Community Curation for GeneView

Studienarbeit

<table>
<thead>
<tr>
<th>Project</th>
<th>GeneView</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervisor</td>
<td>Prof. Dr. Ulf Leser</td>
</tr>
<tr>
<td>Author</td>
<td>Alexander Konrad</td>
</tr>
<tr>
<td>Date of issue</td>
<td>September 2014</td>
</tr>
</tbody>
</table>
1. **Motivation**

The latest discoveries of diseases and their diagnosis or treatments are mostly published in scientific literature. The fast growth of published biomedical articles led to a strong ambiguity of disease names meaning a traditional keyword-based search for biomedical articles will not lead to satisfying results [DL12]. This problem does not only exist for the terms of diseases, it is a problem for most names of biomedical objects like genes or chemicals too [TSV+12]. As long as data grows exponentially, novel Bio-medical Informatics approaches and tools are needed to retrieve the data [FMM07]. Efficient search tools are crucial for biomedical researchers to keep abreast of the biomedical literature relating to their own research [DM+09]. GeneView is such a tool. It was created as a comprehensively annotated version of PubMed articles and abstracts. PubMed\(^1\) is perhaps the most popular information retrieval tool in biomedicine based on MEDLINE\(^2\) repository [S10]. GeneView is a semantic search engine for PubMed using a multitude of state-of-the-art text mining tools for recognizing instances of ten entity classes and protein-protein interactions (PPI) [TSV+12]. GeneView provides possibilities to search for scientific biomedical articles with a number of features.

Advanced systems like GeneView use automatic methods to annotate biomedical articles automatically. But the precision and recall of the used algorithms still leave plenty of non-detected (false-negatives) or wrongly detected (false-positives) entities. As example the gene name recognition uses the tool GNAT which has a precision of 82% and a recall of 82% for abstracts and 54/47% respectively for full articles. To close the gaps of not detected entities a manually curation of the results is necessary. Leitner et. al show in [LCA+10] that annotations made by systems and manually achieved annotations through authors or curators could assist each other to improve the overall performance. Currently there is no possibility to manually curate annotations in GeneView.

GeneView's web front-ent visualizes articles/abstracts with entity highlighting. Highlighting annotated entities is a clever way showing up interesting links.

---


between entities. This could be extended by having all information regarding one entity, across all relevant articles, in one place. Many knowledge resources are compiled by manually curated knowledge extracted from biomedical literature and other sources [BB08]. As GeneView has a large repository of information about entities this knowledge can be examined through creating a new view on the data. Both the need for curation and the desirable new view on the existing data go hand in hand. Having all information about one entity at one place could help the process of curation and having a curated set of data leads to a huge benefit of the quality of the annotations. This work is dedicated to model and implement this new view in GeneView next to developing a model for community curation for annotated entities.

2. **GeneView**

This work is based and will be implemented on GeneView. GeneView is a semantic search engine for the Life Sciences. GeneView identifies instances of ten different entity types and two relationship types and indexes approximately 21.4 million abstracts and almost 360,000 full texts. To identify these entities and relationships GenView uses a group of state-of-the-art tools. The idea of GeneView is to bundle the best available algorithms into a complex pipeline to analyse every article with different methods to benefit from all individual algorithms. GeneView is updated regularly pulling new articles by PubMed. All entities and relationships extracted by the pipeline are stored in a relational database while the articles (or abstracts) are separately stored in Lucene\(^3\) which serves as storage, query and ranking engine. Aggregated information for each entity type is also stored in Lucene after the pipeline has finished [TSL12]. Users can access GeneView through a web interface and query the Lucene index to search for articles. Once a user has sent its query, GeneView provides a list of matching articles sorted by the ranking facilities of Lucene (by default this is the publishing date). Each article is shown as a short excerpt with additional information such as title, authors, released magazine, publishing dates and other

\(^3\) [http://lucene.apache.org/core/](http://lucene.apache.org/core/)
facts specific to the article. A search for a specific entity will be shown like that as well as a free text search. A click on an article leads to the full text version respectively the abstract. The display of result of a search query is document-centric and provides a list of relevant articles from where the user has to pick the articles which look interesting.

GeneView provides no possibilities to gain all information about an entity on one site. A way of displaying information cumulated is currently not implemented. All entities stored in the relational database corresponding to the article are highlighted in a specific colour. The highlighting is achieved through different layers, each containing highlighting for precisely one entity type. Each layer is an HTML representation of the article. For each entity additional information is provided through a small opening pop-up window when the entity is clicked on. The pop-up shows links to external reference databases and provides more services in the case of proteins or genes. For these entities the pop-up window shows information on pathways and protein-protein interactions the gene/protein participates in. This specific information is aggregated in the background and requested through an AJAX call. Currently there are no curating possibilities provided for GeneView [TSL12].

3. **Community Curation**

The biological data is growing massively and is asking for new models of storage and distribution [PKI+08]. There could be an answer that will scale with the challenge: community curation. To realize curation and annotation for the fast growing amount of data the concept of community curation was developed. This chapter discusses the necessity and definition of curation. It will be pointed out where the difference to annotating is. Further the current situation in the biggest scientific field where community curation is applied, the bioinformatics, will be discussed. It will be pointed out why this scientific field is predestined as leading application of annotation and curation. The benefits and problems of community curation is another key topic in this chapter.
3.1 What is community curation

Curation has to be distinguished from annotation. Both expressions are very often used in the same context as they are related to each other in practice although they mean different things.

Curation is the process of identifying, collecting, analysing, determining the information and producing meaning out of data. Data can arise in various forms. It can be published as written documents or as data in a more structured way like results from experiments. It is the task of curation to gain information out of an undefined range of published knowledge or results [NL12]. Curators are often experts in their domain as all activities of curating demand a certain level of content expertise [RDH+12].

Annotation is a process of tagging an article [NL12]. It can be undertaken for a variety of reasons including description, correction, classification, reviewing, interpretation, and augmentation of the data. Curation and annotation are related to each other. Both have the objective to add information and provide structure. Curation without annotation can be difficult due to missing additional information. Annotation without curation can be automated but misses human selection and combination of different information.

Community describes a voluntary association of a large number of users collaboratively contribution to a same type of effort [RDH+12]. Community curation describes a concept allowing users to participate in annotating and curating scientific data. Curation is in these communities their raison d’être [RDH+12].

Bioinformatics has special requirements to curation and annotation. Data such as nucleotides and protein sequences, protein crystal structures, gene-expression measurements, protein and genetic interactions have to be located within published articles and have to be annotated or extracted. There are controlled vocabularies necessary as well as a reporting-structure standard and for example approved gene symbols. As each article uses different methods and ways to illustrate results, extracting this information requires a great effort. [HCG+08].

To cover the increasing data with annotation and curation methods it is necessary to introduce community based models. Due to the speed at which
humans are able to curate there is a growing gap between the published and the curated information. As especially experimental technology raises its output in an exponential manner why this gap is emerging dramatically [LMK10]. Sticking on the currently existing models without community will increase this gap. Huss III et al. state that "the research community will need to be involved in the annotation effort to scale up to the rate of data generation" [HLM+10].

The growth of a popular technology is able to influence the way of exchanging research information: The new Web 2.0 technologies and the concept of Wikis introduce new possibilities. This includes the interpretation of research data which is now able to be transformed while distributing among other scientist and will cut away separated and repeated processes every single scientist have to do by his own. [AB+08]. There is as well a demand of exchanging information in an open access format or to share information through independent organisation which provide possibilities for collaborative activities [PKI+08]. Following Altman et al. there is an increasing interest in community-based annotation and corresponding supportive tools [AB+08].

Curation communities are influenced by popular online social communities like Flickr\(^4\) or YouTube\(^5\). These communities are as well built to generate and curate content, but their understanding of curation is quite different to curation in terms of science. Target of these collaborations is to detect popular content, there is no special expertise needed to annotate or curate content in the most cases. Many social communities are focused on their own content and format and raise no claim to completeness [RDH+12].

Some collaborative work communities are widely known like Wikipedia or in more general from open source software development. Every community needs people who build the community. Creating such a community is next to defining a concept of participation a big challenge. It would be possible to increase the number of potential annotators/curators by allowing anyone with a web browser to participate. This would include the general public [HCG+08].

---

\(^4\) https://www.flickr.com/ is an online image hosting platform with possibilities to share photos with a social community

\(^5\) http://www.youtube.com/ is an online video hosting platform providing services to share private videos to other users
Everybody can become part of a community to provide effort necessary to handle the growing amount of data. Starting from professional researchers with fundamental knowledge to students with little knowledge or people from other subjects. As an example the familiar web-based format of WikiPathways greatly reduces the barrier that prevents participation in pathway curation. More importantly, the open, public approach of WikiPathways allows for broader participation by the entire community, ranging from students to senior experts in each field [PKI+08].

The community could provide the necessary power to bear the workload as well as the necessary knowledge distributed along all scientists. The principles of community intelligence can be used as a driver for new systems. The main idea is to exploit the collaborative contribution to synthesize new knowledge [WOB+09]. Community intelligence benefits from the “Long Tail” of users making small adjustments or contributions which are individually small but sum up to larger contributions [MNJ+10].

Community projects like Wikipedia proved the principle concept on broadly oriented topics with strong benefits for the community. Rico et al. sees strong signs during developing their WikiPathways that a lot of scientists are willing to participate [PKI+08]. And it is obvious that many scientist intensively use databases but have no possibility to contribute to them [MNJ+10]. Following Howe et al. this only works if participation is appropriately rewarded. They state that substantial contributions should be taken into account. Further they urge to consider such contributions with the same reputation as peer reviewed publications especially in terms of promotion, salary and hiring [HCG+08]. This could evaluate as an indication that providing such possibilities scientist at least would be devoted to capabilities of such tools. Involving the community always means to step down in terms of control over the databases by the owners. If users join the side of contributors they haven’t to be considered as pure data providers but as fair partners with power of decision making. Therefore technical and logistical barriers must be reduced and tools have to be provided which hold a reasonable balance between administrative principles and user participation [MNJ+10]. But bringing in the community can’t be solely – it can only be a complementary to current non-community systems [HLM+10]. Hybrid solution like Barend Mons advises BioCreative to focus on tasks leading to more efficient
tools for combined computer and community annotation [AB+08] could be an option. A broader look into the future of curation and annotation will take part in chapter 4.5.

3.2 **Usage of community annotation tools**

The concept of community annotation / curation sounds promising. In theory this seems to be a reasonable way to solve the issue of growing data by using the cumulated intelligence. It is questionable if this is this only a theory or if examples applied in practice can be found.

As an example serves the Gene Wiki project\(^6\) where community curation is successfully used. The Gene Wiki project is part of Wikipedia. This project has the goal to create a Wikipedia article on every notable human gene. It is open for the public to participate like all other articles. The success of this project is related with the fact that annotating is not the one and only purpose. Rather than annotation new gene structures it is another important purpose to add descriptive text to attribute functionality to existing gene structures [LG12]. To measure the success of this project Huss III et al. connect the Google Trends information to the page views as shown as example in Figure 1. Google Trends measures search terms over time relative to the whole number of Google searches. The figure shows the linkage between the numbers of views of Wikipedia articles to the amount of Google searches of the same topic. The more Google searches the more Wikipedia views are generated. The spike of Wikipedia page edits at the same time indicates a large user involvement. The spikes of Google Trends in these two examples are due to reports in the popular press. This figure indicates therefore that Gene Wiki is relevant right now to the general public and also to a growing number of scientists [HLM+10].

All projects regarding community curation haven't got a long history and have still potential in developing and there are different ways to approach community participation. Current projects differ in the amount of user participation and in the amount of available content. But why are some projects more attractive to users

---

then others [SD+10]? This is a difficult question to answer as every project is different in its structure. One could argue that this question is impossible to be answered that due to the different topics every community project is dealing with.

![Figure 1: Timelines of Wikipedia views and Google Trends information for HCG (A) and catalase (B). Size of blue dots shows editing activity.](image)

Stehr et al. worked out similarities to community based projects especially considering wiki style projects. They state two key factors when it comes to the user. First the value of community based contribution should be made obvious to instantly gain motivation and to silence all concerns. It should be apparent that the community factor will be a catalyst to scientific effort. Second the project should deliver directly benefits for the users to engage to contribution. The potential benefits for a scientist shouldn't appear only when spending a big amount of time diving into the project; it should be present immediately [SD+10]. If one considers joining a project or even doing some small contributions the benefit must be obvious and immediately paid back.

Stehr et al. also point out that these projects are successful even without traditional mechanisms of reward. [SD+10]. The incentive must still be there; new methods have to be designed as reward is necessary to keep serious scientists involved. It seems to be that scientific tribute or simple personalized recognition
drives scientists to contribute to community curating systems. Beside incentives there are several other ways of getting new people to contribute. Wiki style community curation projects are experiencing the advantage of an already working model. This does not mean they are automatically flooded by users. Other kinds of project have to work even harder to find new users. WikiPathways\(^7\) tries to organize some community curation events where they introduce new users to their system and motivate them starting to contribute. This includes the useful side effect of improving their data quality by working on productive data. For future they will raise this effect of introducing people to the project and coincidently adding annotations [PKI+08].

Another point which is very difficult to change is the image of data curation as it appears for non-familiar users as being a data entry job. But as Burkhardt et al. state this is simply not the case due to complexity of biological databases. Another way WikiPathways follows to encourage users is creating user pages and portals where users can present themselves and respectively their work. Portals conduce to collect users from same interest and focus of work to build up sub-communities [PKI+08].

### 3.3 Wiki based solutions

In this chapter wiki based solutions will be discussed as these solutions have been found in nearly each publication concerning community curating. Regarding curation and annotation there are a number of tools being developed. A lot of these tools are dedicated to support experts annotating a text and partially suit as well for curation [NL12]. In terms of getting the community involved there has to be a different angle of view to almost the same problem. The difference is obvious as not only one user is annotating but many users are annotation or curation on the same text or topic. To get this requirement managed the concept of wiki seems to be the most dedicated one. Other concepts are often strictly controlled by supervisors, follow exact rules and address special areas or use their own ontologies [GSH+08].

\(^7\) [http://wikipathways.org/index.php/WikiPathways](http://wikipathways.org/index.php/WikiPathways)
The term wiki has been used frequently over the past years and was mostly applied to websites using the MediaWiki software which is the software powering Wikipedia. The term wiki also refers to software dedicated to be used by many people working together on the same document or sharing information in-between each other. Wiki software is often public and allows anybody to contribute by adding or changing content. This kind of software has caught the attention from scientists adapting it to community curation and annotation [YBS+12]. These open access-kind of software tend to have a high coverage but lag in standardisation or quality control [GSH+08].

Many community curation projects using MediaWiki as fundamental software built their own add-ons and customizations. These customizations support the users to contribute content. Also they help to generate an standardized format of data. As the most wikis have free-text entries this is one important factor to be considered. WikiPathways for example adds a graphical pathway editing tool and integrated databases covering major gene, protein, and small-molecule systems [PKI+08]. The integration from external and standardized databases is also considered in the most wiki style projects. Many solutions give these integrated sources a special status to flag what is manually contributed and which is automatically contributed. Integrated information can be structured or free text as well [AB+08].

Since annotation is very data focused what means it would be possible to annotate much more information but it wouldn’t help scientists who are not text-mining specialists. Furthermore the large data volumes have to be stored – this needs complex and well-designed solutions without forgetting to provide easy user interfaces for end-users as well as contributors [AB+08].

Another approach is taken by BioGPS[^biogps] a centralized gene portal for aggregating distributed gene annotation resources. BioGPS realized that not all information could be stored in one database and not all scientists want to have the same kind of information. Further they want to bring in the opportunity for users to easily add new sources to the portal and build up their own reports. The user can contribute

[^biogps]: http://biogps.gnf.org
in three ways, first they can add external sites to access through plugins, and second they can write their own new BioGPS-plugins and add them library or third contribute by adding new reports to be used by the community [WOB+09].

There are several concerns when it comes to wiki styled projects. One big concern is the fear of chaos inside a wiki project as the regulation of a wiki is very difficult without destroying its open character and flexibility. Following James Hu et al. “this should be balanced by the realization that individual curators cannot fully encompass the collective expertise of the larger scientific community” [HAB+08]. What means, avoiding chaos by regulations or limitations will avoid to gain new information as well. This information is not yet well structured and is only available in unpublished notes from scientists.

Another point is the amount of users necessary to reach a critical mass. The critical mass is achieved if enough people taking part to keep the project rolling and to encourage new users [HLM+10]. Using Wikipedia as basis, like GeneWiki does, immediately reaches a theoretical critical mass of users by engaging the Wikipedia users.

Following the concept of possible anonymous contribution leads to fears of inaccuracy and bias [HLM+10]. This is closely tied to the demand on a peer review process which is missing. Without any monitoring the quality of entered data or further the qualification of contributors this model would not be ideal for scientists who seek for reliable information [YBS+12].

The less regulated character of wiki styled projects can lead to benefits as at WikiGenes where descriptive text could be added to gene structures to provide additional information [LG12]. WikiGenes describes itself as “the first wiki system to combine the collaborative and largely altruistic possibilities of wikis with explicit authorship.” [WG13]. A deeper look at the trust of community created content is taken in Chapter 3.4.4.

Another argument for wiki styled project is that one could go the other way

---

9 http://www.wikigenes.org/
around and see a wiki as good starting point for deeper research. The initial step of new gathering knowledge could be easily climbed and new scientist could be brought up to speed very quickly. Also unpublished notes from scientist could find their way into wikis otherwise would get lost in individual knowledge [HAB+08].

3.4 Problems

The previous section explained the concept of community curation, use-cases and concrete examples. The promising concept of community curation has some problems which will be described in the following section. As already mentioned the lack of participation is one of the biggest problems. The authority management of building a community is important too. Another serious concern is the quality of the contributions accompanied by the question of vandalism and trust.

3.4.1 Lack of participation

One of the main problems is the lack of participation. Community curation is faced with two main reasons which hold users back to participate.

1. Missing incentives

Why should a scientist participate in a curating or annotating community? What would be his benefit of it? The scientist would enter his own knowledge in a manner it would be useful to other scientist. This would cost additional time and effort without increasing the value of the resource to the contributor. Incentives should promise new information and insights in their own field of research [HCG+08].

Unsupervised communities are struggling in motivation and incentive issues [MNJ+10]. Wu et al. state that a positive feedback loop is required for success of communities. Three components are necessary to establish a feedback loop: scientific utility, community usage and community contribution – where these components are missing community intelligence would never achieve the critical mass of users and activities [WOB+09]. Huss Ill et al. come to the same result that there must be a connection
between content value, content use and content production as community projects are mostly volunteer-driven [HLM+10].

2. Missing reward

A related point is the missing reward. Without getting any personal recognition a scientist would not invest time in contribution his knowledge into a community. As long as there is no improvement in academic reputation or career advancement scientists hesitate to participate [HCG+08]. These concerns may be assuaged if community curation will be considered as an included part of the publication process by academic departments or funding agencies [HCG+08]. Models like GeneTests\(^{10}\), a medical genetics information resource, tightly connects authors with articles allowing them to gain appreciation for their intellectual contributions [YBS+12]. Further codes of conduct on intellectual property must be adapted and developed for community based projects [OEH+10].

### 3.4.2 Process Management

Every community project is faced with the same question of authority: Who is allowed to contribute? The decision on this question could impact the whole project. Allowing anonymous contribution involves the risk of bringing in wrong and non-reviewed information but on the other hand is one of the main drivers for success in terms of the amount of contributing users. Other requirements are monitoring of changes, guidelines how to handle doubtful changes and versioning the content helping to make it transparent. The most projects demand for realizable policies which should be explicit and computationally [GSH+08]. But following Goble et al. “Attention to versioning, privacy, intellectual property protection and security is still rudimentary or neglected in most cases” [GSH+08]. The Research Collaboratory for Structural Bioinformatics (RCSB)\(^{11}\) deals with these questions in the way that all entries are followed by the same annotator from the very beginning to gather a crucial advantage in terms of personal responsibility [BSO06].

\(^{10}\) http://www.genetests.org/

\(^{11}\) http://www.rcsb.org/pdb/home/home.do
3.4.3 Quality

If several people working together on one piece of work coming from different directions and subjects a common ground is hard to find. Quality has different aspects: e.g. contributions in terms of curation can be wrong, annotations boundaries could be set incorrectly, the classification of annotation could be wrong, annotation can be missing. Curated information could also be outdated or don’t match the context (if we think on a flexible free text wiki style models). Quality is hard to measure GeneWiki investigates their logs about creating new entries and editing activity. They observe a shift from creating new articles to editing entries [HLM+10]. The growth of editing activity could be seen as evidence for increased quality of data. As well GeneWiki involves external trusted data sources to keep their entries up-to-date.

Content coverage has to be balanced against content quality [GSH+08]. The standard of curators could be different and therefore they would annotate / curate different things [NL12]. One way to improve the quality is the same as in authority: Give the contributor a kind of ownership and responsibility for the part he is contributing to [GSH+08].

Gobel et al. recognize four categories of curators: Expert curators, self-curators, community curators, automated curators. Whereby expert curators just curate in their distinctive subjects, community curators try to provide knowledge to the whole community, self-curators work in closed areas where only they are allowed to curate and automated curators which are a piece of software which crawl and try to identify submissions from external sources [GSH+08].

To avoid vandalism there are methods to identify vandalism, using the authors name together with several other signals, GeneWiki presents a rule based vandalism protection system with a recall of 90% and a precision of 43% [HLM+10].
3.4.4 Trust

Trust in the content of community curation is quite similar to quality. The question of quality deals with the demand of a high standard and tries to avoid wrong information, trust deals with the reliability of content. How to decide between high reliability and poor reliability? There are some interesting ideas how to prove trust. Adler et al. propose a way to calculate trust at Wikipedia. They take a calculated numeric reputation of the author and all other authors who edited the words close by and derive on this way the reliability of the edited word [ACA+08]. They assume that the reputation of an author has strong correlation with the quality of the author’s contribution. The authors determine reputation by a complex algorithm taking into account a long term of contribution and can’t be changed easily. The method has been evaluated by comparing trust of deleted content which shows that deleted content has a significant lower trust than available text at articles. A word with more trust is expected to have a longer lifetime [ACA+08]. This kind of concept could be transported to every other community curation project to calculate section specific trust.

Another idea is to announce freshly changed data prominently to catch the attention of other curators. At SNPedia\(^\text{12}\), a wiki based project investigating human genetics, exists four level of curation. First, all changes will be reviewed by the SNPedia users – this may be similar to other wikis – second, they introduced semantic templates which flag missing information (e.g. missing chromosome). These flags indicate potential errors for other users as well for the editor himself. Third, they run bots on the entire wiki to add additional information such as chromosome, position, gene and allelic data but don’t destroy any other information. These bots are often the first ones finding wrong data. Fourth, the content is connected to the Promethease\(^\text{13}\) personal genome reports which are well known by the science community which is able to recognize irregularities [CL11]. This last point is again diving into clear

\(^{12}\) http://www.snpedia.com/index.php/SNPedia

\(^{13}\) Promethease is a companion program to SNPedia and builds personal report based on a file of genotypes information in SNPedia
communication and transparency seems to be key feature improving trust. It is valuable to know from where and whom the entries are coming. An expert curator would be more reliable than a casual user especially if it comes to conflicting annotations [GSH+08].

Wikipedia’s ‘No original research’ policy\(^{14}\) has been abandoned at SNPedia. SNPedia actively invite to bring in original research [CL11]. Original research is on the one hand a highly coveted source but on the other hand introduce some space for uncertainty about the trust of it. At Wikipedia this kind of source is not allowed as it is not proved by a third party has therefore a kind of uncertainty. To be fair at this point: The fear of Wikipedia is that non-serious researchers will publish their own results without having any reliable review of it. This risk should be certainty lower within a community of serious scientists.

### 3.5 Professional approach

Besides the approaches to solve the discussed problems there is a wide spread wish to shift towards professional curators. These professional curators could be the missing link between unstructured data entries floating around and a standardization fulfilled by professionals [MNJ+10]. Professional curators could perform quality control on results collected by the community. According to the Transition Factor Encyclopedia (TFe)\(^ {15}\) the long-term goal is to achieve a mixture of expert-curated and automatically populated content [YBS+12]. Following will be discussed how a professional curator looks like, what the requirements are to be a professional and how could they be educated.


\(^{15}\) [http://www.cisreg.ca/cgi-bin/tfe/home.pl](http://www.cisreg.ca/cgi-bin/tfe/home.pl)
3.5.1 Educate Professional Curators

There are new kinds of education in the subject of curation becoming available. For example, the Graduate School of Library and Information Science at the University of Illinois\textsuperscript{16} at Urbana-Champaign offers a “biological information specialist master’s degree” and a specialization in data curation\textsuperscript{17}. They describe their program as: “[..] develop expertise in the curation of research data.”\textsuperscript{11}

The Digital Curation Centre (DCC) offers free half day workshops to teach curation: “A key goal for these workshops is to bring together different communities of practice to share their experiences and to identify where, when and how they can best cooperate to meet data curation challenges” [DCC13].

Howe et al. state that courses in curation should be included in the study of biology as curation became a common activity and should be trained. As well interdisciplinary courses in computer science and information science could be considered as both sciences match together with biology or medicine into a whole set of curation requirements. Further they state that biologist have to shift their way of thinking and need to start to look more systematically and objectively on their data – this must be learned and separately taught as this will become an important part of research [HCG+08].

Are scientists interested in being a curator instead of a researcher? Spending a lot of time in reading and searching for cross-references can be gruelling and less attractive. Klemens Pichler from the European Bioinformatics Institute (EBI) in Hinxton, UK, sees curation as satisfying work and states that as well other scientists would like to be a curator as curators have access to high class scientific material as well to a whole intellectual community. Also the pressure of publishing would be limited and the scope of specialization would be broader [S11].

\textsuperscript{16} http://www.lis.illinois.edu/

\textsuperscript{17} http://www.lis.illinois.edu/academics/programs/specializations/data_curation
3.5.2 Skills of professional curators

A curator needs a whole set of skills. Some of them could be earned as scientist or even at a lower level of education while others are results from self-teaching and job experience. As well competence in the field of computer science will be helpful as databases, system management and even programming will be useful skills to accomplish successfully data curation [HCG+08].

Kyle Burkhardt et al. describe the work from curators as literally never ending job as authors deposit data 24/7, 365 days a year. They have to process new entries, correct them, handle corrupt data, introducing new structures and often beginning from the very beginning what could be taxing. The expertise from curators and annotators can vary in very different experiences, education field and level of education. It starts from Bachelor’s degrees and ends at postdoctoral experience [BSO06].

Sandra Orchard, a senior scientific database curator at the EBI, says accordingly that a PhD is not required although a curator has to have knowledge in biology with at least an undergraduate degree. Also lab experience could be an important skill [S11].

But not only curators go professional, journal publishers must get involved in the curation process and must provide processes linked to curators. There must be an exchange between curators and authors respectively publishers [HCG+08]. Furthermore curators will expand their task to teach curating to authors.

3.5.3 Involving the authors

The community and the professional curators could not care for all problems and cannot handle all kind of data. Authors must be made aware of the importance of providing structured data additional to scientific articles. In 2007
the Federation of European Biochemical Societies (FEBS)\(^\text{18}\), one of the largest organizations in European life sciences, started a survey on authors to find out how authors and curators could work together to reduce workload, increase accuracy and generally ask about the willingness of authors to help curators. The results are that many authors but not all understand the importance of structured information included in or additional to their articles and are willing to cooperate. Further they accept that it would be possible to add author-generated structured annotations within the editorial process. Also the reliability of the curating process could be increased if authors and curators work hand in hand [LMK10].

Udo Hahn, head of the Jena University Language & Information Engineering (JULIE) Laboratory Jena goes even further and not only involves the authors but a broader range of published material. He considers “not just published material from prestigious, peer-reviewed journals but also material scattered over the web, for example, Wikipedia-type sources, conference proceedings, text books, blogs, mailing lists, and so on” [AB+08]. This may be a huge step ahead and bringing in a completely different complexity.

\(^{18}\) www.febs.org/
4. **MAKING GENEVIEW READY FOR COMMUNITY CURATING**

Annotations in GeneView are gained through algorithms as briefly described in chapter 3. There is no possibility for a user to curate these annotations in terms of contributing additional information. The following chapter documents how GeneView has been changed to allow users to contribute additional information to annotations – to curate them.

4.1 **Design**

All annotations that GeneView contains are shown highlighted within the article. Additional to the visual effect it is possible to click on an annotation to view more information about the annotation. A pop-up window opens after clicking on the annotation. Within this pop-up window it seems to be a good place to incorporate community curated information as well as implement a submit panel to curate this annotation. In the next subchapters will be described how curating works in GeneView and how the curations are displayed.

4.1.1 **Allowed users**

It is only registered users allowed to curate. A registration can be requested via email to Philippe Thomas\(^\text{19}\). Users can then login by using the login menu on top of GeneView. To avoid violations and to raise trust to other users, it was decided to make a registration mandatory if a user wants to participate in the curation community. The user-id will be stored against every curation made in a database but won’t get shown to users at the front-end. A non-registered user can still view all curations made by others.

4.1.2 **Curation options**

This work is dedicated to enable curation on current existing annotations within GeneView. Adding new annotations is out of scope of this work.

\(^{19}\) thomas@informatik.hu-berlin.de
Introducing curation on existing annotations is supposed to show the possibilities to change GeneView and open it for community contribution. Add new annotations by users could be the next step to integrate a community into GeneView but is technically sophisticated as GeneView was not designed in the first place to allow user interactions. It is desired to allow users of GeneView to curate annotation within a pre-defined range of options. These options are supposed to avoid different wordings and intensions and give a clear curating structure which can be stored in a database. These options are set to:

1- Wrong annotation
2- Entity wrong
3- Annotation boundary wrong

Figure 2 shows the select form within the pop-up window which opens after a click on an annotation.

All these options are stored in the database by a reference id (please see chapter 6 for technical information) and it is therefore possible to extend the range of options if the current available options appear to be insufficient.

The first option “Wrong Annotation” should be chosen if a user detects an expression which should not be annotated. Since it is possible that the algorithm pipeline of GeneView detects an expression wrongly as entity it is a human task to curate these wrong annotation and report it as wrong.

If an entity is annotated but wrongly categorized for example it is annotated as Gene instead of Chemical the second option should be taken.

The third option “Annotation boundary wrong” is supposed to mark annotations which contain a correct entity but are incorrectly annotated in terms of their boundaries. For example an subsequent letter like the plural “s” is part of the annotation or a concatenated expression is annotated as one entity instead of two.
4.1.3 Highlighting curations in an article

The biggest problem regarding community curating is how to deal with wrong curations. For GeneView it was decided to shift this problem transparently back to the community. Instead of manually validating curations by bringing in an administrator level double checking every entry, the concept of highlighting an annotation is adopted to highlight curated annotations as well. A new sub menu has been implemented in GeneView allowing switching on and off highlighting of curated annotations. Figure 3 shows the new sub menu.
The first option is to highlight all community curations. The highlighting will mark a curated annotation by adding a red line underneath. Figure 4 shows the result after checking the box. In that example the annotations PTEN and CTNNB1 have been curated by a user and are marked. This doesn’t show any detailed information about the curation; it only shows that there is more information available.

Figure 4: Result of “Highlight community curations” button – curated annotations are underlined in red
The second button “Hide wrong annotation” has to be used carefully. After clicking on this button every annotation which got curated by at least one user as wrong will be hidden. “Hidden” in this context means that the original highlighting will be removed and the annotation is shown as normal text. Figure 5 shows the result after hitting this button. In this example CTNNB1 got curated as wrong and gets hidden. Since curation on an annotation is not limited to one user it is possible that several users have curated this entity. It is not possible to override a curation of another user. As there is no control mechanism dedicated to the correctness of curations the concept of visualization has been developed. In the first place annotations which are curated as wrong are still shown highlighted. Only if a user decides to trust the curations of other users and to hide all annotations which are curated as wrong, it is possible to hide them through checking the “Hide..” box. If the user does not decide to hide all these annotation but wants to get notified if a curation has been done for a annotation it is possible to activate the “Highlight community curations” functionality described above. These both possibilities play back the full control to the user how multi-user community curations will be displayed. For future developments it can be possible to develop a kind of supervisor front-end to finally decide if an annotation is correctly curated as wrong and to permanently remove the annotation of the GeneView database but this is out of scope of this work.

Figure 5: Result of “Hide wrong annotations” button – CTNNB1 is no longer highlighted

The last button is used to highlight annotations curated as wrong entity by adding a blue line on top of the annotation. Figure 6 shows the result after checking this box. In this case PTEN got annotated as wrong entity (which is true as it is annotated as chemical but is in fact a gene).
4.1.4 Detail view of curations

After highlighting curations it is possible to see the curations in detail. All curated information is located in the same pop-up window as the curation panel. The pop-up window opens if clicked on an entity. Figure 7 shows an example how curations are displayed. It is basically a list of all curations attached to this specific annotation. The value and the date of curating are displayed. There is no information about the user who did this curation.

![Figure 7: List of curations located at pop-up window](image-url)
4.1.5 Changing curations

It is possible to change the own curation as well as to delete them. If an annotation is already curated by the current user and the user sends a curation value again, the system will ask if the user wants to override the current curation. Figure 7a shows the message after submitting a curation on an already curated annotation.

Figure 7a: Warning message if overriding a curation

It is only possible to curate an annotation with exactly one available option. It is not possible to curate an annotation with different options in parallel. The amount of curations to an annotation by a single user is limited due to the intersection of available options. A “wrong entity” annotation can also be a “wrong” annotation – or a “wrong boundary” annotation could also be curated as “wrong”. It is desired to avoid multiple different curations to the same annotation and force the user to choose the best fitting one. Of course this doesn’t resolve the multi-user problem discussed in the previous subchapter.

It is possible to delete own curations. The list of curations indicates curations of the current user in red. A link below the list provides an option to delete own curations. Figure 7b shows the link and the red marked curation.

Figure 7b: List of curations with option to delete
5. **ENTITY SEARCH BY SENTENCES**

The current search method is keyword-driven and returns a list of articles containing the keyword. If for example someone is interested in a specific gene it can be useful not only having a list of *articles* containing this gene but a list of *sentences* containing the gene. This provides a more detailed overview and can help to decide which article could be relevant for a user.

Therefore a new kind of search has been developed. Following it is described how this new search works in GeneView.

5.1 **How to search by a specific entity**

As described in chapter 4 an entity is highlighted and a pop-up window with additional information can be accessed. At this point a link has been added to run a search for the specific entity. Figure 12 shows the pop-up window and the new link (highlighted). The result of the search is shown in Figure 13. All sentences containing the entity are shown and clustered by article. The user can access the full article with a click on the article title.

![Figure 13: Result of sentence based search](image)
Figure 12: Pop-up window for BRCA1 gene with new link to sentence based search
6. **IMPLEMENTATION**

The implementation affects the front end of GeneView. To be consistent with the current architecture the implementation also involves a new web-service running on an own (virtual) server. The most of the logic and processing is located at this service. This chapter briefly describes how the hardware and software architecture of GeneView looks like and how a new service was built and included to incorporate the requirements.

6.1 **Technical starting point**

The pre-processed annotations is stored in an Apache Lucene database and a MySQL database. Both databases run on different servers. The front-end of GeneView is runs on a third server. Following the current hardware architecture will be briefly described as well as different software pieces building the whole GeneView frontend.

6.1.1 **The hardware architecture**

GeneView’s frontend runs on a webserver unrestricted available for HTTP request connected to the internet. The Lucene Index runs on a different server only accessible within the intranet of the Humboldt-University. This is also the case for the MySQL server which as well sits within the intranet. Figure 8 shows the hardware architecture. Server 1 is part of the Humboldt-University network and hence can connect to Server 2 and Server 3. This setting was to be considered when setting up a new service.

6.1.2 **The GeneView software architecture**

The font-end of GeneView is written in PERL, HTML, JavaScript and uses the Catalyst framework. Catalyst is an open-source framework following the Model-View-Controller (MVC) concept dividing three basic concerns of software development: Access and manipulate data, present data and process user input [HF05]. Communication to the Lucene Index is realised using a XML-RPC protocol based service.
6.2 Technical Approach

Following will be described what had to be introduced on hardware and software side to fit into the current setting of GeneView.

6.2.1 The hardware architecture

It was necessary to install a new server accessible external and sitting within the Humboldt-University network. The used server is in fact a virtual server but physically the same like the GeneView webserver (Server 1). It was decided to make the server external accessible as it opens the possibility to open the service of sentence search for third parties. Figure 9 shows the extended hardware architecture. Server 1.2 is newly introduced and can connect to Server 2 and Server 3 in the same way like Server 1.
6.2.2 The software architecture

The newly introduced service to cover all curating activity is written in PHP programming language. The connection to GeneView is realised by using an XML-RPC client-server library. It would have been possible to use any other service specification like e.g. SOAP but it was decided to use RPC since the Lucene Index is only accessible by using an XML-RPC request and a RPC client was therefore necessary. A service library was therefore already in place and was extended to work as a XML-RPC server as well.

The communication between GeneView and the curating service is reduced to a minimum. Figure 10 shows the MVC implementation necessary to fit GeneView's Catalyst framework requirements. A new controller, model and view have been developed. The most work in GeneView itself does the controller where the logic is implemented which methods to be called. The controller is faced to three possible scenarios: If an article is displayed all curations have to be included, if a pop-up is opened all detailed curations attached to the specific annotation have to be displayed and third to handle the user curation activity. The Curate model does in fact only the
communication to the curating service. The view is chosen by the controller and depends on the three cases explained above. The view itself calls templates to render the data. The templates contain pre-defined HTML and include the given results of the service model.

Figure 10: Process of MVC software architecture

All necessary logic is implemented at the curating service. To store community curations a new database at Server 3 was created. To match a curation to an annotation it is necessary to store the document id (PMID), entity id, entity type, begin and end position within the article in the database table. Figure 11 shows the whole table set up.

Not only the logic is placed at the curating service but also the rendering of results and messages to be integrated into GeneView is done by the service. The rendered results are delivered as valid HTML and JavaScript pattern in the XML-RPC return message. It was more practical to do the rendering at service side and not using the full specifications of XML-RPC to deliver abstract patterns to be rendered at GeneView.
### Table 1: Curations database table

<table>
<thead>
<tr>
<th>Field</th>
<th>Type</th>
<th>Null</th>
<th>Key</th>
<th>Default</th>
<th>Extra</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID</td>
<td>int(5)</td>
<td>NO</td>
<td>PRI</td>
<td>NULL</td>
<td>Auto_increment</td>
</tr>
<tr>
<td>ENTITY_ID</td>
<td>int(7)</td>
<td>NO</td>
<td>NULL</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PMID</td>
<td>int(9)</td>
<td>NO</td>
<td>NULL</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BEGIN</td>
<td>int(5)</td>
<td>NO</td>
<td>NULL</td>
<td></td>
<td></td>
</tr>
<tr>
<td>END</td>
<td>int(5)</td>
<td>NO</td>
<td>NULL</td>
<td></td>
<td></td>
</tr>
<tr>
<td>USERID</td>
<td>int(5)</td>
<td>NO</td>
<td>NULL</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TIMESTAMP</td>
<td>Timestamp</td>
<td>NO</td>
<td>NULL</td>
<td>CURRENT_TIMESTAMP</td>
<td>On update CURRENT_TIMESTAMP</td>
</tr>
<tr>
<td>TYPE</td>
<td>varchar(255)</td>
<td>NO</td>
<td>NULL</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VALUE</td>
<td>varchar(255)</td>
<td>NO</td>
<td>NULL</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ENTITY_TYPE</td>
<td>varchar(255)</td>
<td>YES</td>
<td>NULL</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 11: Curations database table**

### 6.2.3 Implementation sentence search

The implementation of the sentence based search described in chapter 5 is similar to the curating and uses the same communication structure. The curating service also serves as search engine and can deal with incoming requests processing the sentence based search. Figure 14 shows how the service has been integrated into GeneView. A new controller fetches HTTP requests and calls the new `SearchService` model managing the communication with the service outside of GeneView. The results are rendered by the service and directly passed through to the view and the respective toolkit template.
Figure 14: Process of MVC software architecture for the sentence search
7. CONCLUSION

In this work has been discussed how community curation could help to curate and annotate the increasing amount of scientific data. The most common solutions are based on wikis to let users contribute their data and curate them. That is not very surprising in terms of user participations as the probably most famous wiki Wikipedia did a proof of concept of user participation but it is in terms of the requirements of science in general and bioinformatics especially. A wiki is predestined for free text entries, for content which is not structured and doesn’t need a precise structure; while the bioinformatics requires solutions incorporating for example annotation of named-entities, annotation of relationships, standard vocabularies, precisely defined guidelines about the facts to be annotated, pre-calculated annotations, structured output formats as well as a range of input formats allowing to contribute own content easily. Further there is the need of biomedical named-entity extraction, document categorization, customized search engines as well as more general things like support and software licences that suit to the work of scientists. To implement all these special and customized requirements into a community curation system is difficult and a great challenge. A further challenge is building the community considering incentives for scientist. Traditional scientific incentives like reward awareness have to be taken into account but as well could new methods like immediate benefit during contribution can help to encourage users. Concerns within community curation are related to trust and quality, where different approaches have been introduced. There a methods which give an author ownership of an article which adds responsibility and trust regarding quality and validity. Other ideas are about vandalism protection systems taking the editing history of content and the computed trust level of curators into account. Of course there are ordinary review models implemented in some community curation systems. This work showed as well other approaches next to communities or additional to communities like professional curators. Professional curators could work solely on curation and annotation or could be appointed to be supervisors at curation communities.
The practical part of this work implements community curation possibilities to GeneView. As the Community Curating is newly introduced there is no witness that the idea of curation fit the requirement in a productive environment but with that structured design of community curation it faces the most discussed questions, which are: How is the quality of curation – we make it transparent and push it back to the users, can I trust the curation – same answer, we make it transparent and leave consciously the option for contradictive curations, is there a process set up – introducing checkboxes and select fields state an obvious process. GeneView community curation can't offer the same flexibility as wiki style system and is reduced in a way of limiting the possible curations on an annotation to provide at the same time methods to view community curations. This also involves the problem of wrong curations what is solved by pushing it back to the user to interpret what could be wrong and what is useful. Other questions like lack of participation can’t be predicted but the threshold is set very low to start contributing.
8. **Sources**


[S10] Sarkar I. Biomedical informatics and translational medicine. J Transl Med. 2010


[Figure 1] The Gene Wiki: community intelligence applied to human gene annotation, Jon W. Huss, III, Pierre Lindenbaum, et al., 2009
[Figure (2-13) \ 10] Screenshots GeneView, 2014

[Figure 10, 14] MVC Process in GeneView, 2014