonomy of data science processes that could be elaborated into a more complete description of skills and abilities that we could use here.

Beyond these resources, a starting set of high-level skill categories could include: simple comprehension, concept mastery, basic procedural execution, procedural troubleshooting, and procedural implementation. It may then be necessary to further specialize these abilities within domains, programming languages, data science processes, etc.

6.2e Knowledge/expertise. Here, we follow the encoding provided by the Bloom Taxonomy/Mastery Rubric methodology (Bloom, 1956; Tractenberg et al., 2010), which specifies expertise as an ordinal scale with the following categories (in ascending order of expertise): a) beginning, b) novice, c) competent, and d) proficient. Subsequent incarnations of the Mastery Rubric specify a trajectory from a) Novice, to b) Beginner, to c) Apprentice, to d) Journeyman.

6.2f Interest. Learners may have only a general idea of what they are interested in (e.g., “data science” or “bioinformatics”), or they may have specific interests (e.g., “using R to analyze genomic data”). Part of a functional learner model is to record which concepts – at different levels of granularity – a learner would like to learn more about. Wherever possible, this interest should be assigned to the particular skill: e.g., conceptual vs. procedural. Our model of learners’ interests is a mixture of what they tell us (e.g., by defining their goal), and what we infer (e.g., from identifying themes in learning resources they access).

6.2g Leveraging existing resources and implementation. ERuDiTe (Ambite et al., 2017; Van Horn et al., 2018) consists of an index of both manually and automatically identified data-science resources – predominantly online courses and videos – which are tagged with relevant concepts from DSEO. Ongoing work seeks to identify natural sequences of resources based on prerequisites, building on the architecture to do so from the TechKnAcq project (Gordon et al., 2016).

The EDAM ontology (Ison et al., 2013) provides an extensive representation of four branches: a) Data (with Data Identifiers) denoting types of information that may be processed computationally; b) Operations that may be applied to those data; c) the many Formats in which the data may be expressed; and d) the
7. Possible applications and use cases

7.1 Making recommendations

When learners first use this learning portal, it is proposed that they would be required to complete a “user profile”, where they can state their prior experience/expertise, career level, their interests and, if known, explain what their end goal, career-wise, will be. It is envisaged that this information could then be used in one of two ways:

1. The learning portal will have predefined “learning paths” to allow learners to progress from one level of expertise to another (e.g., from bioinformatics user to bioinformatics scientist6). Using this, the system could therefore recommend a specific set of courses for the learner to take.

or

2. Once a learner has attended certain courses, others can be recommended to them based on previous learners attending the same series of courses.

The theory behind the second case is that learners, sometimes, do not know what courses would be complementary or useful for their work, and do not always know what other courses are available. The courses would be recommended informally, in a similar manner to shopping for items on Amazon, using phrases such as “other learners who attended this course, also attended this course...”. Just as with Amazon ‘recommendations’, users would be free to choose which of the offerings resonated most with their personal training needs. These recommendations are not intended to be required for the accomplishment of a learning goal—they are a gentle guide for people who may be new to the topic or resource or for people who are uncertain about what resources they should view next. It is important to note that interest does not equate to expertise, and before making a course recommendation, the system would check that any prerequisite courses have been completed by the learner. Additionally, by filling out a comprehensive user profile when first using the platform, it is hoped that the recommendations would be accurate, and would ultimately allow more learning paths to be created.

To create personalized learning recommendations, we must also capture the relations between concepts in the domain. For example, before recommending a resource about the concept Machine Learning, we would like to ensure that the learner has familiarity with the concepts of Probability and Algorithms, as these concepts are joined by prerequisite relations. If a learner indicates interest in Machine Learning, we should suggest resources about Supervised Learning, Unsupervised Learning, EM Algorithm, etc., as these are all connected to Machine Learning by hierarchical part-of relations.

8. Innovating across international efforts

There is no “one size fits all” when it comes to learning, and a useful learning path for one scientist may not be useful for another. However, the creation of personalized learning paths and personalized recommendations could more readily guide learners than if they had to navigate the sea of available courses alone. So many resources and courses are offered across multiple sites that it can be daunting and overwhelming for learners to know what they should look for.

8.1 What are the commonalities across international data science training efforts?

The huge volumes of data now being routinely generated from high-throughput instrumentation (imaging data, data from sequencing platforms, etc.) have generated an insatiable demand for bioinformatics and data science training among bioscientists. There are many online training resources, including videoed lectures from conference talks, tutorials, short courses and degree programs, and materials on GitHub and other online repositories. However, while much effort has been placed in creating these materials, it can still be very difficult both for students, who wish to improve their skills, and for trainers, who wish to share and re-use other trainers’ materials, to find what they need. By following the FAIR Data Principles, international initiatives like the ERuDite and TeSS portals seek to make it much easier to find and re-use training materials by identifying, indexing and collating them into central resources. Complementing these largely automated approaches are dedicated, manually maintained training portals, such as the one provided by the GOBLET Foundation. By working together, international training efforts can both identify commonalities and complementarities, and pinpoint synergistic opportunities to drive the field forward.

---

6https://tess.elixir-europe.org/workflows
7https://bio.tools
8https://fairsharing.org
9http://journals.plos.org/ploscompbiol/article/file?id=10.1371/journal.pcbi.1004943&type=printable
To make training materials FAIR, they must be adequately annotated and described so that they can be indexed appropriately. Currently, nascent annotation standards are not yet widely adopted, and there are no widely accepted quality metrics to indicate whether courses are delivering the requisite core competencies to trainees (Welch et al., 2016). Furthermore, to make a course truly reusable, all the information required to re-use the associated materials (including data-sets and tools used for practical exercises, and the exercises themselves) needs to be made available to both students and to fellow trainers. Jupyter notebooks, for example, containing code snippets and notes describing the code, are readily reusable, and there are many Jupyter notebooks on GitHub. But how can these be described so that they can be indexed and readily found? International collaboration could help to define agreed guidelines and standards for course and material description to alleviate this problem.

Alongside the challenge of making training materials FAIR is the growing demand for training, which is outstripping our ability to readily satisfy the changing skill requirements of bioscientists. To meet this demand, more instructors are needed, and new trainers must be adequately prepared and trained to train others. The Software Carpentry (SWC) Foundation has developed a model to ensure that trainers who deliver Software Carpentry- or Data Carpentry-branded workshops are trained to do so by undertaking an online or on-site (for SWC members) course, where they learn about pedagogy. Similarly, ELIXIR and GOBLET have developed a Train-the-Trainer (TtT) program (Via et al., 2017), which was piloted across Europe in 2016, to help build trainer capacity. ELIXIR has also extended these TtT course to Australia via BPA-CSIRO and EBI collaborations set up in 2012. Since then, more than 20 trainers have been trained to deliver bioinformatics training in Australia (Watson-Haigh et al., 2013). Against this background, the international data science and bioinformatics communities could together develop common guidelines for trainer training that would help to raise the quality of training delivery and increase the pool of available trainers.

Having participated in a course, a common theme among students is the desire for some form of certification. Without any aspect of assessment in the course, this will most commonly be a certificate of participation. However, many students are interested in a certificate of accomplishment; this is particularly common with MOOCs, such as those delivered by Coursera, EdX or Udacity. Here, on completion of end-of-course assessments to a set standard – and payment of a fee – students receive a certificate of accomplishment; those who opt to take the course without paying the fee are not certificated. There are practical limitations to the length of a course that determine whether and how it is meaningfully assessable. Nevertheless, for short courses, no assessment guidelines are yet available for universal adoption. Combining our joint experience and expertise, an opportunity therefore exists for the international community to make a set of recommendations for trainers who are developing course assessments, in order for some sort of universal certification to be created.

When reusing online materials, it is important to both cite the sources and give credit to the authors. Materials can be released under several different Creative Commons licences, each of which has different features. Understanding the differences between each licence, and when a particular one should be used, would be beneficial to course developers. Across the international bioinformatics and data science communities, offering guidance to trainers on what licence to use, and allowing end-users to understand their responsibilities when using licensed materials, would be very valuable, and may encourage more trainers to make their resources available online.

There are different types of training courses, including those that i) focus on skills development, ii) tackle data-analysis problems, and iii) develop and encourage creative thinking. Skills-development workshops are the most common, where students follow lectures and tutorial exercises, and perform tasks for which tools and data have already been prescribed: e.g., how to analyze RNASeq data. In data-analysis workshops, which are outcome-driven (and sometimes also referred to as ‘hackathons’ or ‘data jamborees’), rather than follow pre-set, generic examples, participants actually create something meaningful to them using freely available data. H3ABioNet has used the hackathon concept for different scenarios, including the development of Docker containers for selected data-analysis workflows and for collaborative analysis of specific data-sets, bringing together scientists from diverse backgrounds. Extending this idea, ‘creative thinking’ workshops bring together people with different expertise (e.g., bioinformaticians and mathematicians, or bioscientists and data scientists), but without a fixed agenda. Here, the idea is that by discussing issues that one group may have with their data, the other group may propose solutions. Ultimately, the outcome may be, for example, a research proposal for a grant application, leveraging the interests and expertise of both groups.

Data science in biology and medicine can learn from training approaches in other disciplines: e.g., astronomy and its enormous data-sets bring familiar challenges in data analysis and stewardship, and the same basic principles and underlying skill-sets apply. Bioscientists, however, generally have little exposure to the computing and data science skills now necessary for their work, and are likely to seek training much later in the data lifecycle (Attwood et al., 2017). IT professionals play a fundamental role in supporting the life sciences, as their skills are required to install bioinformatics tools on HPC platforms, and to troubleshoot the many problems that are likely to arise. However, rapid advances in technology mean that the current best-practice software for a given task changes rapidly, making it difficult to keep pace. Understanding what bioinformaticians do with these tools, how quickly they change, why they may need so many tools to analyze a given data-set, the demands of storing and analyzing their data, and so on, would be hugely informative to the IT professionals who support their work.

Internships offer another route via which bioscientists may augment their data science skills. The BD2K TCC supports – most commonly for junior faculty staff – the opportunity to partner
with senior data scientists (for 2–3 weeks, or up to multiple months) in order to learn how best to analyze their data. At the end of these “road trips”, the bioscientists return to their own institutions and deliver training courses on what they have learned during their internship. H3ABioNet offers internships to acquire skills from experts at the host institution while working on their own data. The EU offers COST Action grants\(^\text{10}\), which exploit a similar training paradigm – so-called “short term scientific missions”. Finding out about such professional development opportunities can be very difficult, so providing a resource that indexes them would be very useful; this could also expand the program to become a globalized “road trip”, facilitating exchange of expertise internationally if sufficient funding were available. A combined international index could also offer a “matchmaking” service, connecting those who wish to gain a particular skill-set with those who are willing to host them and help them do so (rather like the Knowledge Transfer Program\(^\text{11}\) advocated by the Centre for Proteomic and Genomic Research in Cape Town). Typically, such programs yield research collaborations, funding proposals and publications that last long after the road trips have ended. Consequently, junior researchers are not just supported in developing their skills, but may also benefit from new international collaborations that support them along their career paths.

Increasing the visibility of data science training will facilitate an increase in diversity and enable under-represented groups to better exploit these opportunities. Indeed, some initiatives are already underway to represent and encourage more participation by women in the data sciences. WiDS (Women in Data Science)\(^\text{12}\) is a global initiative to inspire and educate data scientists worldwide that specifically supports women. Their 2017 conference was hosted at Stanford University, and was webcast live or had a delayed broadcast to 80 locations, in 30 countries, worldwide. Similarly, RLadies\(^\text{13}\) is an organisation that promotes gender diversity in the R community, and has multiple city-based chapters across the world.

Aside from connecting international training communities and efforts, a clear value of making training material more visible and discoverable is in sharing experience and expertise, reducing redundancy, and combining efforts to make individual contributions greater than the sum of their parts. Frequent meetings among the groups that represent these communities offer further opportunities for communication that will help each group to understand the other’s work, and to identify skill or knowledge gaps. However, these activities cannot continue without funding. Training bioscientists in data science is an essential investment for 21st-century bioscience. Continued investment in data science training may be possible through large funding initiatives, such as NIH’s, or through large-scale multinational projects like ELIXIR. However, data science skills are also sought after by industry, with companies like Google, Microsoft and Amazon becoming active in biomedical and bioscience research. It may be possible to leverage this interest and partner with them to increase the pool of bioscientists skilled in data science. Furthermore, charitable organizations like the Bill and Melinda Gates or Gordon and Betty Moore Foundations, which fund grand challenges in bioscience, may also be interested in partnering to address the global skills shortage in data science.

9. Conclusions
Modern bioscience increasingly depends upon data science, advanced biostatistics and informatics approaches in order to

---

\(^{10}\)http://www.cost.eu/COST_Actions

\(^{11}\)https://www.africancentreforcities.net/programme/mistra-urban-futures/the-knowledge-transfer-programme/

\(^{12}\)http://www.widsconference.org/about1.html

\(^{13}\)https://rladies.org/

---

**Figure 1.** Diagram of the high-level structure of relations among learners, concepts and skills, and resources.
manage, model and understand the ever-growing amounts of
data being accumulated. This challenge has drawn the attention
of major science consortia in Europe, Africa, Australia and the
United States. Importantly, this necessitates consideration of
how best to provide training in the modern science of “data”,
which is not necessarily part of the undergraduate or even,
perhaps, graduate-level bioscience curricula.

As discussed above, a multitude of factors must be considered in
ensuring efficacious training and educational programs that focus
on the role of training for biomedical data science, and instilling
of core competencies in the development of career paths.
Value exists in standardizing and sharing data science training
resources across national boundaries to ensure commonalities
across communities. However, challenges will always remain
in encouraging adoption of training material standards, which
indicates that such training should be compelling, engaging and
comprehensive. In so doing, the availability of the Internet affords
that resources and learning experiences can be personalized
and customized to the individual. This works to accommodate
the widest set of possible stakeholders in the modern bioscience
enterprise, providing skills that fill important niches in learners’
understanding. International partnerships can help to inform the
world on data science training in ways that are seen as positive
for all such stakeholders. We look forward to continued efforts
between the BD2K’s TCC, ELIXIR’s TeSS, bioCADDIE,
GOBLET, EMBL-ABR and H3ABioNet to promote biomedical
data science through the recommendations made herein.

Data availability
No data is associated with this article.

Grant information
This work was supported in part by National Institutes of Health
(NIH) grant U24 ES026465 02, part of the NIH Big Data to
Knowledge (BD2K) initiative.

The funders had no role in study design, data collection and
analysis, decision to publish, or preparation of the manuscript.

Acknowledgements
The authors wish to thank our respective colleagues for providing
critical comments on an earlier draft of this manuscript. We also
wish to acknowledge the encouragement and support of
Dr. Carol Shreffler of the National Institute for Environmental
Health Sciences (NIEHS).

Figure 2. Example model of a learner.

<table>
<thead>
<tr>
<th>CONCEPT</th>
<th>SKILL</th>
<th>EXPERTISE</th>
<th>INTEREST</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genetics</td>
<td>Comprehension</td>
<td>Novice</td>
<td>Just wants to get started</td>
</tr>
<tr>
<td>Cluster Analysis</td>
<td>R Programming</td>
<td>Proficient</td>
<td></td>
</tr>
<tr>
<td>Cluster Analysis</td>
<td>Python Programming</td>
<td>Novice</td>
<td>Need to learn for a project</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

References

Discovery Index for Data Science. The Fourth WWW Workshop on Big Scholarly
Data: Towards the Web of Scholars (BigScholar). 2017.


Brazas MD, Blackford S, Attwood TK: Training: Plug gap in essential

Corpas M, Jimenez RC, Bongcam-Rudloff E, et al.: The GOBLET training portal:
a global repository of bioinformatics training materials, courses and trainers.


Brazas MD, Blackford S, Attwood TK: Training: Plug gap in essential


Brazas MD, Blackford S, Attwood TK: Training: Plug gap in essential


Brazas MD, Blackford S, Attwood TK: Training: Plug gap in essential


Schneider MV, Griffin PC, Tyagi S, et al.: Establishing a distributed national research infrastructure providing bioinformatics support to life science researchers in Australia. Brief Bioinform. 2017; bbx071.


Van Horn JD: Opinion: Big data biomedicine offers big higher education opportunities. Proc Natl Acad Sci U S A. 2016; 113(23): 6322–6324.


Open Peer Review

Current Peer Review Status: ✔️❓

Version 1

Reviewer Report 03 April 2019

https://doi.org/10.5256/f1000research.20081.r45371

© 2019 Kennedy D et al. This is an open access peer review report distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

David N. Kennedy
Eunice Kennedy Shriver Center, Department of Psychiatry, University of Massachusetts Medical School (UMMS), Worcester, MA, USA

Julianna F. Bates
Eunice Kennedy Shriver Center, Department of Psychiatry, University of Massachusetts Medical School (UMMS), Worcester, MA, USA

Overall, this is an informative manuscript that sets a large number of training issues in big data in the context of the many national and international big data projects. The organization of the manuscript, however, is rather uneven, and the presentation logic is at times, confusing.

Specific Points:
1. Neither the outline of discussion points in the abstract, nor the list of focus questions (p.3) are cleanly aligned with organizational structure of the rest of the paper. It is difficult to reference either one and then try to find the related content in the conference proceedings.
2. Abstract does not set up problem clearly - why do the challenges to conventional methods for sharing, managing and analyzing constitute a training issue? What are the impediments on translational impact: very confusing as worded. The outline of conversations in the abstract appears to be different here from the outline of questions in the Introduction, and yet again different from the actual organization of the paper content. Even though much of the discussion ultimately considers the topics in each of the various outlines, it does make things hard to cross-reference.
3. The structure/organization would be much more accessible if outlined by the headers, and also with the addition of headers to block some of the very long sections into meaningful subunits. Modify headers to conceptually align with the body of the text and to allow the reader to target content areas.
4. Section 2 (Training in biomedical data science) needs a topic sentence or some overview introductory statement to orient the reader to what is coming (descriptions of 7 different big data initiatives, from different regions of the world), and why.
5. The transition from initiative summaries in section 2 to what collectively seems to constitute a section on inter-initiative collaboration is a very awkward transition in the section as currently structured and seems like a non-sequitur. Section 3, which is a single paragraph titled 'International
interactions' seems to be closely related to the preceding few paragraphs on inter-initiative collaborations.

6. Re: Section 4, it is very unclear to the reader why the discussion shifts from an overview of many initiatives to a) a specific and detailed comparison of online portals from two of those initiatives, and b) the conceptual basis of a FAIRness comparison. Some topic sentence here would be important.

7. What is the primary conclusion from the FAIR principles comparison for TeSS x ERuDite?

8. Identification of and utility for stakeholders are mentioned as priorities in both the Abstract and the Introduction, if this is indeed an intended priority for the readership audience, it would be helpful to summarize conclusions in a section with a corresponding header.

9. Very confusing use of the term ‘translational impact’ in the training context. Translational impact implies ‘impact of translational research’ – it isn’t obvious how this relates to training - there is a missing link between big data and the need for people to be trained to work with it in order to enable translation between current cutting edge biomedical data science and research outcomes.

10. Section 8, pages 10-13, would be more accessible with some from some conceptual blocking/sub-headers e.g Need for Standards, Training and Trainers, Certification/Credentialing of Trainers/Trainees, Licensing of Materials, Courses and Training Paradigms, Data Science-interface between bioscientists and IT.

Overall, the content is valuable, but the organization needs some work to improve the clarity of the presentation.

Is the topic of the opinion article discussed accurately in the context of the current literature? Yes

Are all factual statements correct and adequately supported by citations? Yes

Are arguments sufficiently supported by evidence from the published literature? Yes

Are the conclusions drawn balanced and justified on the basis of the presented arguments? Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Neuroimaging, Nuroinformatics, development

We have read this submission. We believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.
This article summarizes the discussions from a meeting of several organizations from different parts of the world that all develop and collect materials for education in biomedical data science.

One of the main strengths of the article is that it includes a world-wide perspective on training in the intersection of computing and the life sciences, with representatives from four different continents. As expected from such a forum, the paper covers a lot of ground, ranging from theories of learning, through technical development of databases for learning resources and through topics such as interactions with stakeholders. This is a major strength of the article, but it also makes the reading of the article slightly daunting, and I worry that it would make it hard for a potential reader to navigate. I propose adding a roadmap in the beginning of the article, which would direct readers towards their topic of interest. I also think that the article would benefit from relegating some technical details, such as the detailed FAIRness assessment in section 4, or the details of the schema in section 5.2 to appendices. The former (section 4) could possibly also be summarized in a table. Similarly, some theoretical background could possibly be elided, and references provided for readers who want to (e.g., Bloom’s taxonomy in section 6.1).

Similarly, at the very end of the article, I think that it would help to summarize the main conclusions and recommendations made in the article and reiterate them.

One topic that I would have liked to see addressed more is the ability that relatively decentralized systems have to provide objective assessments of users' skill levels, progress and outcomes. The work of the Johns Hopkins Data Science Group on this topic is particularly compelling. For example, articles that they published about their MOOC series: Kross et al., 2017 and Hadavand et al., 2018, and work that they have done on systems for automating production of educational resources, and primarily the Swirl project (https://swirlstats.com/), which provides a uniform interface for creating data science courses that run in the RStudio console.

One additional topic that could be further addressed is the motivations that lead researchers to contribute training materials to the repositories that are mentioned in the article. This is addressed in 5.1.C. Another option that is not mentioned is providing venues for publication, and mechanisms to cite and provide credit and feedback, to the creators of OER. One possible venue for this is the recently-started Journal for Open Source Education, https://jose.theoj.org/ (with the full disclosure that I am a member of the editorial board of JOSE’s partner journal for open source software (JOSS) and a contributor to JOSE).

I found the assessment of interoperability between projects developed by different organizations in sections 4 and 5 to be compelling, and I was wondering why bioCADDIE’s DDI was not included in these assessments.

Minor:

In the abstract, the sentence: "Discussions focused on the role of i) training for biomedical data science; ii) the need to promote core competencies, and the ii) development of career paths" should be "Discussions focused on i) the role of training for biomedical data science; ii) the need to promote core competencies, and iii) the development of career paths”

Introduction, I believe that the sentence: “In May 2017, during a two-day meeting between the BD2K Training Coordinating Center (TCC), ELIXIR Training/TeSS, GOBLET, H3ABioNet, EMBL-ABR,
bioCADDIE and the CSIRO, convened in Huntington Beach, California, to describe our respective activities and how these might be leveraged for wider impact on a global scale* is missing something (maybe a verb?)

On page 9: I believe that the sentence “The example in the figure depicts the situation of a learner who wants to find out about the field of genetics but does not know where to start.” refers to Figure 2. It would be good to spell that out. Also, in the following paragraph: “The key aspect of this underlying model is that the table shown above would need a large number of rows that codify the various ...”. Because of the layout of the article, the table in the Figure is actually several rows “below” this paragraph, so it would be good to be explicit here as well.

References

Is the topic of the opinion article discussed accurately in the context of the current literature? Yes

Are all factual statements correct and adequately supported by citations? Yes

Are arguments sufficiently supported by evidence from the published literature? Yes

Are the conclusions drawn balanced and justified on the basis of the presented arguments? Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Neuroinformatics, data science, cognitive neuroscience

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.
The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com