NHECD

knowledge on the health, safety and environmental impact of nanoparticles

Creation of a critical and commented database on the health, safety and environmental impact of nanoparticles – Challenges and objectives

Oded Maimon[1], Abel Browarnik[1], Einat Saruk[1], Rafi Korenstein[1], Francois Rossi[2], Hildo Krop[3], Pieter van Broekhuizen[3], Hanno Wittig[4], Anna Maria Lemor[4]

NHECD Goal

The goal of *NHECD* is to build a free access, robust and sustainable system including a knowledge repository on the impact of nanoparticles on health, safety and the environment.

Introduction

The majority of electronic knowledge repositories (such as databases and content management systems) currently existing worldwide, in the field of health, safety and environmental impact of nanoparticles, focus mainly on what can be named metadata. Moreover, those repositories are manually operated, thus, allowing only a limited amount of data being processed, and utilizing a rather unsystematic taxonomy and ontology, which guide the documents classification and information extraction processes.

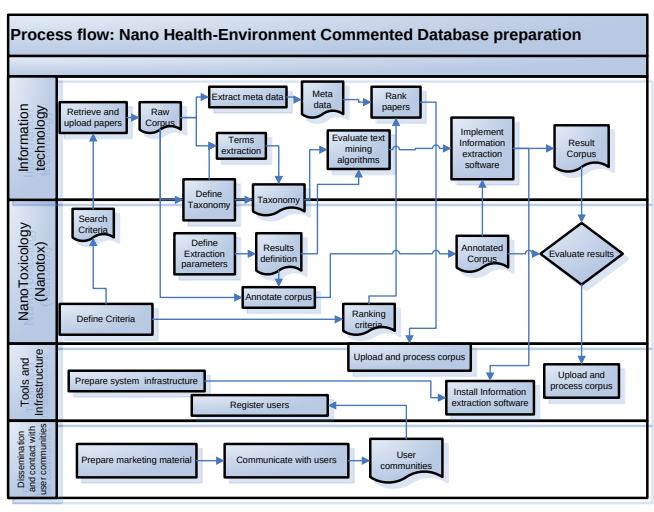
The European Commission 7th frame funded project, NHECD, is intended at converting the unstructured body of knowledge produced by the different groups of users (such as researchers and regulators) into a database of scientific papers and reviews (e.g., whitepapers), augmented by layers of information extracted from the above papers.

NHECD will include a robust content management system (Documentum) as its backbone, to hold unstructured data (e.g., scientific papers and other relevant publications). It will also include a mechanism for automatically updating its documents repository, thus enabling the creation of a large, updated and developing collection of published data on environmental and health effects following exposure to nanoparticles. The repository will be harmonized to be compatible at the metadata level with existing databases.

NHECD will be based on text mining methods and algorithms that will make possible the transition from metadata (such as author names, journals, keywords) to the data itself. The transition will be implemented by using innovative and automated text mining techniques. These methods and algorithms will be implemented to specifically extract pertinent information from a large amount of documents. NHECD will include the development of a systematic domain model of concepts and terms (i.e., a wide set of domain taxonomies) that will support the classification of papers. It will also include the development of the information extraction process. Particular domain-specific zoning and text mining algorithms will be applied to reach the defined goals.

The unique features of this database will allow different user groups – academics, industry, public institutions and the general public – to easily access, locate and retrieve information relevant to their needs. The creation of the NHECD knowledge repository will enrich public understanding of the impact of nanoparticles on health and the environments; will support a safe and responsible development and use of nanotechnology; and will represent a useful instrument for the implementation of relevant regulatory measures and legislation.

The *NHECD* model



The process starts with a collection of documents (such as scientific papers) gathered by means of a search using criteria given by Nanotox experts. The documents are accompanied by the corresponding metadata (e.g., authors, publication dates, journals, keywords supplied by the authors, abstract and more). The process requires Nanotox taxonomies. Taxonomies are classification artifacts used at the information extraction stage (taxonomies are also used in NHECD for document navigation). Taxonomy building tasks are "located" at the boundary between the Nanotox experts and the IT experts, due to its interdisciplinary nature.

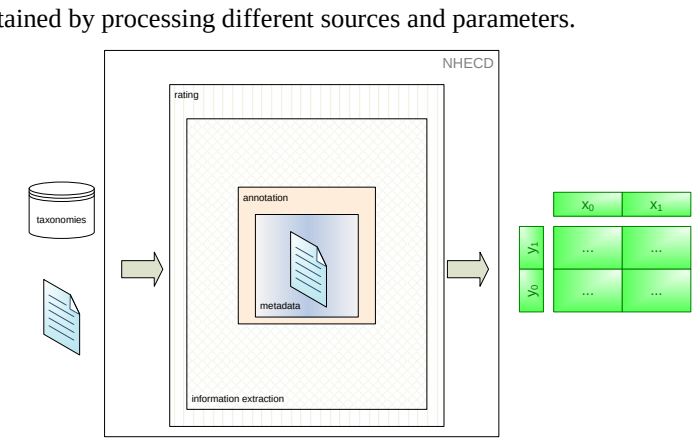
Nanotox experts annotate papers to train the system towards the information extraction stage. This stage is implemented using text mining algorithms. Further to the information extraction process, a set of rating algorithms is applied on the documents to provide an additional layer of information (e.g., the rating).

NHECD implementation

- **Taxonomies**: essential to the NHECD process, which uses taxonomies to classify papers, both for information extraction and later on for repository navigation
- **Crawling**: The process of automatically obtaining scientific papers and data about the paper (such as the name of the author or authors, the publication date, the name of the journal, keywords, abstract, and in general any detail made available with the paper itself) by visiting scientific paper repositories available on the web (whether restricted to subscribers or available to everyone) and searching by keywords on the paper's text
- **Information extraction**: methods and algorithms used to enable users to ask specific questions about attributes and receive answers and a link to the paper along with a pointer within the document where the information exists, and allow, in the future, data mining on the extracted information patterns
- **Scientific paper rating**: one of NHECD contributions, seen as an additional *comment* layer. Obtained by processing different sources and parameters.

The result of the process consists of:

- A corpus of results, updated on an ongoing, asynchronous basis.
- A *commented* collection of scientific papers. By commented we refer to the added layer of metadata, rating and other information extracted from the document.



Conclusions

NHECD will provide two important *products*:

- An extensive and commented repository of scientific papers and other publications in the Nanotox area, searchable using taxonomies and full text search. The scientific papers will be rated according to published NHECD criteria, to help users better estimate their findings. Such a repository will significantly expand currently available repositories due to the fact that it goes beyond the mapping of existing research in Nanotox (as most current initiatives do). NHECD will give access to the research papers results, extracted from the sources using text mining algorithms. Access to scientific papers will be granted to visitors following copyright and restrictions as imposed by publishers.This NHECD result is intended for Nanotox scientists, regulators and for the general public.
- A set of structured results extracted from the scientific papers populating the NHECD repository. Using these results it will be possible to perform data mining on the results. Data mining will result in validated results and further knowledge discovery. This part of NHECD results is targeted at Nanotox scientists and regulators.

Challenges

- Automatic population
- Information extraction
- Repository up-to-date
- Build/update taxonomies
- Papers rating
- Intelligent retrieval

Find more information on www.nhecd-fp7.eu

NHECD has received funding from the European Commission's Seventh Framework Programme [FP7/2007-2013] under *grant agreement* n° 218639. Neither the European Commission nor any person acting on behalf of the Commission is responsible for the use which might be made of the given information.

[1] *Tel AvivUniversity – Tel Aviv, Israel* [2] *Joint Research Centre Ispra – Ispra, Italy* [3] *Environmental and Occupational Health Research Institute – Amsterdam, The Netherlands* [4] *tp21 GmbH – Saarbruecken, Germany*