Drug Discovery Today • Volume 00, Number 00 • December 2019 PERSPECTIVE

ELSEVIER

feature

Road to effective data curation for translational research

Wei Gu[1,2,‡], **Samiul Hasan**[3,‡], **Philippe Rocca-Serra**[4,‡] **and Venkata P. Satagopam**[1,2], venkata.satagopam@uni.lu

Translational research today is data-intensive and requires multi-stakeholder collaborations to generate and pool data together for integrated analysis. This leads to the challenge of harmonization of data from different sources with different formats and standards, which is often overlooked during project planning and thus becomes a bottleneck of the research progress. We report on our experience and lessons learnt about data curation for translational research garnered over the course of the European Translational Research Infrastructure & Knowledge management Services (eTRIKS) program (https://www.etriks.org), a unique, 5-year, cross-organizational, cross-cultural collaboration project funded by the Innovative Medicines Initiative of the EU. Here, we discuss the obstacles and suggest what steps are needed for effective data curation in translational research, especially for projects involving multiple organizations from academia and industry.

Introduction

Billions of dollars are spent annually on generating clinical and translational research data. Yet despite such significant levels of investment, the amount of data that are reused remains surprisingly low. Besides the legal constraints, two main reasons explain this deficit: first, the difficulties in finding and accessing data sets themselves, and second, the effort required to harmonize data sets both syntactically and semantically. The perpetual need for added curation efforts highlights a lack of interoperability between digital assets. This is symptomatic of the dichotomy plaguing the life sciences domain when it comes to data management and asset handling. On the one hand, there is an operational gap in terms of training life-science researchers in the necessary skills, tools and culture required to turn data sets into FAIR (findable, accessible, interoperable, reusable) resources [1]. And on the other hand, there is a cultural divide between the worlds of academia and industry, which translates into diverse and frequently divergent attitudes towards the adoption of data-management standards (Fig. 1). This is compounded by an array of often competing standards specifications sponsored by different stakeholders, whose views of the world are not necessarily discordant, but are not entirely aligned, either. Hence, unless a life-science *lingua franca* emerges, realising the vision of a FAIR data exchange requires strategic thinking with a good understanding of existing standards resources and a clear commitment to using them.

Furthermore, funders, sponsors, program managers, data analysts and scientists need to fully appreciate the added burden of format conversions, language translations and mapping between terminologies. They will also have to bear in mind that standards are not static artefacts but evolving entities designed to reflect the growth of domain knowledge. Therefore, they need to plan for upgrades, migration and obsolescence strategies, leveraging learnings and practices well-entrenched in other domains in the software industry.

The Innovative Medicines Initiative (IMI) is the largest public–private partnership in the life-sciences domain in Europe, and it is focused on developing better and safer medicines for patients. Central to this strategy is a knowledge management (KM) environment that provides sustainable access to the data in an integrated manner. To meet this challenge, the IMI European Translational Research Infrastructure &

https://doi.org/10.1016/j.drudis.2020.12.007

Please cite this article in press as: Gu, W. et al. Road to effective data curation for translational research, Drug Discov Today (2020), https://doi.org/10.1016/j.drudis.2020.12.007

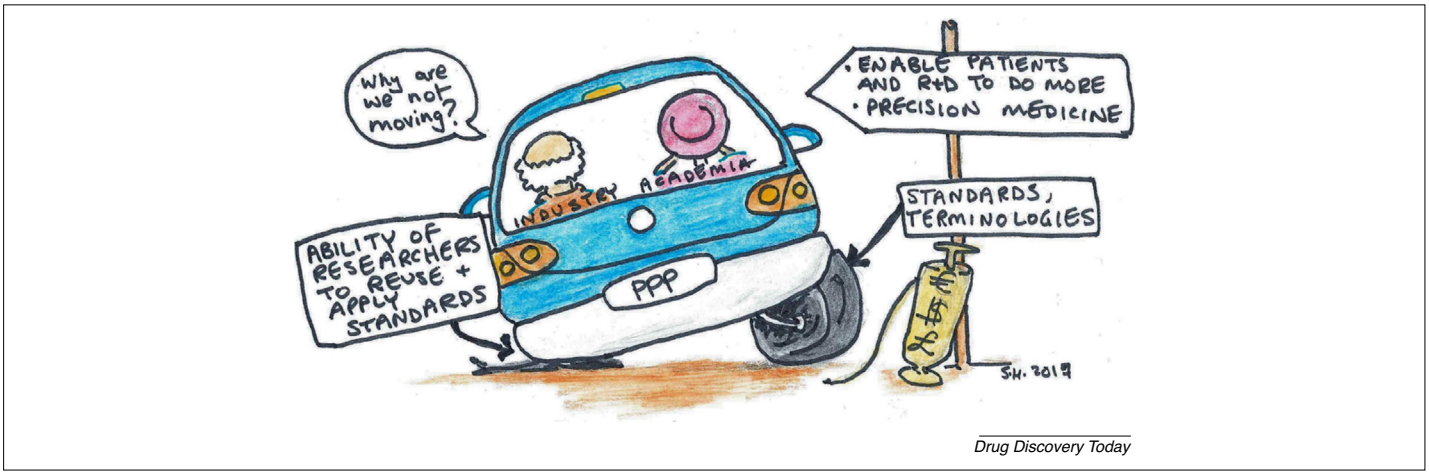

FIGURE 1

Fundamental data curation policies, technical infrastructure and cultural adoption are necessary to help public–private partnerships deliver new and better-structured clinical and omics data to support personalized medicine.

Knowledge management Services (eTRIKS) project focused on building a sustainable KM platform and provided support at the level of data management to other IMI projects. Curation, as an essential part of data management, connects key components, such as vocabulary management and data hosting, in a longer 'value chain' that leads to a substantial improvement of the utility of data for research. Based on our learnings from eTRIKS, we outline the key steps to define, specify, roll out and enact robust, proven, standards-aware data-management plans (DMPs). These experiences are relevant to all data-intensive clinical and translational studies. This work builds on the eTRIKS standards starter pack (SSP) [2], which surveyed, evaluated and compiled a collection of data standards of relevance to the field.

Our aim is to share the practical knowledge thus gained to assist organizations to maximize the value of their own and public data and to practically assess when to implement data curation in translational-research operational processes. This involves outlining a comprehensive concept of data curation as an integral part of the global concept of a DMP. It covers data handling processing from end to end, introducing the use of data standards from the study planning stages to data-set socialization through planned releases.

Experience of data curation in IMI eTRIKS supported projects

eTRIKS directly engaged with IMI projects by providing support in data curation and training. This facilitated data access to consortium members via data-sharing platforms such as tranSMART [3]. The curation activities targeted compliance to standards as recommended by the SSP [2]. Curation therefore covered aspects of metadata structuring (the syntax) and elements of content annotation (the semantics). The IMI projects directly involved were: UBIOPRED [4], OncoTrack [5], RA-MAP [6], ABIRISK [7], APPROACH (https://www.approachproject.eu/about-approach) and AETIONOMY [8].

Initially, eTRIKS adopted a 'full support' model by allocating experienced personnel to curate data for lead client projects. However, as if to confirm the aphorism 'no battle plan ever survives contact with the enemy', this soon resulted in having to make adjustments, often significant ones, owing to the following shareable obstacles:

Obstacle 1: underestimation of the time required to clear legal hurdles

The negotiation of material or data processing agreements between consortia turned out to be a time-consuming process. The complex details resulted in > 20 months spent to grant data access for curation. This restriction was compounded further by apprehensiveness in sharing information about study variables, which resulted in further unanticipated delays.

It is crucial for future consortia to clear the path to data exchange between partners to prevent being bogged down in juridical no man's land.

Obstacle 2: under resourcing of the curation activities and sustainability issues

The number of curators allocated to eTRIKS only allowed us to realistically engage with five to eight lead projects, which would have fallen far short of eTRIKS' engagement goal of 40 projects. The real costs and the scale of the gap between the curatorial resources and the number of IMI projects being funded was quickly and crudely exposed.

To provide sustainable curation support to subsequent IMI projects, an additional 'light weight' support model was introduced. Under this remit, eTRIKS has developed data-curation guidelines and an IMI curator training course [9]. The guidelines and a series of training sessions were provided to supported projects to enable independent curation. Even this model often faced challenges owing to a lack of personnel in many projects.

We strongly suggest that future consortia/projects should plan sufficient curation resources in their proposals to ensure an efficient data flow from data production to analyses.

Obstacle 3: underestimation of the 'cultural divide' of standards

Industry strength standards have a daunting complexity. This means that bringing data managers to the required levels of competence demands adequate training and resourcing. By contrast, data standards used in academia often have a different origin, with pragmatism, agility and flexibility at their core, and with loosely coupled implementation. These distinct attitudes tend to polarize practice and make finding common ground challenging. Having these two worlds exchange, coordinate and develop global standards consumes time, expertise, training and knowledge transfer for harmonization and stability.

We urge scientific leaders and researchers from both academia and industry to reach out to funders and sponsors regarding this issue, so that it will be acknowledged by funding agencies and study sponsors to promote convergence of efforts.

Please cite this article in press as: Gu, W. et al. Road to effective data curation for translational research, Drug Discov Today (2020), https://doi.org/10.1016/j.drudis.2020.12.007

Obstacle 4: lack of a proper data-management plan

A DMP is vital for various activities involving the handling of data or information that is outlined in the protocol to be collected, analysed and shared. It also contains detailed information necessary for the curators to understand the data to be processed. Yet all too often, DMPs are missing or, when available, are too thinly written to bootstrap the curation process.

Because of the essential role of a proper DMP in efficient and effective data curation, we have made some suggestions in detail in a dedicated section 'Support for effective curation within a data-management plan'.

Obstacle 5: lack of metadata descriptors

Although it relates to the DMP, this point is worth separate attention owing to its large practical impact on curation, especially for retrospective data sets for which a DMP might not be available. Assuming that curators have overcome the previous hurdles and obtained a data set, the variables and their value sets are often so poorly annotated that it is near impossible to decipher them without further interactions with the primary researchers, who are typically pressed for time or unreachable. Mostly, it falls on the curators to understand the pre-curated ('straight from the hose') data. This affects the efficiency of curation and frequently results in erroneous interpretations.

We strongly recommend that a proper data dictionary be developed to document the metadata of each variable (name, label, type, unit, value ranges, controlled vocabularies, dependencies, etc.). This information should be collected as early as possible or, if possible, even before the project starts: a key step to make data FAIR by design.

Support for effective curation within a data-management plan

The major research funders have now evolved relatively (or somewhat) homogenous approaches to their DMPs. DMP templates are available from numerous sources, but the UK Digital Curation Centre [10] and ELIXIR DMP-wizard (https://ds-wizard.org/) provide resources that reflect broad current thinking. DMPs are now evolving from simple checklists to documents covering the entire data custody plan from collection to publication, ensuring compliance with the formats and annotation requirements of repositories (e.g., CONSORT, http://www.consort-statement.org/), regulators (e.g., CDISC, https://www.cdisc.org/standards/therapeutic-areas) and European law (EU General Data Protection Regulation). Key performance indicators are being developed to determine the level of adherence to principles such as the FAIR principles both in the EU (e.g., FAIR-DOM.org) [11] and in the United States (US National Institutes of Health DCCPC KC1 FAIR access) [1]. These efforts share a common objective: To select data standards, terminologies or ontologies that will be implemented in the data-capturing phase wherever possible, leading to an ideal limit of 'free of free text' data sets. The upfront declaration of standards, their name and version thus allows the availability of key provenance descriptors, which ought to be associated to metadata capture templates, which themselves ought to be identified, licensed and versioned. This followed the work by Dietrich *et al.* on DMPs [12]. In support of these tasks, efforts such as FAIRsharing [13] and CDISC Share are laudable initiatives.

However, in spite of this progress, many instances of DMPs seem to be too weak, or even an afterthought. DMPs need to be prepared at the early stages to provide a means to deliver prospective, design driven, compliant data-collection blueprints that are digital artefacts in their own right, both machine readable and actionable. To reach this stage, it is essential that from inception, data controllers and data processors (including users and statisticians) should be closely involved, along with study designers. They should ascertain that the protocols embed sufficient safeguards to account for bias and document how lurking variables are dealt with; this is crucial, and the lack of such information in legacy or public data sets is usually a cause for their exclusion in meta-analyses and reviews, as the use of insufficiently described data sets carries a risk of misinterpreting signals.

Some data modalities present new challenges, such as the high-dimensional data sets generated by omics technologies. These challenges are based on the emergence of new platform and technology descriptions and signal-processing methodologies (e.g., normalization and batch-effect corrections). In addition, the computational workflows and information technology (IT) infrastructures for storing and executing the computations are evolving, adding to the complexities.

The DMP must be augmented with detailed data-access conditions and terms of use (such as the Data Use Ontology GA4GH standard (https://www.ga4gh.org/news/data-use-ontology-approved-as-a-ga4gh-technical-standard) and the European Genome-phenome Archive terms [14]) to be compliant with legal frameworks and to safeguard patients' rights and privacy. Hence, data processors should provide (if they host the data) technical implementation plans to host (transfer), process, analyse and share data in a secure and scalable manner. They ought to make available a clear picture of the data flow and provide standard operating procedures for data access. Models such as Data Tag Suite (DATS) [15,16] and the data access components could be followed.

The IMI published the 'Guidelines on FAIR Data Management in Horizon 2020' [17] with a FAIR DMP template to harmonize and improve curation activities. Another good resource is the 'Good Clinical Data Management Practices', which aims to help implement sustainable curation activities [18].

Support for effective curation within a curation community

But beyond a DMP, what would benefit the translation-research community most is an open, precompetitive curation community. This ought to include academia, data managers, bioinformaticians, standards development, IT developers, physicians, statisticians and curators. In the case of the IMI, this includes European Federation of Pharmaceutical Industries and Associations (EFPIA) partners, academia and small and medium enterprises. For that, members of the research community should establish and maintain the following key pillars.

People and organizations

Build up a network of people with diverse expertise (domain knowledge, IT skills, algorithms, legal and ethical requirements and coordination) and create a forum for all curation stakeholders to exchange ideas, requirements and resources. Despite initiatives like the Biocuration Society and the Pistoia Alliance, data created by the pharmaceutical industry are still treated as trade secrets, leading to a lot of wheel reinvention in every organization.

Projects and ideas

Seek out opportunities to apply joint grants to improve curation methods, resources (e.g., metadata and reference-data repositories) and infrastructures.

Data sets

Release training data sets of variables, their value set and actual 'dirty data'. As machine learning and artificial intelligence (AI) methodologies are gaining importance, it is crucial that a community is built around assembling training data sets that can establish the basis for tools geared to bootstrap curation efforts. To that effect, obtaining common data elements from the National Cancer Institute (NCI) (https://www.cancer.gov/), US National Institutes of Health,

Please cite this article in press as: Gu, W. et al. Road to effective data curation for translational research, Drug Discov Today (2020), https://doi.org/10.1016/j.drudis.2020.12.007

Federal Interagency Traumatic Brain Injury Research (FITBIR) (https://fitbir.nih.gov/) and many others, organized by clinical domain, would be essential to seed an AI engine.

Training

Provide training not only to curators, but also to principal investigators, clinical researchers, IT staff and legal experts to help them understand the importance of curation and other people's views and roles in the whole data-management process related to curation.

Support for effective curation within infrastructure developments

Thanks to its widespread use of genomics and high-throughput or high-resolution imaging techniques, translational research sits firmly in the realm of big data science. Effective and efficient data curation requires now more than ever strong support from infrastructure, reference (meta)data resources, and tools so that FAIR principles can be implemented. The goals of such infrastructure are to reduce the overall time of curation at the same time as improving the quality of the outcome. Other benefits include lowering the barrier of curation and being able to trace and reproduce curation steps. A modern curation infrastructure requires care and proper set-up, following the stepwise process shown in Figure 2.

We would like to emphasize that many of the functionalities mentioned here are already available piece by piece in existing software applications. For example, some of the components needed are already provided by ontology support tools (in Europe, https://www.ebi.ac.uk/spot/; in the United States, https://bioportal.bioontology.org/), OpenRefine (http://openrefine.org/) and many community versions of commercial tools, such as Pentaho-Kettle (https://github.com/pentaho/pentaho-kettle) and Talend Open Studio (https://github.com/Talend/tbd-studio-se). When a curation infrastructure is to be set up, one should reuse and integrate the existing tools as much as possible to avoid reinventing the same solutions.

Conclusion

This article delivers an honest review of the lessons learnt from our experiences in supporting several IMI projects with their data-curation needs. The outcome of these efforts has resulted in the reuse of data in several multi-party projects within IMI [4–8] and beyond. For example, the H2020-SYSCID project (http://www.syscid.eu/) reused the curated data related

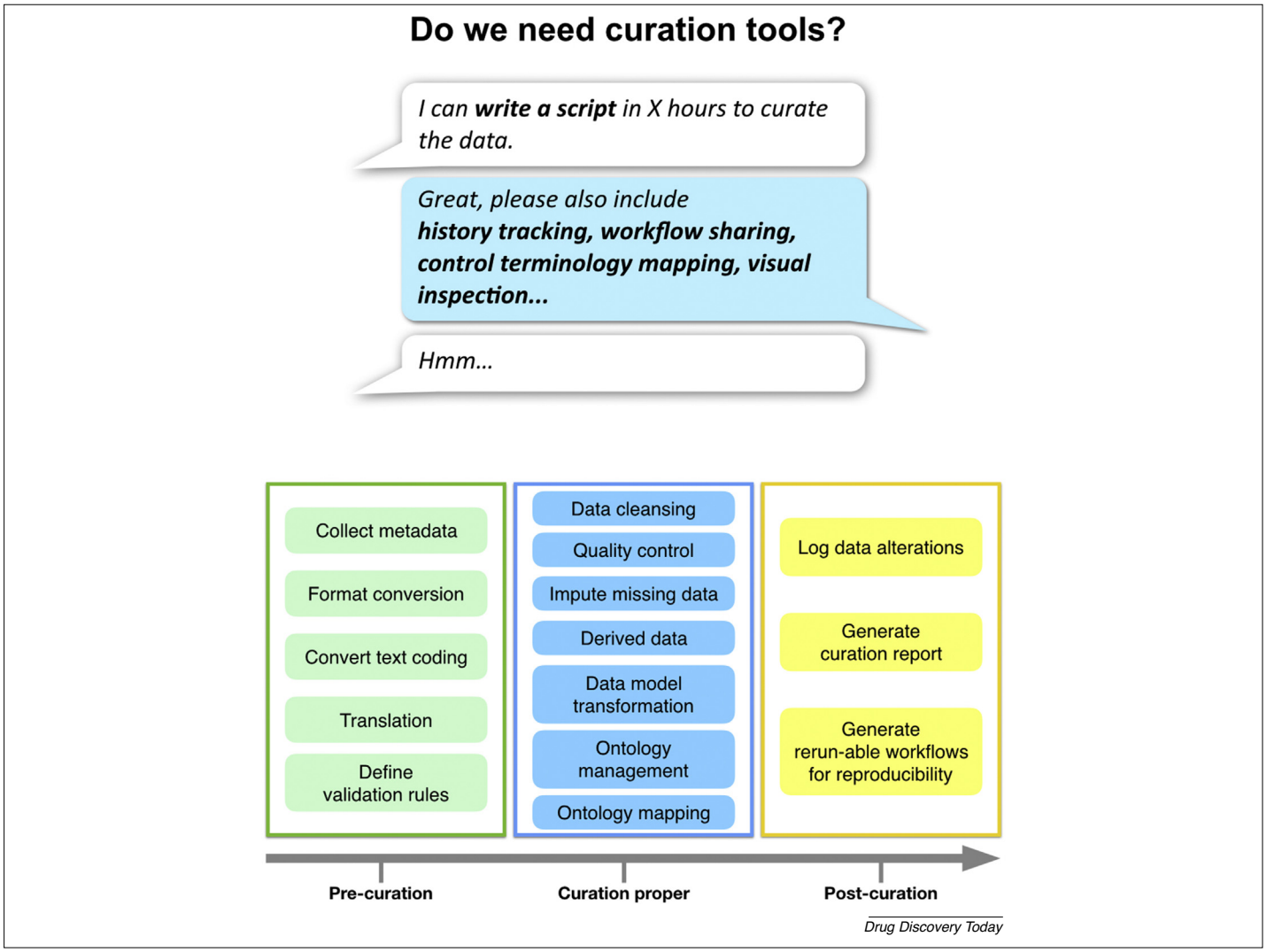

FIGURE 2

Tools such as an efficient IT infrastructure are needed for effective and efficient data curation. The upper example conversation is a typical starting point for researchers (or data managers and data analysts) who try to perform data curation without the proper tools. The lower part is a list of functional components for an ideal curation infrastructure throughout the life cycle of data curation, from pre-curation to curation proper and post-curation.

Please cite this article in press as: Gu, W. et al. Road to effective data curation for translational research, Drug Discov Today (2020), https://doi.org/10.1016/j.drudis.2020.12.007

to autoimmune disease, and the NCER-PD project (https://parkinson.lu) reused the curated data related to Parkinson's disease. Data curation remains the main technical challenge for the reuse of data in clinical and translational research.

Cleaning and scrubbing data is not only a time-consuming and tedious process, it is also an expensive one. Yet projects still consent to that expenditure, even though they are less inclined to agree to a more robust, 'front-loaded' approach to data management. Indeed, possibly the most effective way to wrestle the problem is to form a well-structured DMP at the start that covers all the steps from data generation or capture to data analysis, with the highest possible precision, definitions and discipline. The resources, roles and responsibilities of each party should be clearly defined and planned. Necessary infrastructure should be deployed and, if needed, developed to accelerate and improve the curation activities. A curation community is needed to sustain and continue developing knowledge, technologies and expertise. Last but not least, awareness of the importance of data curation and the risk of a lack of planning should be disseminated to the clinical and translational research field, so researchers can avoid making the same mistakes time and time again.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We highly appreciate the essential contributions of the eTRIKS consortium members: Dorina Bratfalean, Adriano Barbosa-Silva, Serge Eifes, Francisco Capdevila Bonachela, Ibrahim Emam, Manfred Hendlich, Paul Houston, Chris Marshall, Paul Peeters, Kavita Rege, Fabien Richard, Martin Romacker, Andreas Tielmann, Michael Braxenthaler, Susanna-Assunta Sansone and Reinhard Schneider. This work has received support from the IMI Joint Undertaking under grant agreement no. 115446 (eTRIKS), the resources of which are composed of financial contribution from the EU's Seventh Framework Programme (FP7/2007–2013) and The European Federation of Pharmaceutical Industries and Associations (EFPIA) companies' in-kind contributions (www.imi.europa.eu).

References

1 Wilkinson, M.D. *et al.* (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3, 160018
2 Rocca-Serra, P. *et al.* (2016) eTRIKS Standards Starter Pack Release 1.1 April 2016. *Zenodo* . http://dx.doi.org/10.5281/zenodo.50825
3 Szalma, S. *et al.* (2010) Effective knowledge management in translational medicine. *Brief. Bioinform.* 8, 68
4 Wheelock, C.E. *et al.* (2013) Application of' omics technologies to biomarker discovery in inflammatory lung diseases. *Eur. Respir. J.* 42, 802–825
5 Gu, W. *et al.* (2019) Data and knowledge management in translational research: implementation of the eTRIKS platform for the IMI OncoTrack consortium. *BMC Bioinform.* 20, 164
6 Cope, A.P. *et al.* (2018) The RA-MAP Consortium: a working model for academia–industry collaboration. *Nat. Rev. Rheumatol.* 14, 53–60
7 Bachelet, D. (2016) Occurrence of anti-drug antibodies against interferon-beta and natalizumab in multiple sclerosis: A collaborative cohort analysis. *PLoS One* 11, e0162752
8 Hofmann-Apitius, M. *et al.* (2015) Bioinformatics mining and modeling methods for the identification of disease mechanisms in neurodegenerative disorders. *Int. J. Mol. Sci.* 16, 29179–29206
9 Marchetti, G. and Jullian, N. (2017) *eTRIKS Final Training Curriculum: Deliverable D6.6.* eTRICKS
10 DCC (2013) *Checklist for a Data Management Plan, v.4.0.* Digital Curation Centre
11 Wolstencroft, K. *et al.* (2016) FAIRDOMHub: a repository and collaboration environment for sharing systems biology research. *Nucleic Acids Res.* 45, D404–D407
12 Dietrich, D. *et al.* (2012) De-mystifying the data management requirements of research funders. *Issues Sci. Technol. Librariansh* 70 . http://dx.doi.org/10.5062/F44M92G2
13 Sansone, S.A. *et al.* (2019) FAIRsharing as a community approach to standards, repositories and policies. *Nat. Biotechnol* 37, 358–367
14 Lappalainen, I. *et al.* (2015) The European Genome-phenome Archive of human data consented for biomedical research. *Nat. Genet.* 47, 692–695
15 Sansone, S.A. *et al.* (2017) DATS, the data tag suite to enable discoverability of datasets. *Sci. Data* 4, 170059
16 Alter, G. *et al.* (2020) The Data Tags Suite (DATS) model for discovering data access and use requirements. *Gigascience* 9 giz165
17 European Commission (2016) *H2020 Programme: Guidelines on FAIR Data Management in Horizon 2020.* European Commission
18 Society for Clinical Data Management (2020) *Good Clinical Data Management Practices (GCDMP).* Society for Clinical Data Management

Please cite this article in press as: Gu, W. et al. Road to effective data curation for translational research, Drug Discov Today (2020), https://doi.org/10.1016/j.drudis.2020.12.007