Final Report Summary - NHECD (Nano health-environment commented database)

NHECD is free access, robust and sustainable web based information system including a knowledge repository on the impact of nanoparticles on health, safety and the environment. It includes a robust content management system (CMS) as its backbone, to hold unstructured data (e.g. scientific papers and other relevant publications). It also includes a mechanism for automatically updating its knowledge repository, thus enabling the creation of a large and developing collection of published data on environmental and health effects following exposure to nanoparticles.

NHECD is based on text mining methods and algorithms that make possible the transition from metadata (such as author names, journals, keywords) to more sophisticated metadata and to additional information extracted from the scientific papers themself. These methods and algorithms were implemented to specifically extract pertinent information from large amount of documents. NHECD created a systematic domain model of concepts and terms (i.e. a wide set of domain taxonomies) to support the categorisation of published papers and the information extraction process within this project.

The unique features of NHECD allow different user groups - academics, industry, public institutions and the public at large - to easily access, locate and retrieve information relevant to their needs. The creation of the NHECD knowledge repository enriches public understanding of the impact of nanoparticles on health and the environments; it supports a safe and responsible development and use of engineered nanoparticles and represents a useful instrument for the implementation of relevant regulatory measures and law making.

Project context and objectives:

Background
Nanotechnology has led to advances in many diverse areas, including medicine and healthcare, information technology (IT), energy, household and consumer products, to name but a few. Nanotechnology has diffused in the marketplace and has certainly revolutionised scientific areas such as organic chemistry, molecular biology and semiconductor physics.

This thriving marketplace, however, has also led to an increase in the awareness of possible risks to environmental safety, human health and safety when using these new materials. For example, certain nanoparticles, upon entrance into the body, have not only been shown to cause serious lung problems but are also transported to other organs like the brain or the heart without knowing whether they pose additional risks there. Nanoparticles or nanomaterials (a nanometre being one millionth of a millimetre) pose a potential risk due to their minute size, meaning they function and behave differently from their larger counterparts (e.g. bulk matter). Uncovering the potential harmfulness nanoparticles (nanotoxicology) remains a very active and ongoing area of research, and this has called for deeper investigation into their impact.

The rise of potential health hazards has led to the creation of a new discipline. Nanotoxicity - the study of toxicity in nanomaterials - is considered to be an essential part of nanotechnology development. This can be a lengthy process as all nanomaterials are different and need to be investigated individually and separately. Therefore, some nanomaterials are likely to demonstrate more biological effect than others.

Interest regarding the impact that nanoparticles have on human health and the environment has started to grow, and the potential of nanotechnology to progress will be largely marked by the strategies we use to ensure safety in these key areas. The European industry's interest surrounding nanosafety has enlarged due to the number of nanoproducts being sold commercially. Also, environmental and ethical groups are concerned with the regulatory measures being taken during product development. Despite the large amount of money that is applied to nanosafety research and development (R&D), there are still issues with quantifying and qualifying all of the information on nanotoxicity. A successful evolution of nanotechnology is therefore dependent on our ability to attest for all precautions and preoccupations, as well as keeping policy makers, the general public, industry, scientists and researchers, companies and governments in-the-loop and aware of the hazards.

New and improved approaches towards the safety of nanotechnology have become a high priority on the agendas of many scientists and political groups and to meet this growing demand, a multitude of organisations and institutions are tackling the issues. One challenge facing many of those exploring nanosafety is that the sheer extent and volume of information is incredibly high. In order to collect this data effectively and efficiently, investigators will need to invest in a suitable method for dealing with a variety of data types. Therefore, there was an urgent need to have a free access structured information repository on the impact of nanoparticles on health and environment that will possess the capabilities of employing semi-automatic tools for update and comprehensive analysis of the published data.

Our project

The European Commission (EC)-funded project NHECD converts the unstructured body of knowledge produced by the different groups of users (such as researchers and regulators) into a repository of scientific papers and reviews (i.e. whitepapers) augmented by layers of information extracted from the papers.

The goal of the NHECD project was to build a free access robust and sustainable system than can meet the challenge of semi-automatically keeping rich and up-to-date scientific research repository, enabling a comprehensive analysis of published data on environment and health effects following exposure to nanoparticles. Methodology and automated ranking and commenting processes was set up handling unlimited number of detailed research results so that a large scale knowledge base can be built and maintained on a long term basis. This process was accompanied and supported by automated data and text mining techniques capable to extract the results from unlimited number of published papers.

The NHECD approach was based on the integration of the following features:

- innovative text mining tools designed specifically to extract information from scientific research papers in the nanoparticles domain;
- automated extraction of electronically published scientific research papers keeping quality results;
- leading toxicology domain knowledge provided by the NHECD partners;
- website as a front-end for hosting facilities as well as a web based user interface for the data base application;
- effective public relations operation to expose the repository to various audiences such as scientific community, regulatory bodies and
the general public, dealing with generating white paper and production of summaries for interaction with stakeholders and dissemination to Industry and public.

Project results:

NHECD has been gathering data from existing literature in the area of nanotoxicity and has created a novel system known as the 'Nano health-environment commented database' (NHECD). It's a free access web-based information system that includes a knowledge repository on the impact and effects of nanoparticles on the environment, health and safety (EHS). Ensuring that all the different data sources that are a part of the database (e.g. scientific papers, white papers) are categorised into a solid and coherent structure is no easy task. This is why we used a robust CMS, which acts as the centre point of the entire system. All of the information collected thus relates to the environmental and health effects of nanoparticles, and can be easily accessed, thus, reducing time consumption and filtering inappropriate content.

Coming to the end of the project, the system has already made some landmark achievements. Using a mixture of unique features, robust and concrete database management and expertise within the team has allowed for a solid web-based product, enabling users to access information quickly and directly.

How does it work?

The system was constructed keeping in mind the major goal of facilitating the extraction of relevant and suitable information from a vast range of documents (such as scientific papers). It provides a consolidated resource, bringing together articles from various sources into one repository. It is available to academics, industry and public institutions, in accordance with their particular needs and requirements. Moreover, the resource is available to the general public, in a bid to generate more awareness about the effects of nanoparticles and the hazards that are associated with its development.

NHECD database is reachable at general project information via http://www.nhecd-fp7.eu

Project achievements:

To date, the system includes a number of tailored features that support both users and administrators:

- The interface of NHECD to the three communities targeted for it, namely nanotox scientists, regulators and the public. NHECD website is a comprehensive solution designed to allow users to search for relevant information using a state-of-the-art graphical user interface (GUI) matching diverse types of users (regular users, sophisticated users and more). The GUI allows for taxonomic search, simple / advanced search, full-text search, intelligent search (a unique method that enables researchers to search for the information extracted from the scientific papers) and any combination of the above search methods.
- A backend system based on a robust CMS and its accompanying modules such as classification, full text search and more.
- A crawling system designed to effectively navigate selected sources (automatically or semi-automatically) for the purpose of obtaining data materials related to NHECD.
- An information extraction component that allows users to view list of relations from each scientific paper found in the repository.
- A rich set of computer based taxonomies related to the NHECD target areas to support the categorisation of papers and easing users search queries.
- A body of classified knowledge consisting of scientific papers related to in-vivo / in-vitro, ecotox and occupational nanotoxicology. The corpus currently contains around 10 000 papers on NHECD related subject.
- A robust infrastructure which carries out administration and maintenance activities effectively also after the project has officially ended - and many more.

Data and data sources at the basis of the NHECD repository:

NHECD features peer reviewed papers on the different aspects of engineered nanoparticles toxicology. It also contains Nanotox white papers manually uploaded. NHECD crawls all publicly accessible sources, such as pubmed and biomed. The summaries are made available to the different user communities, while the papers itself are kept confidential, according to intellectual property limitations as set by legislation in place.
What can be found?

We have gathered around 10 000 open source articles thus far; all are in English. Our repository of scientific papers related to Nanotox, augmented by metadata provided by authors and publishers, metadata extracted from the papers using text mining algorithms, and ratings for the articles based on methods adopted by NHECD, to help users better estimate their findings. All the above are indexed using NHECD taxonomies. As a result, it is possible to retrieve scientific papers using sophisticated queries and full text search. Within the database, information regarding specific characteristics of nanoparticle's toxicity can be extracted automatically and converted into structured data. As such, anyone who wishes to access a specific article can get more detailed analysis in an automatic way; if, for instance, one asks a specific question regarding the use of certain nanoparticles under in vivo or in vitro, one can get a list of detailed information, in order to assess key questions. All the data can be easily accessed through the system via one of the search methods: basic search, advanced search, intelligent search or taxonomic navigation.

Who could use the data and for which purpose?

NHECD is intended for three diverse communities of users:

- nanotox researchers can profit from NHECD capabilities towards reviewing literature while preparing research papers or reviews;
- regulators and non-governmental organisations (NGO) can resort to NHECD to find relevant information on the effects of nanoparticles on health and the environment, both generally as well as related to issues such as occupational health;
- the general public can use NHECD data and reviews to mitigate worries related to nanotoxicology.

What are the system limitations?

NHECD limited coverage - NHECD relies on the public or academic availability of information resources such as academic papers, governmental or institutional publications and more resources related to the toxicology and effects of nanoparticles. The open dissemination of such knowledge is jeopardised by the accumulation of isolated islands of knowledge that tend to put barriers to the wide use of resources for economic profit. This is the case when trying to obtain information (even for non-profit initiatives such as NHECD) from sources such as Google, Thomson Reuters, Springer, Elsevier and others.

NHECD made every possible effort to overcome these limitations while making sure that no legal issues are incurred. Currently, NHECD covers peer-reviews article from PUBMED and BioMED.

The main effort in the near future will be to get in touch with major scientific editors (Elsevier, Wiley, Thomson Reuters, ACS) in order to get access to their article database, thus expanding significantly the repository coverage.

The information extracted is restricted to the abstract and result section of the peer reviewed article. Scientific advances together with further development of NHECD may lead to more extensive information extraction. Also, further development is needed in the relation extraction part to enhance the quality of the results.

The NHECD user interface:

The system includes basic, advanced, intelligent and taxonomic level search features, depending on the specifics required during the retrieval process. This provides a comprehensive solution to wading through copious amounts of information, and provides the user with an option to make either a general or specialised search.

- A basic search interface to perform keyword and/or taxonomy based queries.
- An advanced search interface for experienced users to extend the capabilities of the basic search using advanced search features such as logical operators (and, or, not), as well as allowing to search through extended metadata.
- The intelligent search is a unique method especially adapted to the needs of researchers in the nanoscience field. It is specifically aimed at targeting a particular piece of information required by the user. The search method was initially created for researchers’ nanoscience needs and is capable of allowing the user to search for content by cell or / and animal model, experiment or by distinguishing characteristics. By applying an intelligent search one will be able to identify links between the physico-chemical properties of nanoparticles and their specific biological effects. As a result the user gets a set of structured facts extracted from the scientific papers in tabular format.
Further functionality allows for:

- features such as save searches, recall of recent queries, managing a clipboard and organising the data in collections;
- tutorials guide the user to efficiently use the interface.

In addition to the search functionalities provided, a variety of further information resources related to the health, safety and environmental impact of nanoparticles is available:

- a White Paper forum allowing registered;
- users to comment on white papers entered by a moderator;
- further features include an introduction to the field, information on corresponding rules and regulations, news and events, and inform the users on relevant developments and activities related to the potential impact of nanoparticles on health, safety and the environment.

NHECD activities were conducted as three main work packages including:

1. provision of IT;
2. toxicology;
3. tools and environment.

NHECD success was a consequence of the synergetic relationship between the work packages results, themselves a result of the synergy between the partners. When possible, joint activities that were not initially planned were held to ensure that NHECD achieves its goals in an optimal way.

1. Provision of IT

The objective of this part was to provide the IT infrastructure enabling to establish and maintain on an on-going basis an automated retrieval, indexing and extraction of relevant results from scientific publications. It includes a robust CMS (Documentum) as its backbone, to hold unstructured data (e.g. scientific papers and other relevant publications). It also includes a mechanism for automatically updating its document 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 data is categorised using domain taxonomies specially designed by domain experts for efficient navigation of the repository by novice users.

The main tasks under this work package included:

- design of documents database metadata and data structures for effective and efficient retrieval and reporting;
- preparation of training and test corpora of expected extraction results;
- implementation of the taxonomy and categorisation of the corpora;
- evaluation and development of efficient text mining software for automated information extraction from free text, based upon the selected algorithms;
- initial and on-going retrieval, indexing, analysis and update of relevant scientific publications.

Retrieval process:

At first, we implemented a process to obtain the preliminary training and test corpora required for the development of a crawler, an automatic device designed to:

- search the web given specific keywords;
- fetch matching articles with their meta-data;
- upload results into NHECD repository according to its unique architecture.

The crawler was created in-house. Few publishing sites were targeted as web repositories to be harvested: Pubmed and Biomed. These web databases allow free and unlimited access to their knowledge base by automatic ‘robots’ (such as the NHECD crawler). Using
predefined queries (based on those scientific domains pertaining to NHECD that are dealt with by TAU) a corpus of roughly 5000 scientific papers was formed.

The ongoing retrieval and indexing of publications was accompanied with a thorough validation procedure of the entire corpus. We created an evaluation process in order to check the performance of the procedure and the quality of the corpus. To this end, we developed an easy-to-use tool for the domain experts to perform the validation. Each domain expert got a few hundreds articles to review and had to assess whether the papers were relevant to his domain or to the corpus at all.

The results of the analysis led to the improvement of the retrieval conditions. We used the indexing of medical subject headings (MeSH) from the MEDLINE / PubMed article database. (MeSH is a comprehensive controlled vocabulary used to index journal articles and books in the area of life sciences).

At the end of the project, the repository contains approximately 10,000 research papers. The results of the validation procedure led to an excellent result. The quality of the corpus was improved significantly with 80-90% relevancy of papers.

Information extraction (IE):

NHECD is based on text mining methods and algorithms that make possible the transition from metadata (such as author names, journals, keywords, abstract) to the data itself. The transition is implemented by using innovative and automated text mining techniques. These methods and algorithms are implemented to specifically extract pertinent information from a large amount of documents. It included the development of a systematic domain model of concepts and terms (i.e. a wide set of domain taxonomies) that support the classification of papers. It also included the development of the information extraction process. Particular domain-specific zoning and text mining algorithms were applied to reach the defined goals. Information extraction is a type of information retrieval whose goal is to automatically extract structured information from unstructured and/or semi-structured machine-readable documents. In most cases, this activity concerns processing natural language texts. Information extraction has numerous potential applications. For example, information available as unstructured text can be transformed into traditional databases that users can probe through standard queries.

The aim of the NHECD IE component is to extract, from every scientific paper gathered by the NHECD crawler, a comprehensive, full and precise list of relations. NHECD text mining tasks are in fact information extraction tasks, namely, to extract entities and relations (which are, by nature, structured information) from unstructured nanoparticle-toxicity related documents. The information extraction system expected the following entities or relations:

(i) nanoparticle,
(ii) model - cell model or animal,
(iii) attributes - nanoparticle size, Zeta potential, animal age, and
(iv) experiment attributes - mode of exposure, measurement assay.

We provided domain experts with tools to assist their evaluation of the extracted result to enhance the quality of the results. The tool shows the article text and the extracted entities and relations. Using this tool, the domain experts were able to identify the main domain-related problems.

In addition, the IE component was reviewed to reflect the state-of-the-art developments in the literature. It was an on-going process throughout the life of the project. Information extraction yielded a precision of approximately 60%.

2. Toxicology

The overall goal of the toxicology group was to provide knowledge in the domain of the impact of nanoparticles on health and environment for setting a database. This was accomplished by providing appropriate taxonomies, prioritising and ranking of the articles and selection of PDF articles, in addition to the provided meta-data, that will serve for a comprehensive data mining. An additional central goal for the toxicology group was to validate the tools for text mining at each stage of development of the data extraction component and to serve as the quality control of the developed process.

The main tasks of this work package were:
Set taxonomy:

At the first stage we had set-up a prototype for the database taxonomy. A taxonomy is composed of units (taxons). Taxonomies are used to classify items according to a predetermined set of categories, arranged in a hierarchical way.

NHECD used taxonomies to classify the documents in the repository. Each document was linked to one or more taxons in one or several taxonomies. The aim of the classification according to taxonomies was to facilitate navigation and search within the document database.

Taxonomies are also one of the components of the information extraction process that stands at the basis of NHECD text mining algorithms.

The classification process is based on the content intelligence services (CIS) module of Documentum. It requires the development of taxonomies, and the definition of a set of evidences (per taxon). The work included:

(i) identifying the taxonomies,
(ii) developing the taxonomies,
(iii) defining the set of evidence rules for each taxon,
(iv) importing the taxonomies to the CIS module of Documentum,
(v) test, evaluation and validation.

The result achieved with this task consists of 50 taxonomies that represent the domains NHECD deals with (health, environment and occupational effects of nanoparticles). The taxonomies can assist visitors of the NHECD frontend in navigation and search related tasks. It should be emphasised that the taxonomies are an important input for the text mining process that makes NHECD different from traditional search engines. Furthermore, obtaining the taxonomies is an achievement that goes beyond the project itself, and may have influences on the whole domain.

Ranking process:

Ranking of data is a mean of sorting data by a relative score given to an article. This is done by using the metadata of the article. The relative score (rating) is attached in the NHECD website as an attribute to any article in the corpus, allowing the users to consider it in their search for documents. The elaboration, definition and agreement on ranking criteria of data were very important for the further implementation and progress of NHECD. The ranking was done on the scientific papers in the NHECD repository using a set of rating algorithms, developed specifically for this project. The rating algorithms are unique in the way they take into account the citation 'strength' of each citing source. To build the rating, we map the citations from each scientific paper to all other scientific papers in the repository; this mapping leads to a time-dependant function of citations for each paper. Using this function, we prepared a per-paper rating score based on several 'strength' parameters per citing source. The advantage of this method is that papers published on less-known journals cited on better-known journals, will receive a good score. Another advantage is that papers cited consistently over time will be ranked higher, as compared to papers cited for a shorter period. Furthermore, using the citation mapping in the future, can provide the option to introduce a recommendation system based on papers that are frequently cited by the same sources.

Set criteria for appropriate selection of documents for forming a database for use at three levels:

(i) scientific community,
(ii) regulatory governmental bodies,
(iii) general public:

In order to set criteria for data base, we chose keywords to create queries for the crawler (automatic search engine), the keywords were specific for each level and combinations of queries were chosen to upload the maximum number of domain-relevant articles to the repository. We examined the relevance of the queries by comparing the results to those obtained by using Google Scholar search (taken on this context as a gold standard), Pubmed and Scopus. It appears that queries need to be adjusted for each of the databases (i.e. Google Scholar, Pubmed). Additionally, besides choosing scientific articles for the scientific community, the selection of documents for the two other levels was focused in particular: regulatory governmental bodies and general public. The work was mainly manual.
Outline the parameters needed for data-mining based evaluation of the data originating from in vitro and in vivo studies, epidemiological and environmental studies:

The main goal of this task was to extract the fields (well defined entities) from the paper and to formulate the concept of relations that will be extracted from the document for textual data-mining. Every field has a specific format, and a pre-defined set of valid values. For some of the fields like nanoparticle chemical name - the values were taken from the taxonomies; for other fields, we built a specific vocabulary of valid values. Example of such fields are: nanoparticle name, exposure model, characterisation method for nanoparticle.

Relation extraction appearance defined as number of fields/entities that appeared in a few sentences together or in other words, a relation represents a specific experiment and its result, which described in the article. The final excel file of each extracted article is supposed to present the relevant relation; the relation includes:

(i) nanoparticles name (chemistry),
(ii) experiment model: animal (species) or cell model,
(iii) experiment methods, measurement tools,
(iv) conditions,
(v) experiment results.

NHECD deals with three main expertise areas: in vivo / in vitro, environmental and occupational. Splitting efforts to build exhaustive schemas of all three areas due to the resources needed, would be detrimental to all the three areas, resulting in a more shallow, suboptimal schema; therefore, the main effort in outlining the parameters needed for textual data mining based evaluation and subsequently performing an initial textual data mining was concentrated on the two areas with a wide spectrum of publications (in vitro / in vivo and environmental), reaching a deeper level of knowledge and a detailed list of fields (and thus a more complete schema). For every domain, we have designed specific template / rules (according to the vocabularies) of the extracted relations, dividing it into well-defined subcategories. This stage was followed by initial and manual validation of quality control.

In the first phases of information extraction, we verified that all the fields are well defined, and can be tagged in the articles with reasonable recall and precision. On later phases, we verified that facts / relations could be extracted from the articles, according to the tagged fields and the skeletal structure of the articles. One of the challenges was to define the results fields (such as viability change) and to extract result terms from the papers. When finalising the definition of all fields and rules we were capable of having more accurate and complete information extraction as manifested by the validation process.

Select from meta-data a articles for advanced data mining :

The toxicology group annotated a training corpus of several dozens of representative research papers for information extraction automation; these articles were chosen out of the big repository which contains today (month 48) up to 10 000 articles.

About 400 representative articles were chosen manually for the toxicology domain (part of the chosen 400 articles dealt with gold nanoparticles in order to cover a specific topic for information extraction validation process). For the ecotoxicology domain and the occupational domains, around 200 articles were chosen carefully by the domain experts. The information extraction group picked 136 articles out of the 600 manually chosen representative articles and used them for the developing of the advanced text-mining. The articles were stored in PDF format and converted into text by the information extraction group. The main parameters for the selection of the corpus of articles were domain relevant queries which were improved throughout the project period; the queries included domain relevant queries such as: nanoparticle and toxicity, nanoparticle and model, and more. During the process, it was very important to choose articles as precise as possible. This is because the information extraction group needed to have relevant and appropriate articles for the continuous improvement of the algorithm; the selection was parallel and comparable process with the crawler validation.

During the project period, we carried out an ongoing task of crawler validation using the crawler validation tool prepared by the IT group. This process went on until the end of the project because:

(i) it was very important to improve the queries throughout the project in order to get accurate queries relevant to nano-toxicology,
(ii) it was necessary for the improvement of the selection of ‘large subset of full PDF articles’.
The results of the first and second rounds of the crawler validation task which were done at the last year of the project showed 80 - 90% precision which is an excellent result. Furthermore, we continued to validate the crawler through the end of the project by checking solely each specific query in the different domains in order to reveal where the specific wrong search word is.

During the last period of the project, the toxicology group was also focused on selecting a new set of papers relevant to the specific domains: in vivo - in vitro, ecotoxicity and the occupational domains. The articles were collected from the literature for two purposes:

(i) selection of new, more relevant, larger subset of full PDF articles for the information extraction group,
(ii) improvement of the queries which was done also by providing criteria for the papers selection; the scanning of the mesh terms of the relevant papers was very helpful in improving the query for indexed articles.

Ongoing validation of data mining development to optimise the quality of data extraction:

The main goal of the toxicology group was to increase the precision of data extraction and to validate the results quantitatively as will be detailed below. The purpose of the validation work was to develop the procedure that will serve as a gold standard in the future for machine learning. The quality of the automated information extraction depends very much on the requirements-set validated by the scientific domain experts. Throughout the project period the procedure was always interactive in order to optimise the quality of textual data extraction. The basics of the validation process didn't change throughout the project period; the only changes were introduced in the different technical validation procedures and in the information transfer between the domain expert and the information extraction group. The validation work was done by domain experts (in-vivo / in-vitro domain, ecotoxicology domain and occupational domain). The development of the validation process contained several steps:

- extraction of fields and relations followed by validation of these entities;
- validation of relations by excel tables, meaning - reading the relevant sentence in the article and validating whether each line is correct in the excel table; eventually, this type of validation turned out to be very inconvenient and insufficient and a new tool was developed to validate the relations.

The final and optimal validation approach that was carried out in the last year of the project used the new validation tool. This tool, developed by the information extraction group, extracts 5-40 relations for each article. Each relation contains one or two sentences from the article, reflecting a result from the specific analysed research; this result contains a minimum of three entities:

(i) nanoparticle name,
(ii) cell or animal model,
(iii) effect of the nanoparticle on the model, additionally the relation can contain entities like: assay name, nanoparticle concentration, exposure period of nanoparticle.

All the mistakes and / or additions at the validation procedure are transferred to the information extraction experts and can be applied in the next ‘round’ of validation.

Another change introduced in the last year of the project was to extract relations and validate them employing only the abstract and the results sections of the article. This was done due to the fact that the introduction and discussion sections of the article contain a large number of relation sentences that are not related to the specific research (for example reference citations).

Focusing the extraction only from the Abstract and Results sections was very effective and allowed the domain experts to:

- improve the recognition of the main entities;
- attain higher precision in the extraction from data.

In the last months of the project the data extraction group reported on the main improvements in the validation tool after receiving feedbacks from the experts (several ‘rounds’) as follows:

- correct extraction of abstract and results;
- improved recognition of species and cell models and reduced false identification of cell model and nanoparticles;
Therefore, the number of extracted relations in this version is 50% lower than in the previous version but the relations are more accurate.

One of our main goals for the final validation task was to provide statistics on the precision in order to evaluate quantitatively the data extraction. The precision results yielded an approximately 60% correct relations out of the total relations that were extracted from the articles (from the abstract and results sections). This is considered to be an excellent precision result.

In summary, the validation procedure which is one of the main goals of the toxicology group, was a repetitive iteration between the automatic information extraction and its manual validation (by the experts of the different domains) and turned out to be very successful; the extraction of the textual results from articles was improved significantly; most of the entities are now being recognised correctly and the relations are more accurate. The validation process was an important task in the NHECD project because it reflected the quality control of the developed process which have finished these days.

3. Tools and environment

The objective of this work package was to supply and maintain the tools and infrastructure for NHECD, including hardware, software and networking services. The primary software includes an open source content management, and the database and website front-end of the project, accessible to all the users per their definitions. We provided, on an ongoing basis, system administration and management services including user management (anonymous users, registered users), security and access management, backup and recovery, and more.

The main task of this work package included:

- detailed system architecture design of test and production environments to meet the planned requirements (in terms of: volumes, expected number of users, required quality of service, security requirements), including data base administration services;
- acquisition and setup of system hardware and software as per the actual design;
- setup of system management and administration facilities to monitor the system status and maintain required quality of service on an ongoing 7x24 basis;
- execution of test scripts to validate the software and data quality;
- definition of the requirements and supervision of the GUI design to optimise the user experience and satisfaction for the three types of user communities.

The core component of this work package was the definition of the requirements and the design of the GUI of the system. It required tight collaboration between the partners to reflect the complexity of the underlying scientific subject dealt with. It also required the expertise of both the team members and the GUI experts to obtain a 'product' that reflects the vision of the nanoparticles scientists, regulators and the public at large as they were seen on our preliminary studies.

The GUI design process is by nature an iterative process. During the design, the team together with the GUI expert led many discussions with project partners, domain experts and stakeholders to devise as precise as possible requirements and functionality to be implemented through the GUI.

This process included the following steps: conceptual model stage included figuring out what are the features the system will supply requirements gathering including writing the software requirements specification with the functional and non-functional requirements; UI design by expert based on analysis done by project team.

The implementation of the UI was done by several teams working together to achieve the goal of setting a testing and development environments of the system (development and content wise).

The first public version of NHECD’s website was deployed at JRC site on September 2011. Ever since, there was an ongoing routine of manual tests of the system to ensure the behaviour of the features and the quality of the data. In addition, we performed extensive user
feedbacks on the system. The results and fixes that stem from the outcome of those procedures were incorporate into the production system on a regular basis.

We provided an operational environment that meets the long term required service. The website (final version) is a comprehensive solution that includes many features supporting the user experience and needs.

The final operation consisted into transfering the back-end of the project to JRC and adapting it to the IT environment. A full operational system including front and back end is operational since January 2013.

Now, the project has reached the end of its lifespan, and the JRC will be responsible the database is kept up-to-date as and when new articles enter the public domain. The main effort in the near future (after project end) will be to get in touch with major scientific editors (Elsevier, Wiley, Thomson Reuters, ACS) in order to get access to their article database, thus expanding significantly the repository coverage.

Potential impact:

Nanotechnology has been identified as a key enabling technology (KET) providing the basis for further innovation and new products by the EU. In their second regulatory review on nanomaterials (COM(2012)572Final), the potential benefits also needs to be accompanied by a better understanding of the impact of nanoparticles on health and the environment. The EASAC-JRC report from 2011 (entitled impact of engineered nanomaterials on health: consideration for benefit-risk assessment) indicates that the advance of nanotechnology poses the question how to deal with the uncertainty when there is insufficient knowledge regarding health impacts. Questions emerge at different 'levels' and range from the following:

- very specific scientific queries on how to understand the interaction of nanomaterials with the human body;
- concerns of consumers on the safety of products and the general benefits for the use of nanotechnology;
- policy questions on how to address safety issues and concerns from the regulatory side and how to develop appropriate governance systems to cope with the novelties of nanotechnology.

Activities on the governance of nanotechnologies should encompass all issues related to EHS and take due account of ethical, legal and social aspects (ELSA). This requires the use of appropriate instruments including the following:

- knowledge gathering;
- self-regulation and voluntary measures;
- regulation by adaptation of the existing regulatory framework;
- transnational collaboration.

The NHECD-database has the potential impact to fulfil especially the first point, knowledge gathering, at the different societal levels. The potential impact will be highlighted by the different aspects considered in the project.

Better understanding of the impact of nanoparticles on health and the environment and definition of future actions:

The NHECD system is the first in its kind that is able to automatically extract scientific information from peer-reviewed papers. The information extraction process is applied to environmental health and safety aspects of manufactured nanomaterials. The database is organised in such a way that different user communities can quickly and easily locate and retrieve the data relevant to their respective needs. In order to define future actions, current information must be retrieved and digested. However, the type of information that different user communities needs is different.

The repository includes several options that are considered to be helpful for the different user communities. Impact is often defined by specific parameters that can be found in the developed taxonomies. These parameters can be used as search query in the database.

The impact on health and the environment is related to similar findings from different research groups. NHECD extracts information from peer-reviewed references to establish relations between entities in the collected references. These relations are found in the excel file that is produced and as such indicates what entities are related to each other. The system includes a ranking system for each peer-reviewed reference that is based on the journal impact factor and the journal half-life.
Safe and responsible development and use of nanotechnology:

Innovation of new products in the area of nanotechnology is often hampered by a lack of knowledge on the health and environmental effects of the nanomaterials that are used in the production process. NHECD stores the majority of the peer-reviewed references on EHS of nanomaterials and a collection of other papers and reports (White Papers) in its repository. The taxonomies include a large list of nanomaterials that are used in the described experiments. When using this taxonomy, it is also possible to see whether the retrieved reference has been included in the information extraction process (by clicking on the icon N in the result page). If included the software indicates a relation between different entities. In this way it can be used in an early stage to obtain relevant information on a safe use of nanomaterials in the design phase of the product.

Support to research and regulation; support to regulatory measures and implementation of legislation:

NHECD supports research and regulation in a number of ways. For researchers the system needs to give as much as possible all perceived relevant references and as little as possible the irrelevant ones (noise). NHECD includes a crawling system that retrieves automatically the relevant peer-reviewed papers into a repository. The efficiency of this process reached a level higher than 80 %. During the test phase new users have indicated that the retrieved references from this repository were all relevant for their research. Little 'noise' was experienced. However, only continuous use (and feedback) can verify the relevancy of the peer-reviewed references in the repository.

For regulation the support is in two ways:

(a) stated information can be checked quickly. Since a lot of 'noise' or irrelevant papers are left out the number of papers to check substantially decreases,
(b) the result of the information extraction indicates whether relations between entities in the reference are found.

This is a measure of the relevant information the particular reference contains. Especially for the support of research the quality of the precision of the information extraction is around 60 %. Priority in the development of NHECD is on a high precision so as to avoid as much as possible irrelevant relations.

NHECD contains White Papers. These documents are composed by other institutes and bodies. These documents are uploaded manually and this can be performed by the user after logging in. The repository contains a large number of white papers collected during the running period of the project and organised according to specific headings. These White Papers support then especially the non-scientific community in developing position papers.

The implementation of legislation on the health and safety aspects of nanomaterials in the EU is largely centred on the information required for the REACH registration dossier. NHECD can fulfil a specific and important role in this aspect. The registrant obtains an overview of highly relevant peer-reviewed articles on the requested endpoints without a lot of irrelevant information. The evaluator of the registration dossier can quickly check whether relevant references are missing.

The information extraction process indicates if a relation between relevant endpoints and other entities is present in the references collected in the repository.

Implementation of the EC's action plan for nanotechnology:

Part of the EC action plan for nanotechnology is the development of a regulatory framework. NHECD can give relevant information of specific EHS issues for example by using (a combination) of entities identified in the different taxonomies.

NHECD can fulfil a specific role in new FP7 projects or its successors developed to investigate specific points in the EC action plan of nanotechnology. NHECD can be applied to quickly give an overview of the relevant literature in the area of nanosafety and health but also in a specific work package in the proposal to improve and extend the information extraction of NHECD.

Free access, Support to good governance:
NHECD is accessible free of charge. The different concepts are described in the website’s frequently asked questions (FAQ). In addition, it is possible to upload papers into the white paper section and to request comments on these papers. The high number of relevant references retrieved and the high precision of the extracted relations show a high quality of the software used in the database. The latest user questionnaire indicated that all participants will NHECD in the future in case it is public and free of charge.

Dissemination activities and exploitation results:

Input collected from different user groups (research, regulation and civil society organisations) in two polls to around 1500 persons was used to develop and improve NHECD. This also had the function to increase the awareness of the future system and as a source of specific feedback once the system was in function.

The specific features of NHECD were put forward in presentations given at the Euronanoform meeting in 2011 and the NaNOSH conference in 2009. Live-presentations of NHECD showing how the system works and what it can do were held specifically at ISO and CEN meetings and for the nanoworkgroup of the European Trade Union Institute as example of a civil society organisation. NHECD was included in other presentations as source of information at the OECD workshop of nanotoxicology and occupational safety and health workshop organised by Directorate General (DG) ‘Employment’. NHECD was presented and included in several presentations in the nanosafety cluster meetings between the EU and the United States (US). In these meetings a common action program is defined of which database and data mining and exposure assessment methods in the life-cycle are two focus areas. NHECD can be used as a source of available information that can be retrieved quickly and easily and gives also a quick idea on the relevancy of the information through the use of the information extraction software. NHECD will be introduced by ECHA by one of the advisory board members who is a national representative of the committee for risk assessment.

The number of unique visitor rose steadily to around 80 per month after the system was launched in October 2011. The latest user questionnaire indicated that that more than 90% of the participants would recommend NHECD to colleagues.

Project website: http://www.NHECD-fp7.eu

This project is featured in...

RESEARCH*EU MAGAZINE
Research for regions: a coherent research area for Europe
Issue 3, June 2011

Share this page