

# User-Driven Development of Text Mining Resources for Cancer Risk Assessment

Lin Sun<sup>1</sup>, Anna Korhonen<sup>1</sup>, Ilona Silins<sup>2</sup>, Ulla Stenius<sup>2</sup>

<sup>1</sup>Computer Laboratory, University of Cambridge, UK; <sup>2</sup>Institute of Environmental Medicine, Karolinska Institutet, Sweden.

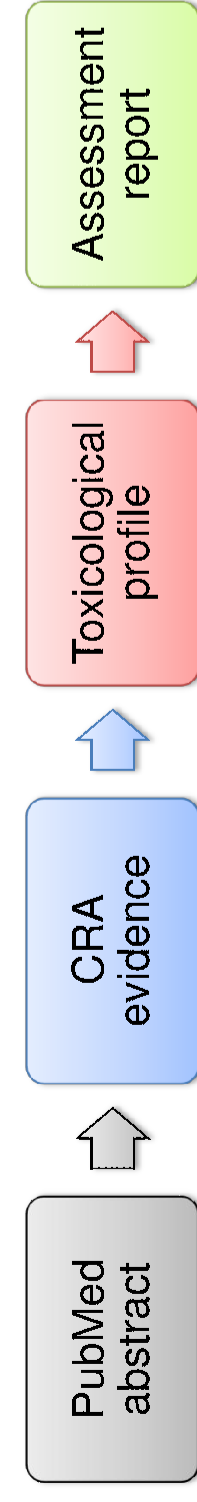
## Introduction

In recent years, the link between environmental chemicals and cancer has become increasingly evident. Cancer Risk Assessment (CRA) is an important task which involves examining published biomedical evidence to determine the relationship between exposure to a chemical and the likelihood of developing cancer from that exposure. Performed manually by experts, CRA can be extremely time-consuming.

We investigated the user needs of CRA and created basic Text Mining (TM) resources for the task. We present a taxonomy which specifies the scientific evidence needed for CRA at the level of detail required for TM. The taxonomy is based on expert annotation of a corpus of 1297 MEDLINE abstracts. We report promising results with inter-annotator agreement tests, automatic classification of corpus data into taxonomy classes, and a user test in a near real-world CRA scenario which shows that the taxonomy is highly accurate and useful for practical CRA.

## User Needs of Cancer Risk Assessment

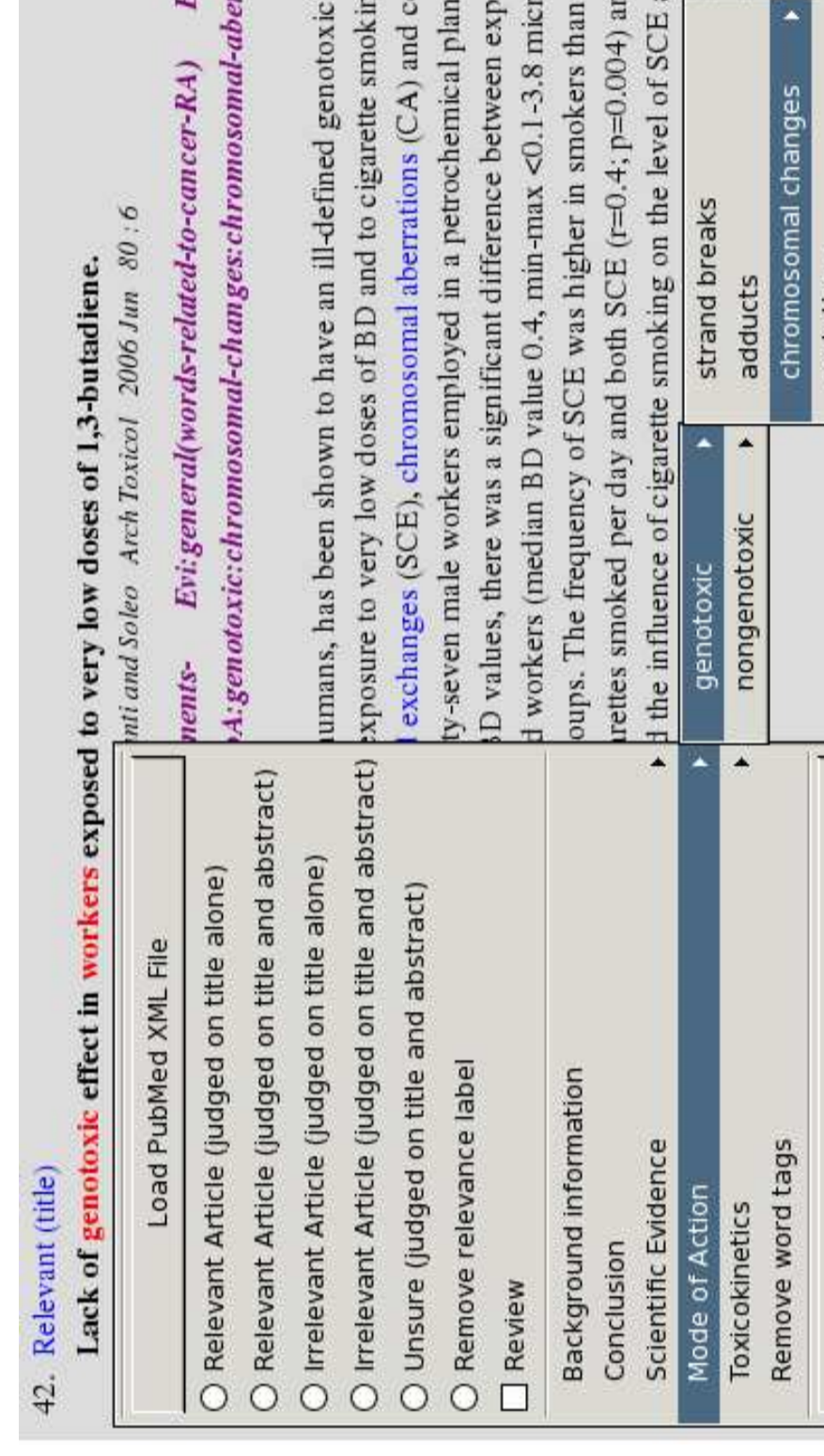
We interviewed 14 experienced risk assessors. They described the different steps of CRA (Figure 1) and reported that locating and classifying the scientific evidence in literature is the most time consuming step. A tool assisting this step would be helpful, but such a tool requires an extensive specification of scientific evidence needed for CRA. No such specification was available. We created it via expert annotation of relevant PubMed abstracts.



**Figure 1:** Workflow of CRA. The process takes 2 years on average for a single chemical. Steps 1 and 2 are the most time consuming.

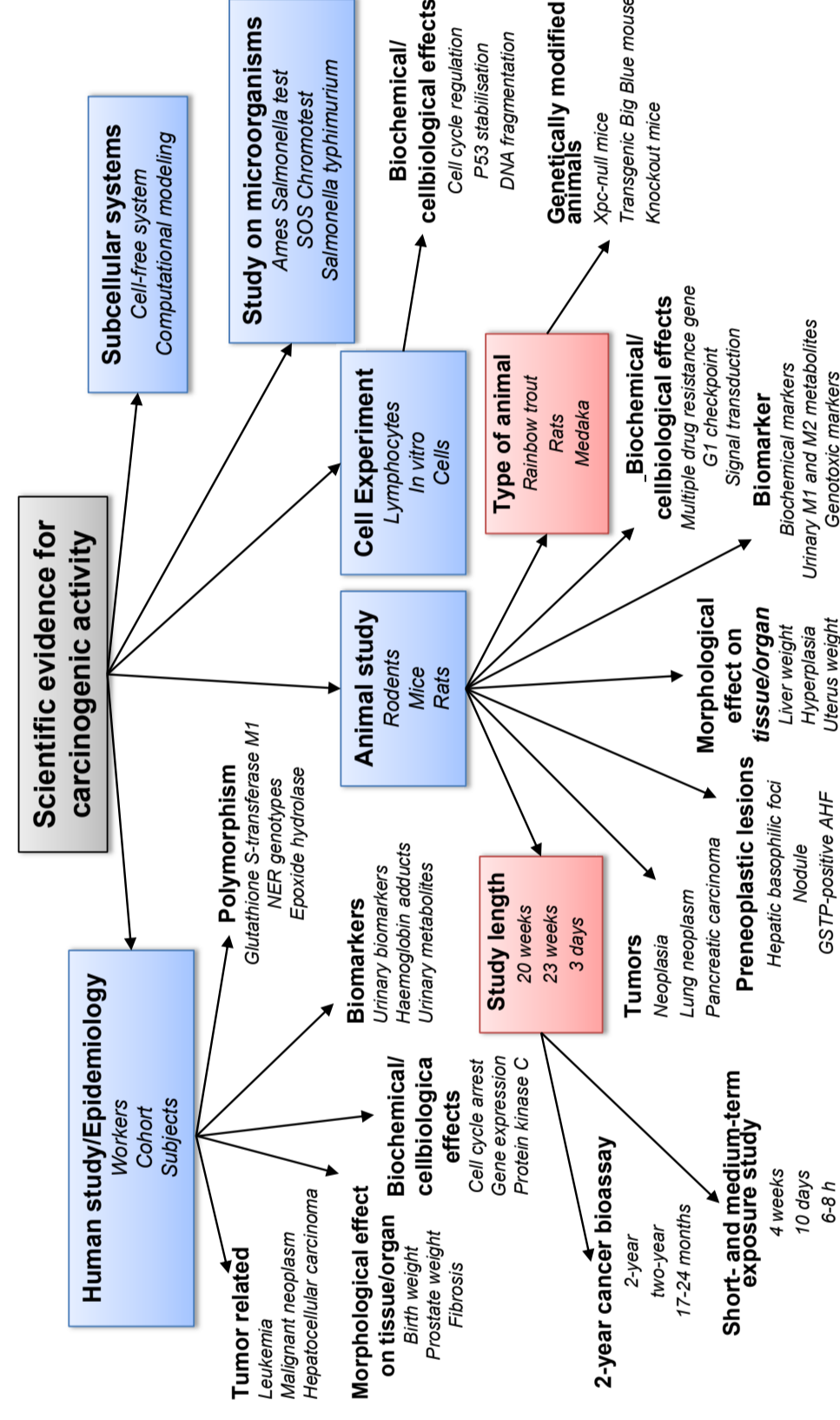
## Cancer Risk Assessment Taxonomy

- We created a CRA corpus. It includes 1297 PubMed abstracts for 8 chemicals which are i) well-researched using scientific tests and (ii) represent the two most frequent Mode of Actions (MOAs) of cancer. The abstracts were retrieved from 15 CRA journals.
- Experts annotated each abstract for (i) relevance and (ii) keywords indicating the types of evidence it provides for CRA. The inter-annotator agreement was 0.68 (Kappa statistics).



**Figure 2:** Screenshot of the annotation tool

- Experts organized the 1742 unique keywords in a taxonomy which specifies the scientific evidence and classifies it into 48 classes. The taxonomy has three top level classes – Carcinogenic activity (CA), Mode of Action (MOA) and Toxicokinetics (TOX). The inter-annotator agreement of assigning abstracts to taxonomy classes is 76%.



**Figure 3:** Taxonomy for Carcinogenic Activity

## Automatic Classification

Abstract classification experiments were conducted to examine whether the taxonomy is machine learnable.

- Experiment** We experimented with two document representation techniques – bag of words (BOW) and bag of substrings (BOS) – and three classification methods: Naive Multinomial Bayesian (NMB), Complement Naive Bayesian (CNB) and Linear Support Vector Machines (SVM).

**Results** SVM has the best overall performance: 0.73 F-measure.

Class	NMB	CNB	SVM
CA	0.91	0.93	0.93
MOA	0.84	0.83	0.86
TOX	0.74	0.75	0.78

**Table 1** Performance of classifiers with for the top level classes

## User Test

A user test was carried out to examine the practical usefulness of automatic classification. The best classifier was applied to the PubMed abstracts of 5 unseen chemicals. The results were presented to the experts. According to their judgement the overall accuracy was high (Table 2). The experts felt that if such a tool was available in real-world CRA, it could significantly increase their productivity and lead to more consistent and thorough CRA.

Name	MOA	$\Sigma$	P	Class	P
Aflatoxin B1	geno	189	0.95	CA	0.94
Benzene	geno	461	0.99	MOA	0.95
PCB	non	761	0.89	TOX	0.99
Tamoxifen	non	382	0.96		
TCDD	non	641	0.96		

**Table 2** Results (Precision, P) of the user test

## Conclusion

The results of our inter-annotator agreement tests, automatic classification experiments and the user test demonstrate that the taxonomy created by risk assessors is accurate, well-defined, and can be useful for CRA.

## Future Work

The current taxonomy is not comprehensive: more data is required especially for low frequency classes, and the taxonomy needs to be extended to cover more specific MOA types (e.g. further subtypes of non-genotoxic chemicals).

- Refine and extend the taxonomy further (e.g. new MOA types) using
  - manual annotation
  - knowledge resources (e.g. MeSH)
  - automatic taxonomy learning
- Use more sophisticated linguistic features in automatic classification
- Develop a TM tool to support the entire CRA workflow

**Acknowledgements:** The Royal Society (UK), the Swedish Council for Working Life and Social Research (Sweden).