A 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

Quality is an important aspect that needs to be managed in databases, as the importance of data is determined by its quality. This draws the attention of many database providers to care about curating their data in order to maintain data quality over time. Also, this leads database providers and researchers to investigate the area of data curation and propose ways to improve it, either through providing tools to automate the process or to support human curators in making changes to the data. However, among all available suggestions to improve data curation, to the best of our knowledge, no a general description of the curation process has been given that also provides solutions to improve it, and that can help database providers to assess how mature their approach to data curation is. To fill this gap, this paper proposes a maturity model, that describes the maturity levels of biomedical data curation. The proposed Maturity Model aims to help data providers to identify limitations in their current curation methods and enhance their curation process.

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 (measurements) 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 prevails 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it can change rapidly or remain current for many years, it is subject to disagreement within the community, and can be superseded as new experimental results come in.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 prevailing view of the field as presented in the scientific literature.

Thus, the task of biomedical data curation goes beyond fixing defects in data (although this is part of the curator's task). Instead, curation must be done by human experts in the domain of the data, who are capable of interpreting the scientific literature, resolving conflicting interpretations, and reflecting the results in the data. The curation task is time-consuming, and it is not always easy to recruit curators with the breadth and depth of expertise to be able to do the job well. The speed of arrival of new experimental results, and new interpretations of these and past results, easily out-paces the amount of curation that can be done by available curators (Baumgartner, Jr *et al.*, 2007). Data curation is a vital task, but one that must be done (and done well) with a fraction of the resources needed to complete the work wholly manually.

This has led curators and researchers to examine and propose ways to speed up and improve the curation process. A review of the biomedical literature for the last 5 years (2012-2017) indicates a number of publications proposing tools for data curation (such as OntoMate (Liu *et al.*, 2015), PubTator (Wei *et al.*, 2013), MIntAct (Orchard *et al.*, 2013) and Data Tamer (Stonebraker *et al.*, 2013)), as well as others describing specific approaches to curation in certain fields (such as using a graph-based approach to improve detecting problems in records (Croset *et al.*, 2016), and proposing a middle layer to unify curation results (Sernadela *et al.*, 2015)).

These improvement efforts are good news for biomedical science. However, individual communities are at different stages in terms of how their curation is performed. Some communities are well established, with documented, agreed-upon processes for data curation and access to a repertoire of curation resources, such as rich ontologies defining agreed shared vocabularies. Others are just starting out, and are following *ad hoc* procedures with few quality controls. This is often the case, for example, when some new experimental technique is developed; it can take a little time before community repositories for storing the results can be created, and for the needs of the communities using the new data to be understood and supported. During this period, curation of data is less of a focus for the burgeoning community than just getting up and running. These communities need a quick and efficient way to introduce curation regimes to protect and amplify the value of this early data.

At present, there is little general advice for curators of biomedical data. An exception is a 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 tasks. 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.

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

Such a maturity model will need the support and assistance of the biomedical community to refine and test. As a first step, we present in this paper our initial version of a maturity model for biomedical data curation. The model was created from a survey of the research literature on biocuration, and in this first version is focussed on literature-based curation. The paper is organised as follows. We begin by surveying the literature on biomedical data curation (Section 2) and on maturity models (Section 3). We then present our tentative Biomedical Data Curation Maturity Model (Section 4) and illustrate its use with an example (Section 5). Finally, we conclude.

2 STATE OF THE ART: DATA CURATION

Scientific data curation[1] is the process of associating semantic information with experimental results, to describe their interpretation in terms of current scientific thought. The semantic information typically takes the form of terms from a controlled vocabulary or ontology, or of links with other databases. In addition to adding new annotations, curators are responsible for the overall quality of the data, including resolving defects reported in the experimental data and in the interpretation annotations added previously, or by other curators. The process is expensive, as we have mentioned, because it is important that any such interpretive data is supported by the available scientific evidence. It is also important that these annotations are as complete as possible, so that data-driven science performed on them produces useful results. Some communities/source owners are able to employ experts to work as full-time data curators, while others must rely on volunteers from the community giving their time and knowledge. Because of this, some communities have created their own specialist processes and tools to try to increase the efficiency and accuracy of data curation.

Recent years have seen an increase in the number of publications presenting research in the area of biomedical data curation. While different aspects of curation are covered by different proposals, all share the same goal of improving the outcome of the curation process, while using the same or fewer resources.

Some authors and communities have attempted to make collaboration between curators easier, to avoid overlapping curation work and to make better use of the curation effort available. Orchard *et al.* initiated a project called MIntAct (Orchard *et al.*, 2013), which merged the IntAct Molecular Interaction database[2] with the MINT database of verified protein-protein interactions[3]. MINT is manually curated by experts from the scientific literature. The MIntAct project focussed on sharing the curation efforts from 11 different databases, to gain the maximum value from the curation work performed at each individual source. Thinking along similar lines, Ravagli *et al.* (2016) created OntoBrowser, an on-line collaboration tool for curators, that allows them to work on a single shared working copy, to avoid redundant curation work. Campos *et al.* (2014) also created a curation tool, called Egas, that allows for real-time collaboration curation from the scientific literature.

Others have implemented tools to speed up data curation by automating aspects of the search for relevant papers in the scientific literature to curate, and the extraction of data from the literature. Liu *et al.* (2015) proposed OntoMate, a text annotation tool that tags abstracts of PubMed articles with terms from 16 ontologies using machine learning. The curators can specify query terms and Ontomate returns the abstracts with matching tags. Ontomate will also filter and rank the resulted papers. Wei *et al.* (2013) implemented PubTator, a web-based tool that also searches for articles in PubMed, retrieves them and adds annotations in order to ease the curators' job. PubTator allows curators to select articles from the list of search results, indicates whether the article is curatable or not, and add specifications to data type and relations.

One of the most important procedures in data curation is adding annotations to the curated data. Verspoor *et al.* (2013) propose a schema for representing annotations describing human genetic variants and their relation to disease. The schema was designed for a specific community but is intended to be more widely used, and to save curators the need to redo the design work when creating a data format for their annotations. Generally, the schema works as a fundamental stage for those, who look for text mining solutions in human variome. More generally, Goldberg *et al.* (2015) emphasised the importance of providing linked annotations between resources. Their claim is that such links assist manual curation, since they give curators ready access to all (or most) available resources that are connected with the artefact under curation.

While the literature contains a variety of proposals to improve the way data curation is carried out, most of the work concerns or serves a specific target community or data source. We were not able to find much in the way of guidance for setting up a data curation activity, nor much work giving general guidance applicable across the biomedical field. We propose our maturity model in an attempt to (partially) address this gap.

3 STATE OF THE ART: MATURITY MODELS

Maturity models grew out of work in the 1980s and 90s on business process improvement (e.g., Crosby's Quality Management Maturity Grid (Crosby, 1979)) and, especially, software engineering (e.g., the Capability Maturity Model (Paulk *et al.*, 1993)). Since then, a variety of different maturity models, covering a variety of different fields and process types, have been defined.

Briefly, a maturity model is a sequence of levels or stages that show the progress needed to reach a mature level of practice in some tasks or areas (Paulk *et al.*, 1993). Each level has specific criteria that need to be fulfilled in order to move from one level to the next. For example, one of the most well-known and well-used maturity models, the Capability Maturity Model for Software, aims to model maturity of software development processes (Paulk *et al.*, 1993). It consists of five levels:

1. **Initial** The software process used by teams at this level is characterised as *ad hoc*, and occasionally even chaotic. Few processes are defined, and success depends on individual effort.
2. **Repeatable** Basic project management processes are established to track cost, schedule, and functionality. The necessary process discipline is in place to repeat earlier successes on projects with similar applications.
3. **Defined** The software processes for both management and engineering activities are documented, standardised, and integrated into a standard software process for the organisation. All projects use an approved, tailored version of the organisation's

[1] `www.dcc.ac.uk/resources/curation-lifecycle-model`

[2] `www.ebi.ac.uk/intact`

[3] `mint.bio.uniroma2.it`

standard software process for developing and maintaining software.

4. **Managed** Detailed measures of the software process and product quality are collected. Both the software process and products are quantitatively understood and controlled.
5. **Optimizing** Continuous process improvement is enabled by quantitative feedback from the process and from piloting innovative ideas and technologies.

Maturity models have several uses (Pöppelbuß and Röglinger, 2011a). They can be used to assess the current level of a group, in order simply to understand how the group is performing relative to the norms in the field. They can be used to compare the performance of two different groups (for example, to look for opportunities for partners for fruitful interactions and discussions — a group may find it more useful to work with a partner one level higher than it in the maturity model than with one at the other extreme of the model). Principally, however, they are a tool for long-term, sustained improvement. By assessing a group's current standing against the model, and comparing this with the group's desired level, a sequence of manageable improvement actions can be planned. With the model's help, the group can target its efforts on areas of its performance where there is most scope for useful improvement. And by looking at the criteria for performing at the level just above it's current performance, achievable improvement steps can be identified, that can be implemented with the resources available.

A large number of maturity models have been proposed, since their inception in the 1980s. A full survey is beyond the scope of this paper, but we mention some representative examples of work in this area, to give a flavour of what is being done.

New maturity models have been proposed in areas that go well beyond the original business process/software process focus of the earliest models Ofner *et al.* (2015), for example, built a maturity model for data quality management at an enterprise level. Another proposed maturity model is called the Student Engagement Success and Retention Maturity Model (SESR-MM) (Clarke *et al.*, 2013). It focuses on helping higher education institutions (HEIs) to provide a good environment for their students. The model covered different aspects that can raise the level of student engagement to improve academic success rates and retention. Yet another model is aimed at innovation capabilities within organisations, and the kinds of support and facility needed to enhance it (Essmann and Du Preez, 2009).

In addition to these business focussed models, a handful of maturity models in the area of scientific data and data management have been proposed. For example, Bates and Privette (2012) proposed a maturity matrix for the quality assurance processes used in managing climate data records. Specifically, the model looks at whether best practice is employed in the task of converting the raw experimental data into a high-quality product. Crowston and Qin (2011) proposed a model based on the CMM for Software but adapted for the management of scientific data. They describe key processes and practices that should be in place for effective data management. A further example is provided by a team at Sandia National Labs in the US, where Oberkampf *et al.* have constructed a maturity model for computer modelling and simulation (Oberkampf *et al.*, 2007). The model includes a check on the tools and techniques used to verify the geometric and physical fidelity of any model created.

Other researchers have studied the whole concept of maturity models, and have proposed ways in which new maturity models can be created and for making better use of existing models. For example, the Institute of Internal Auditors, in the Netherlands, offers a guide for selecting maturity models for use on business process improvement consulting projects The Institute of Internal Auditors, 2013. The guide contains a description of a maturity model, and illustrates how to design a maturity model. Pöppelbuß*et al.* (2011b) focused on investigating the literature of maturity models in business process management. From this investigation, they derived them some general design principles that can help in designing a maturity model.

4 BIOMEDICAL DATA CURATION MATURITY MODEL (BIOC-MM)

In order to create a maturity model for biocuration, it was necessary to gain a picture of the breadth of activity being undertaken (to identify the dimensions for our model), and to gather examples of best practice across different biomedical domains. In order to do this, we reviewed the literature on curation activities in five different biomedical databases, covering a spread of topics across the field:

- UniProt[4]
- BioGRID[5]
- FlyBase[6]
- Saccharomyces Genome Database (SGD)[7]
- Rat Genome Database (RGD)[8]

We also aimed to examine sources from both long-established and more newly established communities, on the grounds that the longer established communities would (typically) have more mature processes in place than those just getting started. (Unfortunately, very new communities are not usually in a position to publish details of their curation processes, and are less likely to have the time or confidence to do so.)

According to our observations of practices in use with these data sources, we found that the curation process mainly takes two forms: data-oriented curation and literature-oriented curation. The data-oriented curation means that the focus of the curation process is to look for defects in the data, whereas literature-oriented curation means curating data when a new related publication appears in the area, by extracting relevant information from the paper and associating it with the data. The literature-oriented curation has three main tasks: searching for new publications, extracting data from the abstract and extracting data from the full-text.

These observations led us to divide our Maturity Model (1) into five components as follows:

1. Adding and editing repository data.
2. Searching for and selecting from new literature.
3. Reading and extracting data from the abstract.
4. Reading and extracting data from the full paper.
5. Documenting curation results.

[4] `uniprot.org`

[5] `thebiogrid.org`

[6] `flybase.org`

[7] `yeastgenome.org`

[8] `rgd.mcw.edu`

We identified 5 broad levels for the maturity model from the literature. At level 1, all curation is performed manually (as might be the case, for example, in a community that is new to curation). Then, the process gradually changes to adopt semi-automated ways to curate data. The final level is not full automation, which is not likely to be possible (or desirable) in the foreseeable future due to the need for expert interpretation and decision making, but instead aims for an optimal distribution of work between the human experts (curators) and the supporting software tools.

The provisional model is presented in Table 1. We now describe each dimension (column) of the model in turn.

Adding and editing repository data: This dimension model levels of practice in finding defects in the repository data and correcting them. At the initial level, the curators perform their job by manually searching for defects in data and fixing them. End users may also report data errors, too. At this level, we do not pay attention to the format of data, as manual curation can deal flexibly with a range of formats, to identify how to access and retrieve the data.

Level 2 focuses on making the curation process more organised and repeatable compared with the initial level, as in this level a number of guidelines to define the process of curation and the things that curators should consider to find defects in data are documented. In addition, curators are asked to add an audit trail when making changes to data, giving the reason behind the decision to make the change.

However, curators need semi-automated or automated ways to help them cope with the rapid arrival of new experimental results needing curation. This leads to level 3, in which automatic or semi-automatic tools that can detect defects in data and suggest solutions for the detected defects are adopted. The curators can monitor the results of the tools (perhaps through some dashboard) and authorise changes if applicable.

The next level, level 4, starts from the idea that a number of communities may be working with the data under curation, meaning that multiple curators might be at work on the data. This leads to the possibility of redundant curation being done. At this level, therefore, we look for some support for collaboration between communities of curators. This can be achieved by providing a common curation platform or provide a sharing mechanism. For example, MIntAct proposed a curation platform which allows 11 different databases to share their curation efforts (Orchard *et al.*, 2013). In case of sharing data, it is important to provide a catalogue that standardises the annotations to be created by all communities. This will help curators to be familiar with the meaning of other communities' annotations.

In level 5, all automatable parts of the process *are* done automatically, including the creation of links between data items in the curated sources, and links to relevant external sources.

Searching for and selecting from new literature: This dimension is concerned with the first step in literature-oriented curation, the identification of the scientific papers that will be the subject of the curation. At level 1, searching for new publications in a specific area is done manually by searching with existing publisher web resources. At level 2, semi-automatic tools are used to check for the arrival of new publications and provide the results. At level 3, tools will also be used to rank papers in order of significance or urgency for curation, and will include some notion of paper quality and readiness for curation (e.g. using tools such as the MiniRECH reporting quality checklist[9]). At level 4, the tools would include some element of learning, based on curators decisions about what to curate previously, that removes some of the search labour for curators. Searches would be run automatically, rather than being triggered by the curators, and work is scheduled across available curators, who are notified of the arrival of papers relevant to them that could be curated. At level 5, text analysis of the paper is used to make good quality decisions as to which papers to curate, leaving curators only the task of choosing from amongst a very small number of papers.

Reading and extracting data from the abstract: This dimension relates to the second step of literature-oriented curation, in which annotations that are supported by the abstract of the paper under curation are decided. At level 1, curators read and extract data from the abstract entirely manually. At level 2, the curation process continues to be manual, but authors of the paper can participate in the process. In other words, authors are given the chance to fill in a form with some information about their publications. At level 3, a semi-automatic tool can be used to highlight and extract data from the abstract. However, at this point, only limited formats of abstract will be covered.

At level 4, tools will support the curator by looking for specific features in the abstract, based on a specification of needs from the curator. For example, specific protein interaction information could be located in the text of the abstract. At level 5, the tool will learn from previous interactions what data needs to be extracted, meaning that the curator does not need to do much configuration of the tool.

Reading and extracting data from the full-text: After extracting data from the selected publication(s), the paper need to be curated in full — that is, the full text of the paper is examined for information relevant to the annotation task. As in the other dimensions, the curation process of the full-text is done manually at level 1. At level 2, the curation process can be assessed using a tool such as Kwon *et al.*, 2014. At level 3, collaboration and sharing tools are brought into play, to assist curators in working together to curate a set of papers, sharing information and avoiding redundant work. For example, one curator might mark up the relevant phrases in a paper, and this markup would be visible to other curators. This collaboration can be done by providing curation platform.

At level 4, we start to use tools that extract relevant information from the paper full text automatically (creating the kinds of mark-up that curators create at level 3, but by software rather than manually). At level 5, the tools used must go beyond extracting data from the text of a paper, but will also highlight relevant figures and tables. Besides, supplementary materials will also be considered and processed for relevance.

Documenting curation results: This dimension focuses on recording and displaying the curation results, which might help curators from varies communities to understand the curation process of a specific community. In level 1, any documentation of curation results is done manually, and at the discretion of individual

[9] `github.com/miniRECH`

curators. At level 2, a semi-automatic tool can be used to highlight recent changes made to data items of interest to curators and end-users, but audit trail information is gathered manually and informally. At level 3, the capture of audit trail information will be documented and standardised across the community, with tools to assist in the capture of this information. At level 4, audit trail information will not only be captured, but will be displayed and be capable of being queried. At level 5, tools will be able to aggregate audit trail information across a data source or set of curators, providing graphs for each attribute and divide the results by change type and reason. This information will be used to identify lapses from the documented curation process, and to advise on areas where more curation effort is needed.

5 USAGE OF OUR PROPOSED MATURITY MODEL

This section illustrates how our proposed Maturity Model might be used in practice, by describing an example. In this example, a community that has only recently started to curate its data wishes to make improvements. They will use BioC-MM to identify possible "quick wins" for improvement, based on their current practices.

The community needs to carry out the following steps:

1. Identify the current maturity level of the community curation process against each dimension in the model.
2. Identify the dimensions where improvement is most needed, and select the desired maturity level of each one. The desired maturity level should be close to the current level for this exercise. The assumption behind the use of maturity models is that there is no point in trying to jump from level 2 to level 5 (say) too quickly.
3. For each dimension where improvement is needed, use the descriptions of the levels between the current level and the target level to plan a series of staged improvements.

Let's consider a simple example of a community that wishes to use BioC-MM to improve its processes. Assume that this community uses a tool downloaded from elsewhere to extract new publications from the literature every week, and that it can semi-automatically detect and extract data from the abstract using a bespoke tool they have developed. The community uses a basic collaboration platform, to curate the full text of new publications. However, the repository data is still edited manually, and no audit trail information is gathered (apart from notes kept informally by curators).

Based on the description of the community mentioned above, this community is at level 1 for dimension 1, at level 2 for dimension 2, at level 3 for dimension 3, at level 3 for dimension 4, and at level 1 for dimension 5. The curators feel they are spending too long searching through new publications to find the ones they need to pay attention to, and are beginning to struggle with the lack of any formal audit trail, as errors introduced by inexperienced curators are hard to detect and correct. So, the goal is set to reach level 3 in dimension 2 and level 2 or 3 in dimension 5. Interest is also expressed in making data changes easier, so a target of level 2 is set for dimension 1.

After deciding the target maturity levels, it is time to go through each dimension which is below its target, to improve it. Dimension 1 should be moved from manually editing repository data to semi-automatic editing. If no existing tool can be found, then a bespoke tool will need to be created. The team might decide that this is not cost-effective for them at the present time. To reach level 3 in dimension 3, the community needs to find a tool that can extract relevant information from the abstracts of paper. They find a suitable text mining tool, but need to put some effort into configuring it to work with their preferred ontologies. The team has access to text mining expertise, and decide to go ahead with this improvement.

The last dimension to be improved is dimension 5. The team decides to jump 2 levels, since they realise that they can adapt an audit trail model from another closely related community, and also make use of tools provided by that community. The maturity model has helped them to make informed and defensible decisions about how to obtain the most improvement value from the available resources.

6 CONCLUSION AND FUTURE WORK

The main goal of this paper 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

Bates, J. J. and Privette, J. L. (2012). A maturity model for assessing the completeness of climate data records. *Eos, Transactions American Geophysical Union*, **93**(44), 441–441.

Baumgartner, Jr, W., Cohen, K., Fox, L., Acquaah-Mensah, G., and Hunter, L. (2007). Manual curation is not sufficient for annotation of genomic databases. *Bioinformatics*, **23**(13), i41.

Campos, D., Lourenço, J., Matos, S., and Oliveira, J. L. (2014). Egas: a collaborative and interactive document curation platform. *Database*, **2014**, bau048.

Clarke, J. A., Nelson, K. J., and Stoodley, I. D. (2013). The place of higher education institutions in assessing student engagement, success and retention: A maturity model to guide practice.

Crosby, P. B. (1979). Quality is free: The art of marketing quality certain. *New York: New American Library*.

Croset, S., Rupp, J., and Romacker, M. (2016). Flexible data integration and curation using a graph-based approach. *Bioinformatics*, **32**(6), 918–925.

Crowston, K. and Qin, J. (2011). A capability maturity model for scientific data management: Evidence from the literature. *Proceedings of the American Society for Information Science and Technology*, **48**(1), 1–9.

Essmann, H. and Du Preez, N. (2009). An innovation capability maturity model–development and initial application. *World Academy of Science, Engineering and Technology*, **53**(1), 435–446.

Goldberg, T., Vinchurkar, S., Cejuela, J. M., Jensen, L. J., and Rost, B. (2015). Linked annotations: a middle ground for manual curation of biomedical databases and text corpora. In *BMC Proceedings*, volume 9, page A4. BioMed Central.

Kwon, D., Kim, S., Shin, S.-Y., Chatr-aryamontri, A., and Wilbur, W. J. (2014). Assisting manual literature curation for protein–protein interactions using bioqrator. *Database*, **2014**, bau067.

Component	Level 1	Level 2	Level 3	Level 4	Level 5
Adding and editing repository data	Manually identify problems in the data records and fix them	- Define criteria to go through each data record and fix data - Adding annotations when editing data (manually)	Semi-automatic tool to detect problems in data and suggest solutions to fix problems. The curator can then go through suggestions and authorise the ideal suggestion	- Providing a catalog that link all types of annotations - Collaboration and Data Sharing providing a common curation platform to share curation efforts between databases	Completely automated way to detect and fix problems in data
Searching and choosing for new literature	Check for new publications in the literature manually	Semi-automated tool to search for literature	The tool can rank and order the extracted literature	Set the tool to work every specific period of time, and search in different sources of literature	Totally automated way to search literature and split the extracted papers by type
Reading and extracting data from the abstract	Reading and extracting data manually	Collaboration allow the authors of new publication to participate partially in the curation process	Semi-automated tool to highlight and extract	The tool can also semi-automatically find protein-protein interaction and relationship	The tool can perform its job automatically
Reading and extracting data from the full-text	Reading and extracting data manually	A tool to asses manual curation	Collaboration collaborative curation platform between communities and curators	A tool to extract data from text semi-automatically	Extend the tool, so it covers tables, figures etc. At least point out if it has something that need to be reviewed
Documenting Curation Results	Does not pay attention for documenting any results	A semi-automatic tool to help in extracting results of the curation for a specific type of data	The tool has extra feature such as specifying the period of time	The tool will display the reason	A tool to analyse the curation results

Table 1. Biomedical Data Curation Maturity Model

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.

Oberkampf, W. L., Trucano, T. G., and Pilch, M. M. (2007). Predictive capability maturity model for computational modeling and simulation. Technical report, Sandia National Laboratories.

Ofner, M., Otto, B., and Österle, H. (2015). A maturity model for enterprise data quality management. *Enterprise Modelling and Information Systems Architectures*, **8**(2), 4–24.

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., Curtis, W., Chrissis, M., and Weber, C. (1993). Capability maturity model, version 1.1. *IEEE Software*, **10**(4), 18–27.

Pöppelbuß, J. and Röglinger, M. (2011a). What makes a useful maturity model? a framework of general design principles for maturity models and its demonstration in business process management. In *ECIS*.

Pöppelbuß, J. and Röglinger, M. (2011b). What makes a useful maturity model? a framework of general design principles for maturity models and its demonstration in business process management. In *ECIS*.

Ravagli, C., Pognan, F., and Marc, P. (2016). Ontobrowser: a collaborative tool for curation of ontologies by subject matter experts. *Bioinformatics*, page btw579.

Sernadela, P., Lopes, P., Campos, D., Matos, S., and Oliveira, J. L. (2015). A semantic layer for unifying and exploring biomedical document curation results. In *International Conference on Bioinformatics and Biomedical Engineering*, pages 8–17. Springer.

Stonebraker, M., Bruckner, D., Ilyas, I. F., Beskales, G., Cherniack, M., Zdonik, S. B., Pagan, A., and Xu, S. (2013). Data curation at scale: The data tamer system. In *CIDR*.

The Institute of Internal Auditors (2013). Practice guide: Selecting, using and creating maturity models: a tool for assurance and consulting engagements.

Verspoor, K., Yepes, A. J., Cavedon, L., McIntosh, T., Herten-Crabb, A., Thomas, Z., and Plazzer, J.-P. (2013). Annotating the biomedical literature for the human variome. *Database*, **2013**, bat019.

Wei, C.-H., Kao, H.-Y., and Lu, Z. (2013). Pubtator: a web-based text mining tool for assisting biocuration. *Nucleic acids research*, page gkt441.