SciCura

Shared, General-purpose Curation
of Life Science Knowledge

"A bold new paradigm for how research information is disseminated"

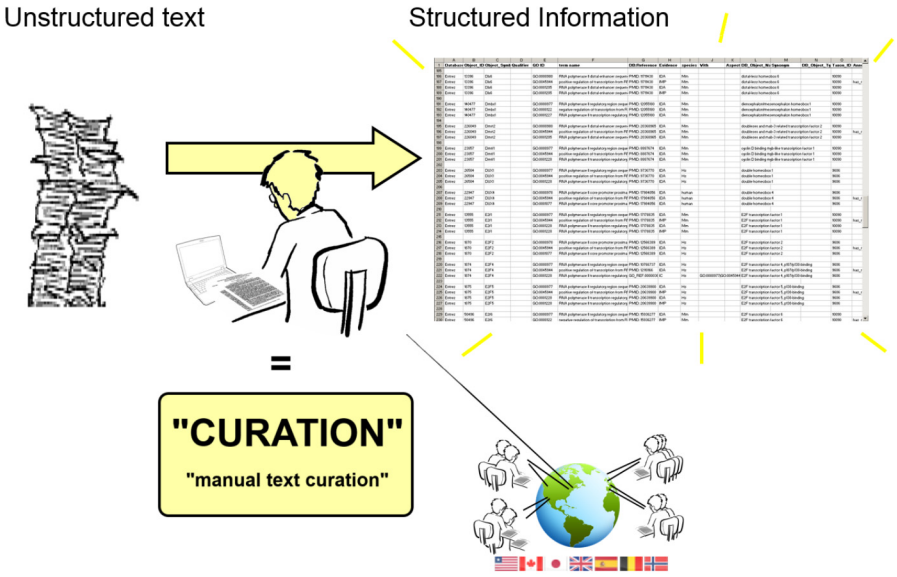
1. The current situation

Most biomedical research results culminate in a publication which only ends up getting buried into an ever growing pile of publications. The content of these papers is functionally invisible for both man and machine: invisible for researchers, because even if you work in a small research niche you should ideally read thousands of papers to know the existing research you need to build upon; and invisible for text-mining algorithms, because information in Life Science papers is too complex for these to comprehend. Only people can deal with detailed nuances of multi-molecular interactions, rich experimental contexts, heterogeneous facts, and often a need to infer what is actually meant.

This is especially important for new approaches using Systems Biology and Systems Medicine, as these fields require holistic and accurate overviews of many molecular interactions and cross-scale observations.

Most people who need some structured overview in their research niche have to create it themselves, and engage, without realizing it, in curation. While reading a selection of relevant papers, they write down a list of useful facts or collect them in a spreadsheet. During a decade of experience with various research labs, we have seen these local curation efforts happen over and over again: e.g. on plant cell cycle and biomass (whole department effort, handicapped by current technology), or gastric cancer network regulation (NTNU research group).

Most of these locally curated facts, however, get lost again and die at the end of a project.



Although institutional efforts by professional curator groups can efficiently carry out semi-automated expert curation strategies, they cannot keep up with the increasing amount of Life Science literature. Over a million new papers get published each year, and only a small percentage of selected information

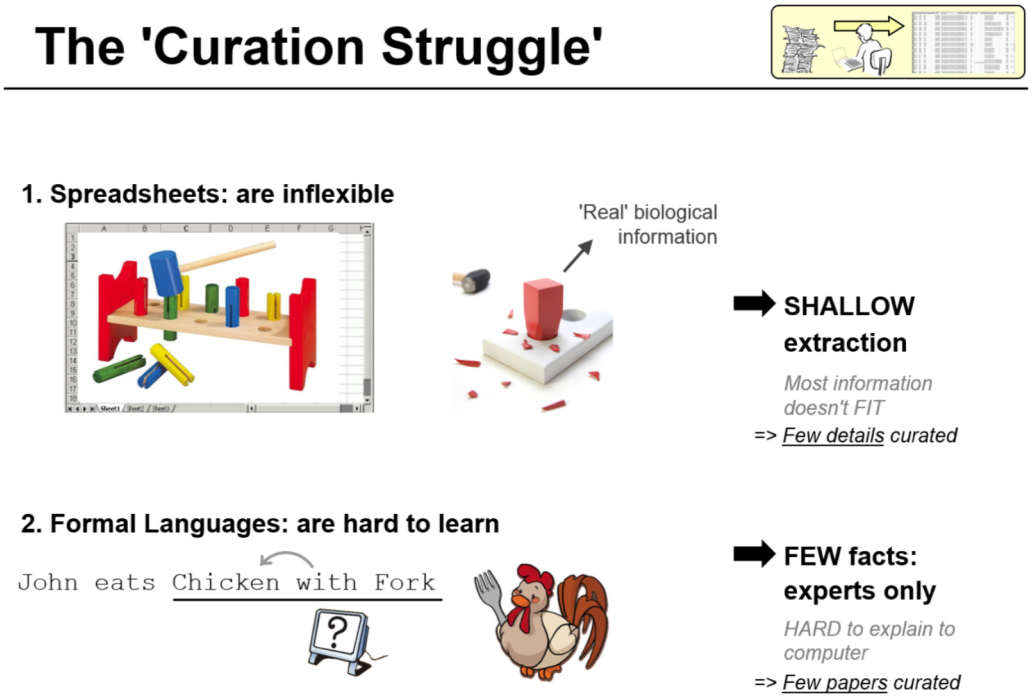
gets structured. Most current curation projects, however, are in fact small-scale, as they are constrained by available *workforce* (limiting the curation volume) and by the supported *scope* (diversity) of content that spreadsheets can easily take in.

2. The key problem

There exists no easy and scalable tool to capture all of Life Science's extremely diverse and flexible (context-rich) information. Instead, a variety of curation approaches are used, often limited to specific types of facts, and with little possibility to capture details about context. Just look at two examples:

To capture details about a biological function with **spreadsheets**, one would not just want to describe e.g. "A activates B", but rather "A, *bound to D*, *transiently* actives a *modified form* of B, and only under certain *conditions*, as shown in *cell type X* of *species* Y, using *experiment Z*". For every type of essential context, you need an extra column. Such an approach soon becomes impractical as there is such heterogeneity of information in research results that a spreadsheet cannot capture it without adding an infinite number of columns. A curator would have to drop essential context, or abuse the form's semantics, or define a 'column X' that becomes a dump of free-text details that remains computationally opaque.

A more flexible format to capture diverse facts is a **controlled language**. However, consider "John eats chicken with fork". The language will have built-in rules that determine how sequences of words are interpreted. As "with" follows "chicken", the computer may capture this fact as "a chicken with a fork", instead of "eating with fork". Extrapolating this to biological sentences of many more words, it is easy to see the learning curve is extremely steep and only few people will take advantage of such languages' power.



3. Our new curation paradigm

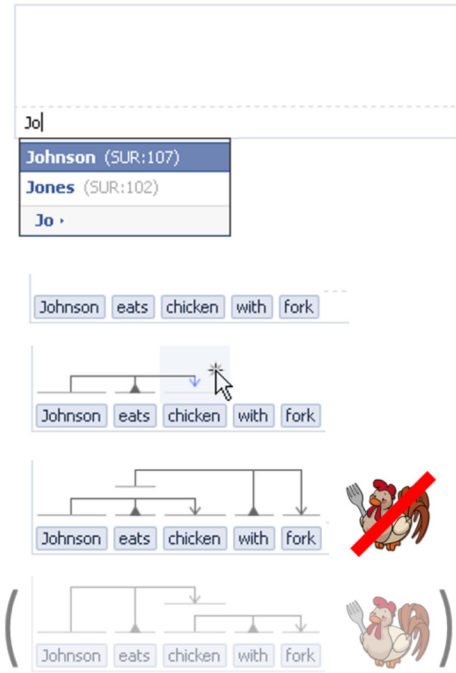
How can we enter any collection of words as an 'information statement' into software, and then make it clear how the words are connected, using only a minimal set of rules as building blocks, used over again?

For this, we developed *VSM: 'Visual Syntax Markup'*. Annotating with VSM is both 1) extremely flexible to capture diverse and rich information, solving the 'spreadsheet' problem, and 2) intuitive to understand and easy to use, solving the 'controlled language' problem.

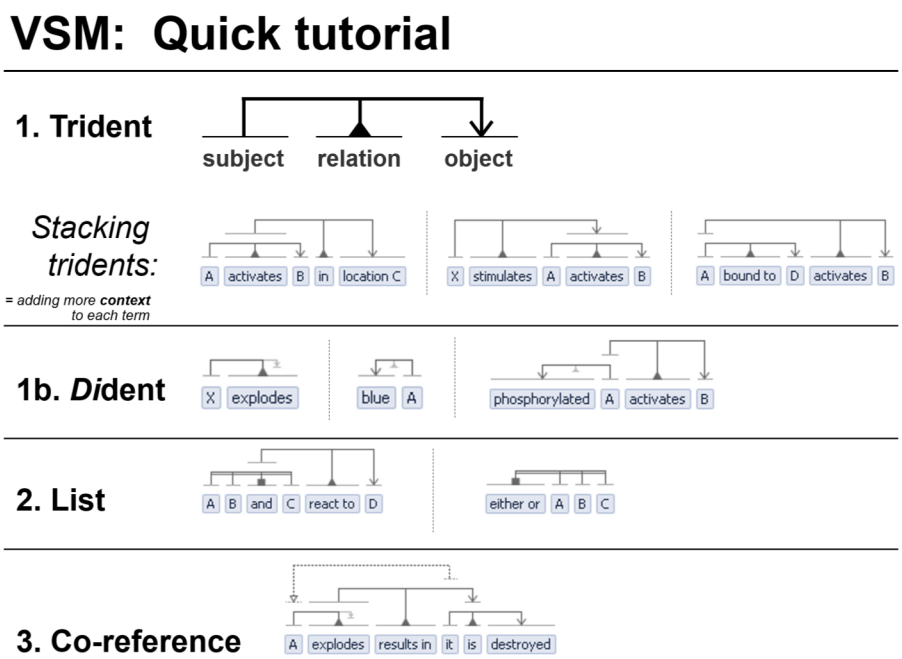
A simple example is shown in the figure below. VSM uses identifier-linked *term*s. Every word or word-group, like "Open Science", must be linked to a semantic ID to make its meaning unambiguous. SciCura's prototype helps people achieve this via an autocomplete list with terms from ontologies and dictionaries, from a back-end database. The example uses a fictional dictionary. A curator chooses a name from this dictionary and adds words from lists of verbs, prepositions and nouns to construct a sentence with 5 'terms', each disambiguated. Next, the curator adds structure, simply by identifying *triples* in the sentence and marking them with a '*trident*' connector. The user-interface enables a curator to do this with three simple clicks: one above each of a set of terms in a triple, identifying the subject, the relation, and the object. The first trident specifies that an individual by the name of Johnson eats an animal of the species 'chicken'. The next trident clarifies that this "eating" is performed with a "fork".

The **key thing to understand about VSM** is the resulting semantics: every time a new trident is added, all connected terms receive extra context. For example, "eat" becomes "the eating of the chicken by Johnson". Likewise "chicken" becomes "the chicken eaten by Johnson", etc. With each concept becoming specific, one can refer to it, and attach further context like: "the eating of .. by ..", that happens with fork. Like this, one can **keep adding further, nested context for every term**.

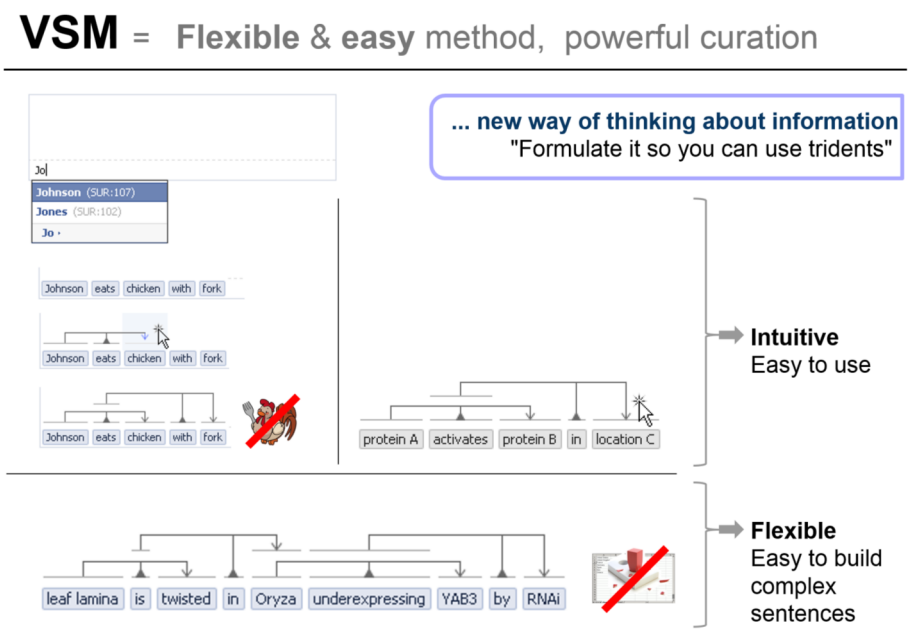
The example shows also that the tridents can be used to clarify the alternative interpretation: the imaginary case where a chicken holding a fork is being eaten. A demo video that illustrates this example step-by step is attached as a media file to our submission.



We designed VSM so it can express just about everything a natural language can, by using three (and a half) connector types: the trident / dident; the list connector; and the co-reference connector that can also refer to terms in other sentences. See figure:

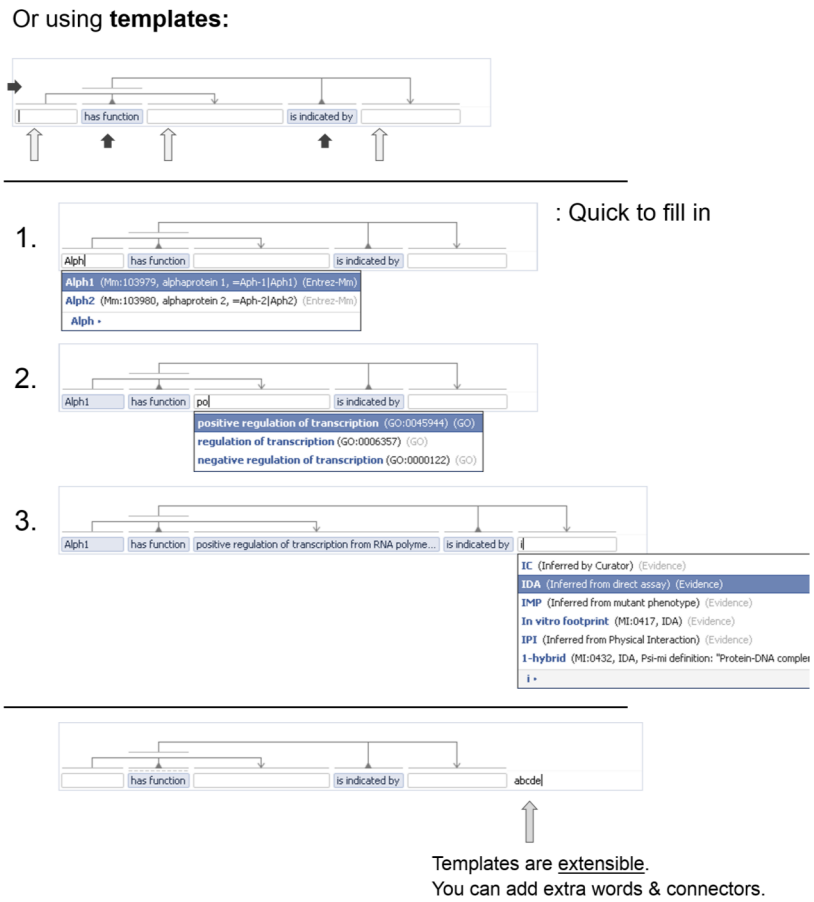


At the bottom of the next figure, an actual fact was built by repeated use of trident connectors adding more detail: "leaf lamina" (the blade of a leaf, an ID-linked term from Plant Ontology) was observed to be "twisted" (PATO term), and this being twisted "pertains to" (=ID-linked synonym of equivalent preposition "in") the species "Oryza" (rice), underexpressing some gene, using RNA-interference (it can be crucial to know which experiment was used to assess a fact's reliability). Such examples quickly become too complex for the spreadsheet method, whereas VSM can capture the biological context.



To be clear: these sentences, or 'information statements', are *not* verbatim copied from a paper. They are reformulated by a user according to their interpretation of some experiment, and represent a disambiguated version of informational essence. In addition, tridents are *not* syntax parse trees generated by algorithms, but are manually added. Although in theory the results of text-mining may help to build such a sentence, this can be investigated later in the development of SciCura.

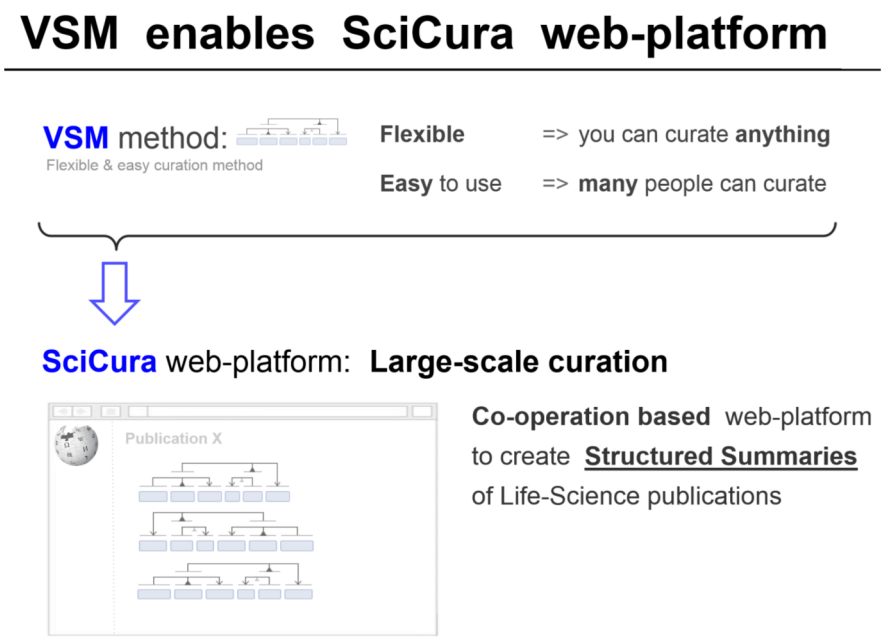
To simplify the entry of information statements using VSM we allow the use of VSM-*templates*, see figure below. These are VSM-statements with some pre-defined terms and connectors that are specific for the information type that should be recorded, with only some additional empty boxes to fill in. These can routinely be used for bulk curation. Yet unlike spreadsheets, empty boxes offer a selection of entity terms or ontology terms linked to their respective IDs. This means that no error-prone identifier-copy/paste steps are needed; only typing a few letters and accepting a suggested term by a mouse-click will do. In addition, templates are extensible: one can still add more words and link them up (avoiding the dreaded 'column X') to specify more context.



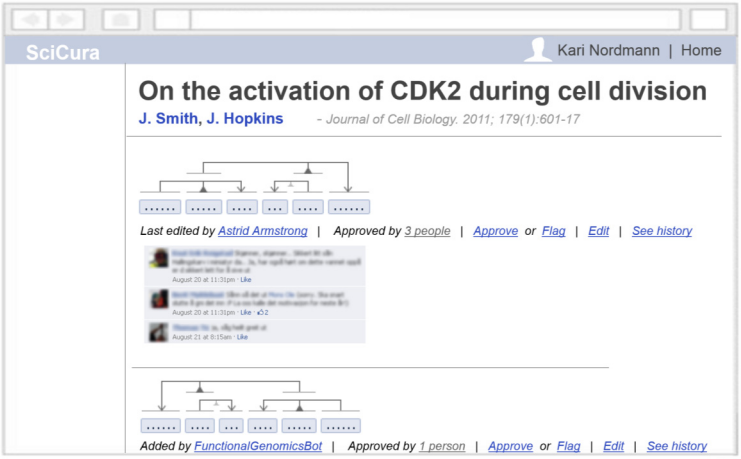
4. SciCura-based community curation

With a flexible and simple curation method that enables a research community to curate whatever they need, we can develop SciCura: a web-based platform for community curation.

Our vision is that after several years it would have one page per paper, holding a VSM-sentence-based *digital abstract* of that paper's detailed findings. This would be similar to a Wikipedia-page, except that the content is fully semantic: the meaning of every term and complex sentence is clearly understood by software.

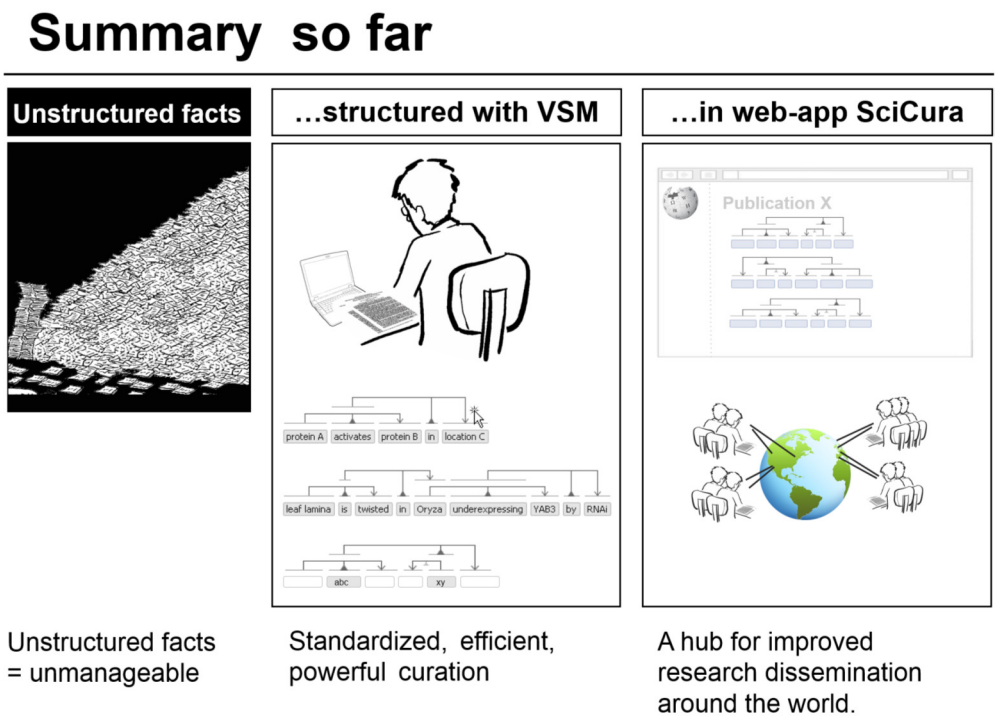


A fully social web platform would also support features like: commenting to discuss interpretations (a real curator need); an edit history; flagging of problems; and reviewing/approving of statements.



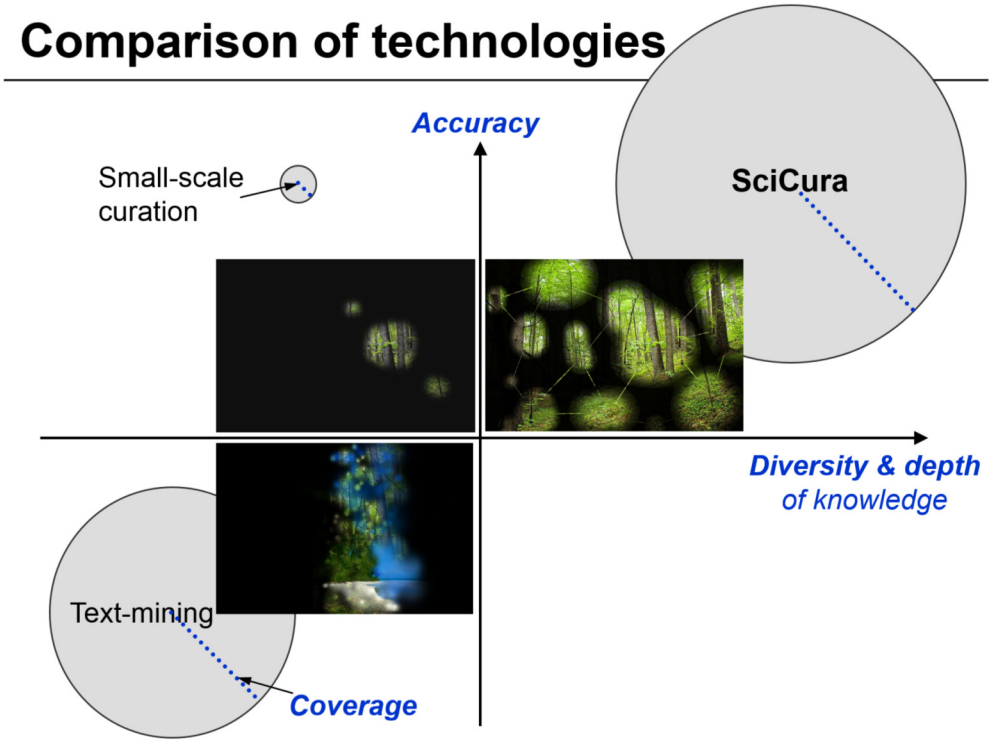
As a comparison: on Wikipedia, 90% just reads, 9% correct typos, and 1% writes content. Similarly, SciCura can grow when most people just fill templates, and some people use VSM's full power and design templates for others.

In summary: diverse, unstructured facts can be structured with VSM. This enables the web-app SciCura to open up previously opaque science results to the world, in a standardized and scalable way.



5. Positioning

A comparison to the current state-of-the-art is shown in the cartoon below. While small-scale curation is accurate and text-mining can boast coverage, both are still limited in scope. Meanwhile, SciCura's has potential for heterogeneous, crowd-sourced manual curation. Small-scale curation may be like walking in a dark forest with a narrow focused flashlight; text-mining is like using a magic lantern that shows many things that aren't really there (errors); SciCura could foster a global community curation effort of high quality and virtually limitless diversity and depth, increasingly showing the full details of the forest.



6. Impact

The impact of SciCura will increase with growing numbers of users. What incentives could there be to get SciCura adopted by a wide community of curators? In our earlier curation experiments, we directly integrated scientists' curations into a browsable overview (see our viewer OLSVis at ols.wordvis.com). The fact that curators could immediately see the results of their work in the context of what else had been entered by their colleagues proved to be an extremely useful 'carrot' for participants.

More generally, SciCura will give scientists access to the sum of curated knowledge through an API, and enable user-written output scripts, plugins, or even a shareable App repository. These additions may elevate SciCura to a community development platform much like the popular Cytoscape platform (cytoscape.org).

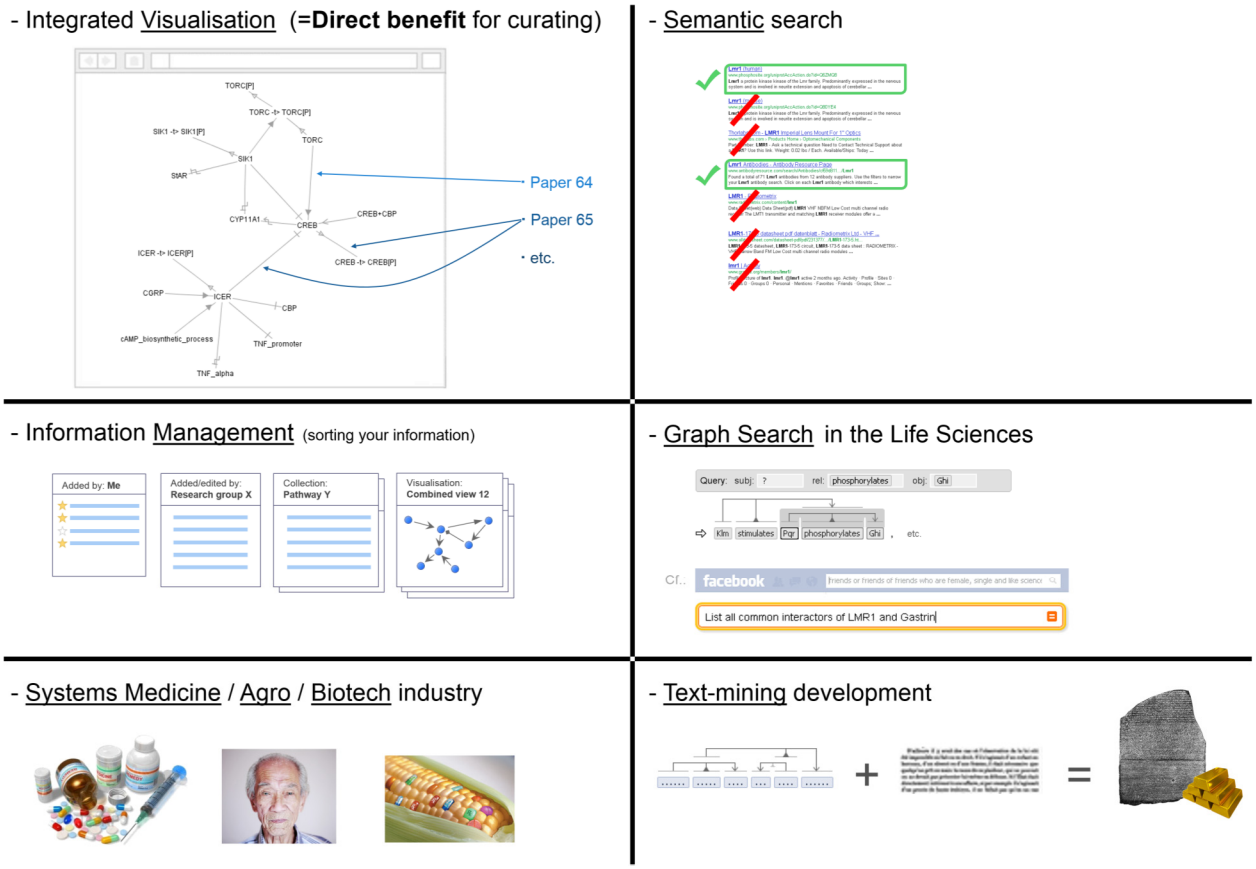
Exports of information to established databases (IntAct, the GO database, other) of the Knowledge Commons will further increase the impact. Several export possibilities are already in place (see section 7), and future export in the semantic web formats RDF and OWL will further the spread of information available through SciCura, including the use of graph searches. OWL exports will also facilitate the integration with the OWL editor Noctua (noctua.berkeleybop.org), and pave the way to apply DL reasoning to the collection of curated facts.

SciCura will also impact the development of text mining algorithms. A collection of structured facts linked to original papers or even sentences is an invaluable text corpus for training. Also, by cross-referencing each VSM statement to the original paper through PubMed IDs, platform users will be able to consult the original papers, which serves as an advertisement for journals and publishing houses.

Research also becomes more cost-efficient. Less duplicated effort of re-interpretation of complex results will happen, and relevant information becomes quicker to find.

Other benefits are shown in the figure below.

In summary, SciCura can be a **transformative project to open up knowledge**, making it computationally available and effectively accessible for Systems Biology, Systems Medicine and Precision Medicine. It can streamline access to open knowledge, deeply understood by intelligent machines. SciCura's growth depends on funding and publicity, making people contribute with code, ideas, and curated knowledge.



7. A First Use Case and Prototype

The current prototype platform was developed in collaboration with volunteer curators at NTNU. Their focus is the curation of all DNA binding transcription factors (DbTFs) of human, mouse and rat; and a curation protocol was developed in collaboration with members of the Gene Ontology Consortium and curators of the Mouse Genome Database at the Jackson laboratories.

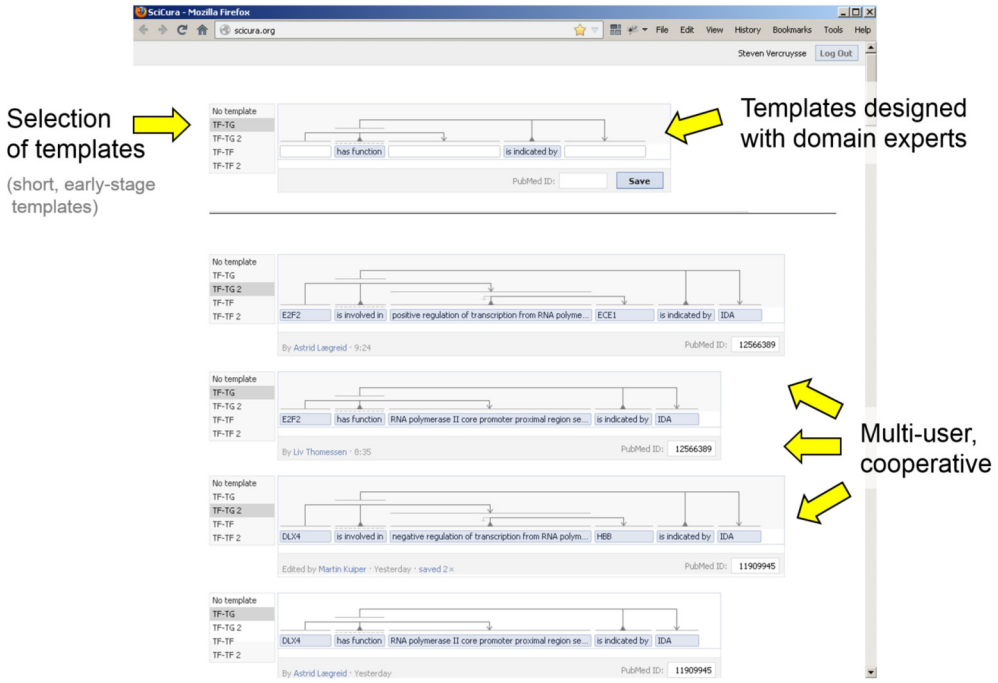
We use an agile, iterative strategy while building the prototype's functionality, in continuous discussion with curators. We prioritize the features that are needed first, while we maintain a balance to keep the web app *generic* enough for next use cases. The SciCura prototype currently features one *curation room* where six people can simultaneously curate DbTFs, their regulation effect, target gene, and experimental details.

We loaded the back-end database with various dictionaries and ontologies needed for the task. As these can be incomplete (e.g. Cell Type), new terms can be added ad-hoc and then reused, and are flagged.

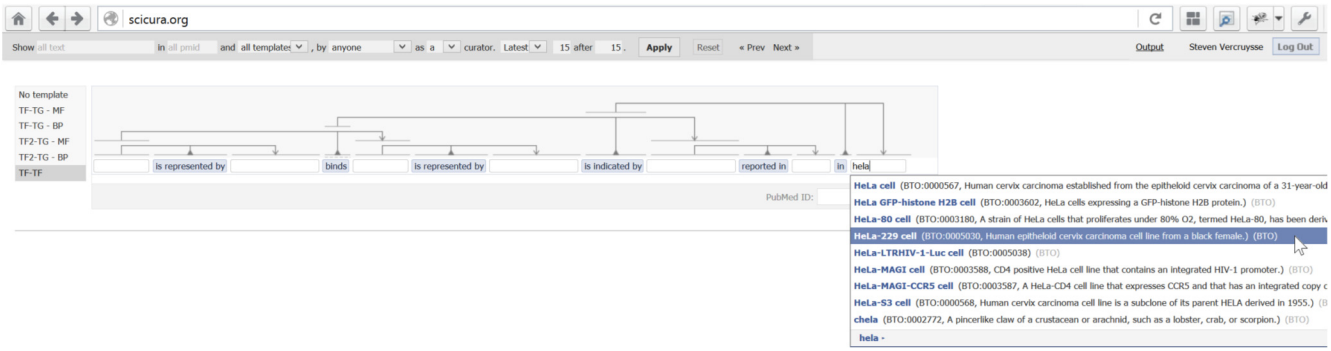
Each curation statement is linked to the original paper, here simply via PubMed ID. We added a place to put the 'original sentence' from which information was extracted, even though facts were often inferred from several parts of text and figures. Curators started using this place to communicate: comments on

interpretation, flagging if re-checked. This shows a user need to develop social and other features around curated sentences.

1) An early version of the curation room:



2) One of five current templates, larger to hold additional details; and an autocomplete list:

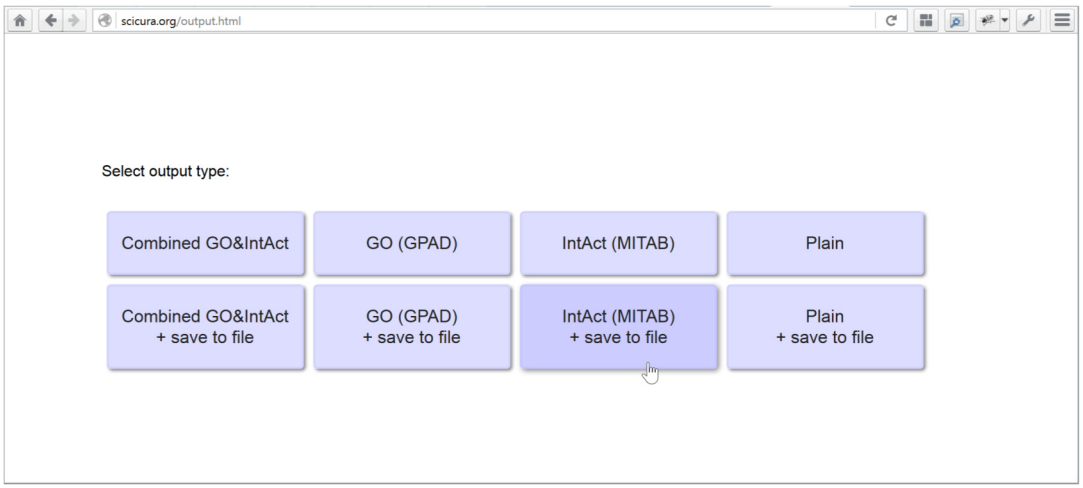


3) A selection from 5000+ curated sentences; under an input box with 5 selectable templates:

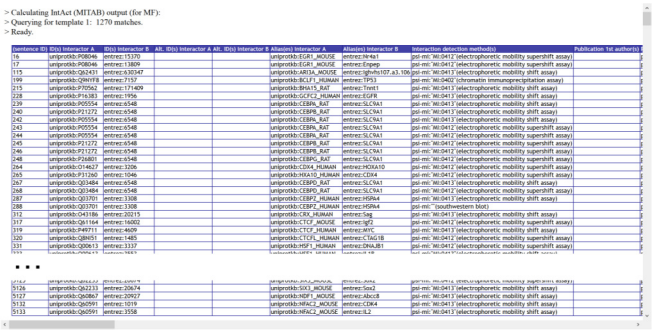
Our users testify they don't find the trident structures difficult at all. In fact, they are excited about all the complexity they can now handle in such an elegant manner, allowing them to focus on the biology without worries about error-prone entry sheets.

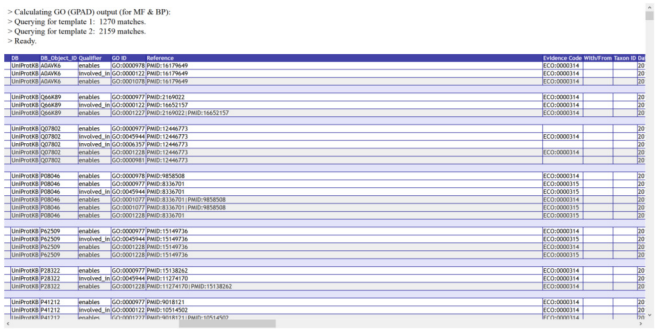
The current prototype can generate output of relevant parts of information to established databases like GO (geneontology.org) and IntAct (www.ebi.ac.uk/intact). We built a graph-query API that can fish out any matching pattern or sub-pattern of terms and connectors in all VSM-sentences. Then we wrote output-scripts to fish out, combine and reformat relevant sets of information.

1) Selection of output scripts:



2) Output results; (grey rows are GO annotations automatically inferred from other ones) :





8. Cooperative work in EU - US partnership

The next development phase is adding the following elements:

A) Curation pilots (NTNU Lægreid group; JAX Blake group):

We enable two use cases:

A1) **Broad curation**: facts from the specific domain of gene expression regulation by transcription factors. Curation templates for this will be improved to meet IntAct standards. Curation will be aided by text mining results (collaboration with Valencia group, CNIO, Madrid), obtained by a broad survey of transcription factor - target gene interactions in PubMed abstracts.

A2) **Deep curation**: a pilot in mouse information curation (Mouse Genome Informatics, JAX). Here the challenge is to develop a flexible template building utility, allowing curators to build their own templates for genetic, genomic and phenotype curation.

B) Integration of ontologies, biological dictionaries, term suggestion service TermGenie (Mungall, BerkeleyBOP):

We will integrate the necessary dictionaries from NCBI Gene, UniProt and NCBI BioPortal into the system. As there may often be a need for new terms/concepts, we will allow users to define candidate terms with a minimal definition and basic categorization, after which candidate terms will be processed/authorised by the TermGenie application (go.termgenie.org).

C) SciCura exports (in collaboration with Sandra Orchard, IntAct, EMBL-EBI; Mungall, Noctua, BerkeleyBOP):

We will further develop the export of facts to the IntAct database. We will in addition enable export to Noctua (OWL format), allowing further refinement of knowledge statements and DL reasoning to classify the information and produce new assertions.

Our work is currently funded by three grants, guaranteeing that by the end of 2016 we will have a platform that can be opened to the general public and open-sourced on GitHub. During this year, our cooperations will help us prioritize between many features we could build, such as:
- multiple curation rooms;
- prefill with statements from text-mining (or from an input-script);
- the infrastructure to support user-defined plugins that query the SciCura content (e.g. to enable custom visualisations). This will set the stage for further community development of the system;
- public/private curations rooms for curation-before-publication, similar to GitHub;
- advanced search based on reasoning.

For continued outreach, we will educate people in use of the platform; present it at conferences, (social) media, and to key biomedical scientists; publish written and video tutorials, and offer training sessions at conferences like ISMB and ECCB, and the International Society for Biocuration.