A Tentative Maturity Model for Biomedical Data Curation

Mariam Alqasab*, Suzanne M. Embury and Sandra Sampaio

Department of Computer Science, The University of Manchester, Oxford Road, Manchester, UK

ABSTRACT

Data curation has become an active area for research, since it is necessary to ensure the long-term, sustained usefulness of scientific data. This has led many communities to adopt data curation practices to improve their data quality. However, at the moment, there is little guidance for curators in reaching and sharing best practice, either for new curators looking to establish a cost effective curation regime or for more established curators looking to catch up with other communities. We propose a tentative maturity model for biomedical curation to help fill this gap.

The maturity model provides a series of stages of curation practice, giving details of the tools and techniques that should be in use by a curation team at each level. Teams can assess their current level, compare their level with other teams, and (most importantly) gain guidance on cost-effective and feasible improvement steps that will raise the quality of their curation without requiring impractical amounts of additional curator time or expertise. This poster will present our tentative maturity model, and invite ICBO participants/curators to rank their own performance and provide feedback on the model.

1 INTRODUCTION

With the growth of data-driven science, the curation of public and community data sets has become a necessary task for ensuring the long term usefulness of scientific data. Scientific data typically comes in two forms: experimental results (objective measures of reality) and the interpretation of those results in the form of statements about the structure, organisation and function of the things being observed. There are curation challenges with both types of data, but the most substantial difficulties lie in the curation of the interpretive data. This data describes the models and hypotheses about reality that prevail within the community that owns the data. As such, it is often complex in form (requiring several ontologies to describe), it can change rapidly or remain current for many years, it is subject to disagreement within the community, and can be superseded as new experimental results come in. Perhaps most significantly, the source of this data is not a machine, which spits out experimental results at high volume but in regular and predictable format. This interpretive data comes from people, in the form of scientific publications. The principal task of a biomedical curator is to ensure that the interpretive data in the resource they curate (sometimes called metadata or annotations) is kept up-to-date with the expert views presented in the scientific literature.

Researchers have tried to improve the process of data curation by implementing tools to speed up the process, such as Canto (Rutherford *et al.*, 2014) OntoMate (Liu *et al.*, 2015), establishing ways to share data between communities or providing a collaboration environment for curation such as MIntAct (Orchard *et al.*, 2013) and OntoBrowser (Ravagli *et al.*, 2017). Communities also produced their own ways to curate their data. For example, FlyBase allows paper authors to participate in the curation process, but some other communities do not.

Despite the various research done in data curation, there is still a need for a common understanding for the curation process.

At present, there is little general advice for curators of biomedical data. An exception is the useful proposal by Hirschman *et al.* for a general biocuration workflow, but even this proposes a one-size-fits-all solution, which may not be appropriate for all communities. Instead, we propose the creation of a *maturity model* for biomedical data curation. A maturity model indicates the different stages of "maturity" of an organisation or group in performing some task. The stages describe good practice (and even best practice) for aspects of the task under consideration, as well as commonly occurring forms of poorer practice. The underlying assumption behind maturity models is that it is not usually possible for a group of people to carry out best practice in a new area from scratch. The need to understand the particular needs of the task and the particular abilities of the group mean that time and experience is needed to learn the best approaches. The maturity model can tell a group where they currently stand in terms of good practice, and can indicate plausible steps for gradual improvement over time. Using the model, newer groups can avoid the mistakes made by other groups, and can improve more quickly. More established groups can identify areas where their (often scarce) resources can be deployed for maximum improvement effect.

In this poster, we will present the current tentative maturity model, and seek feedback from curators and researchers attending ICBO. The poster will be interactive, allowing viewers to rank their own curation performance on the model, and obtain suggestions for improvements. We will also provide mechanisms for participants to leave feedback when the poster is not being "manned".

2 AN OVERVIEW OF OUR PROPOSED MATURITY MODEL

In order to produce our maturity model for biomedical data curation, we reviewed the literature of biomedical data curation in the last five years, and the literature on maturity models. We also investigated how the curation process works in five different real-world communities: UniProt[1], BioGRID[2], FlyBase[3], Saccharomyces Genome Database[4] and the Rat Genome Database[5]. Curation can be done either based on the literature or data available in the community repository. In other words, the curation process will be triggered if a new publication appears in the area, or when defects found in the repository data. We also found that the five

*Corresponding author: mariam.alqasab@postgrad.manchester.ac.uk

[1] http://www.uniprot.org/

[2] https://thebiogrid.org/

[3] http://flybase.org/

[4] http://www.yeastgenome.org/

[5] http://rgd.mcw.edu/

communities, which we investigated, have different criteria to curate data, as each community applies different ways for curation.

According to Paulk *et al.*, 1993, a maturity model consists of a number of dimensions that contribute to the model goal, and each dimension is divided into a number of levels of maturity (typically 4 to 6 levels). Level one indicates low maturity level and the higher level indicates the highest maturity level. Each level contains a number of goals to achieve the required maturity model.

In our maturity model we tentatively propose five dimensions for data curation, with 5 maturity levels. The levels in our maturity model start with proposing manual ways for curation, then gradually develop the curation process until it becomes completely automatic if applicable. The five components of the maturity model are as follows:

1. **Adding and editing repository data.** Some communities do curation based on the data they receive or have in their repository, which means looking for defects in data and fix them. Also, curating literature if the defects in data requires.
2. **Searching for and selecting from the new literature**. This component describes the criteria of searching and choosing among new publications in the area.
3. **Reading and extracting data from the abstract.** When a list of publications is determined, then the abstract of each paper need to be read to extract data.
4. **Reading and extracting data from the full paper.** In the previous component, it need to be determined whether a paper is curtable or not. Then, the paper will be curated in full if needed.
5. **Documenting curation results.** In this component, we do not describe the process of detecting defects in data and fixing them, but we care about highlighting the results of curating data to improve the curation process by allowing curators to visualise how communities curate their data through time.

To use the proposed maturity model, a curation team needs to go through a number of steps. First, defining the level of maturity of the current curation process followed by the community. This is done by determining the maturity level of each dimension. Second, if all dimensions have the same maturity model, then the maturity level of the community should be raised by one level. Otherwise, if the maturity levels of the components are different, then we need to set the highest level to be our target level to achieve. Then, we refer to our maturity model and determine the changes that need to be achieved to raise the maturity level of the component. This will be applied to the rest of the components. Finally, the whole process can be repeated through time until the community reach the highest level of maturity.

3 CONCLUSION

The main goal of this poster is to propose a tentative maturity model for biomedical data curation, with the aim of soliciting preliminary feedback from the biomedical and curation communities. The model gives a general explanation of how to identify the maturity level of each curation step and suggest improvements to reach a sufficient level of maturity. The aim is to achieve the maximum quality of curation with current or fewer resources.

Feedback at this early stage in the work is sought on the overall idea of creating a maturity model for curation, and also on the details of the form the model takes. At this stage, we make no strong claims for this set of levels being the “right ones”, nor for the set of dimensions being complete. Our current work involves gathering feedback from curators and researchers on the model, and incorporating feedback. Once a more stable model has been created, we will create a web resource to allow curation teams to assess their current model, and to obtain suggestions for improvements based on their target maturity levels. We hope that the final maturity model will benefit a range of biomedical communities, by allowing ideas, tools and best practice to be shared and refined.

REFERENCES

Liu, W., Laulederkind, S. J., Hayman, G. T., Wang, S.-J., Nigam, R., Smith, J. R., De Pons, J., Dwinell, M. R., and Shimoyama, M. (2015). Ontomate: a text-mining tool aiding curation at the rat genome database. *Database*, **2015**, bau129.

Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., Campbell, N. H., Chavali, G., Chen, C., Del-Toro, N., *et al.* (2013). The mintact projectintact as a common curation platform for 11 molecular interaction databases. *Nucleic acids research*, page gkt1115.

Paulk, M. C., Curtis, B., Chrissis, M. B., and Weber, C. V. (1993). Capability maturity model, version 1.1. *IEEE software*, **10**(4), 18–27.

Ravagli, C., Pognan, F., and Marc, P. (2017). Ontobrowser: a collaborative tool for curation of ontologies by subject matter experts. *Bioinformatics*, **33**(1), 148–149.

Rutherford, K. M., Harris, M. A., Lock, A., Oliver, S. G., and Wood, V. (2014). Canto: an online tool for community literature curation. *Bioinformatics*, page btu103.