CoRS Curator:
A Text Curation and Annotation Tool

Senger C*, Erxleben A, Grüning BA, Günther S

Pharmaceutical Bioinformatics,
Institute of Pharmaceutical Sciences, University of Freiburg, Germany
**e-mail: christian.senger@pharmazie.uni-freiburg.de*

Text-Curation

Text-curation describes the selection of, care for, and presentation of objects in texts transferred into a collection. Usually, those collections are databases which can be used for evidence based research. Thus, the databases can be utilised for information look-up during research work, reasoning based on machine learning, or in decision support systems. For this purpose, information in texts has to be structured and machine interpretable. Despite of the existence of several filter and text-highlighting tools, text-curation to the greater part is challenging manual work.

CoRS: Compound Research System

Scientific evidence to the greatest part is published digitally in text form. Such texts are usually without detailed structures representing information context. Experts have to relate associated pieces of information, scattered over a great number of articles. Analyses of small molecules is crucial to find e.g. potential drugs or compound-related health risks. The DFG supported project Compound Research System (CoRS) will establish an automated molecule research system and associated modules such as CoRS Curator to extract and connect relevant information distributed in literature and will enable comprehensive data representation.

Introduction

Searching for information in the fast growing body of available biomedical literature (Fig.1) needs efficient algorithms and software. Several tools are publicly available that are directed at this task [2,3].

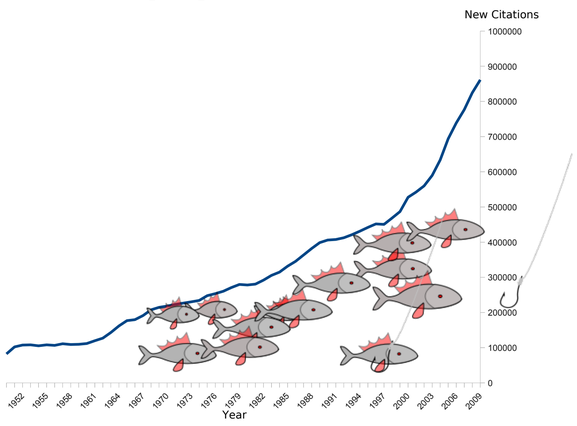

Fig. 1: Growth of MedLine over time [1].

But none of the tools provide a sufficient accuracy to extract reliable data without further review. Thus, these tools have to be looked at rather as filtering step for manual curation.
CoRS Curator uses those tools besides common PubMed and MeSH term searches and provides an interface to enhance subsequent manual curation.

Prerequisites

A software tool supporting manual text-curation, utilised in a small to middle size team should meet several requirements. It should:

- Provide an easy-to-use interface for all kind of scientists, rather than computer scientists only.
- Provide “visual” control, giving direct feedback and a ubiquitous overview of entered data.
- Enable the user to apply a hierarchy of evidence (e.g. meta-analysis, randomized controlled study, letter) for the curated literature.
- Allow the user to define scales and spectra to translate attributes hidden in the context (e.g. `strong`, `medium`, `weak` inhibition).
- Allow the user to freely define entities per project for which text pieces should be collected (e.g. IC_{50}, source organism, assay). Those entities have to be shared with all project participants and have to be unchangeable during one project.
- Enforce the definition of guidelines per project (e.g. step-by-step guide, hints for each entity).

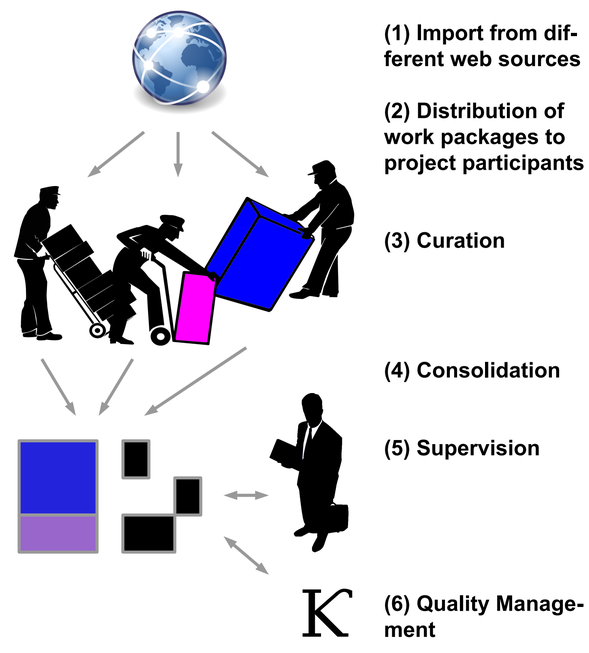

Fig. 2: Important requirements for an optimal curation process.

- Provide a version control to allow comparisons of accomplished work packages and leaping back in the case of errors.
- Support collaboration for teams with differing capabilities and distribution as well as consolidation of (possibly overlapping) work packages of different sizes, enable e.g. the project leader to supervise curated data, and allow for statistical evaluation of the curation quality with e.g. kappa-Tests.
- Support the translation of collected data to vocabularies/ontologies (e.g. `malt sugar metabolic process` → `maltose metabolic process [GO:0000023]`).
- Provide import interfaces from different sources (literature data repositories, automated text-mining, information retrieval tools) and provide export interfaces to different targets (Excel, Relational Database Management Systems, XML, etc.).

Methods

A preliminary version of CoRS Curator was implemented considering the prerequisites mentioned above, as far as feasible.
Functionality and handling has been tested by 1 biologist, 3 students of pharmaceutical sciences (without any curation experiences), and 1 bioinformatician. For evaluation, more than 1500 distinct abstracts were curated to characterise over 500 chemical compounds of the class of Sesquiterpene lactones in 13 categories (entities).
The abstracts were received via interface from the CIL [3] database, such that compound indicating words were highlighted to facilitate identification for the user. The dataset was divided to distribute different work packages and merged in the end of the evaluation.
A comparison of an identical subset of the working packages of the 3 students was used to measure the inter-(cu-)rater agreement via a kappa-like test. Because several chemical compounds were similar, the same abstract had to be curated for more than one compound, leading to 6650 abstracts to be curated. For a more comfortable handling an „inherit“ function was applied, allowing to copy entity annotations from one compound to another if abstracts were identical.
Enhancements compared to a trivial copy-and-paste curation approach using text or Excel files were observed but not evaluated. Recommendations and hints of all curators were collected and used in the iterative further development after careful consideration if reasonable and feasible.
The CoRS Curator software is open source, created with Java 1.6.

A preliminary version is available via email request, sent to: christian.senger[at]pharmazie.uni-freiburg.de.

Results

The curators needed ~90 man-hours to curate 6650 abstracts. The average interrater agreement on the test set was 72%. About 78% of the abstracts contained repeated information and were thus not further analysed. For 11 entities, 2874 pieces of text were annotated, with the highest frequency for “source organism“ with 839 annotations, followed by “therapeutic action” (774) and “assay“ (671). For 2 entities (“literature“, “vendor“) no information at all was found in the abstracts.

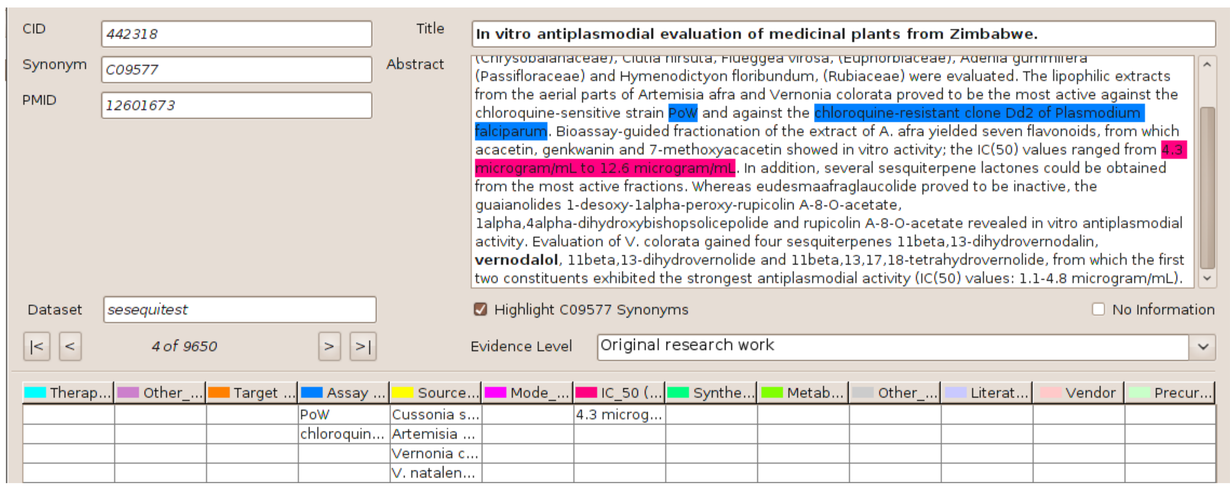

Fig. 3: Example of the CoRS Curator main view.

In this case study, CoRS Curator proved to be a useful tool for manual text-curation work collecting information for chemical compounds from publicly available text sources.

References
1. http://dan.corlan.net/medline-trend.htm
2. Rebholz-Schuhmann D et al. (2008). Text processing through web services: calling whatizit. Bioinformatics 24:296-298
3. Grüning BA et al. (2011). CIL. Compounds In Literature. Bioinformatics 27:1341-2

Acknowledgements:
Thorsten Brink, Julian Haas, and Sebastian Schiffer for their perseverance in the curation work and the hints to enhance the user interface.

Pharmaceutical Bioinformatics
University of Freiburg

The working group of Pharmaceutical Bioinformatics at the Institute for Pharmaceutical Sciences develops algorithms and software for pharmaceutical research. Our fields of research include the modeling of molecular interactions, prediction of biological effects of molecules, and identification of potential new drug agents. The working group is part of the University of Freiburg’s Research Group Program of the Excellence Initiative of the federal and state governments.

Bundesministerium für Bildung und Forschung

UNI FREIBURG