Data Integration: A Case Study

Nina Jeliazkova

Ideaconsult Ltd., Angel Kanchev 4, 1000 Sofia, Bulgaria

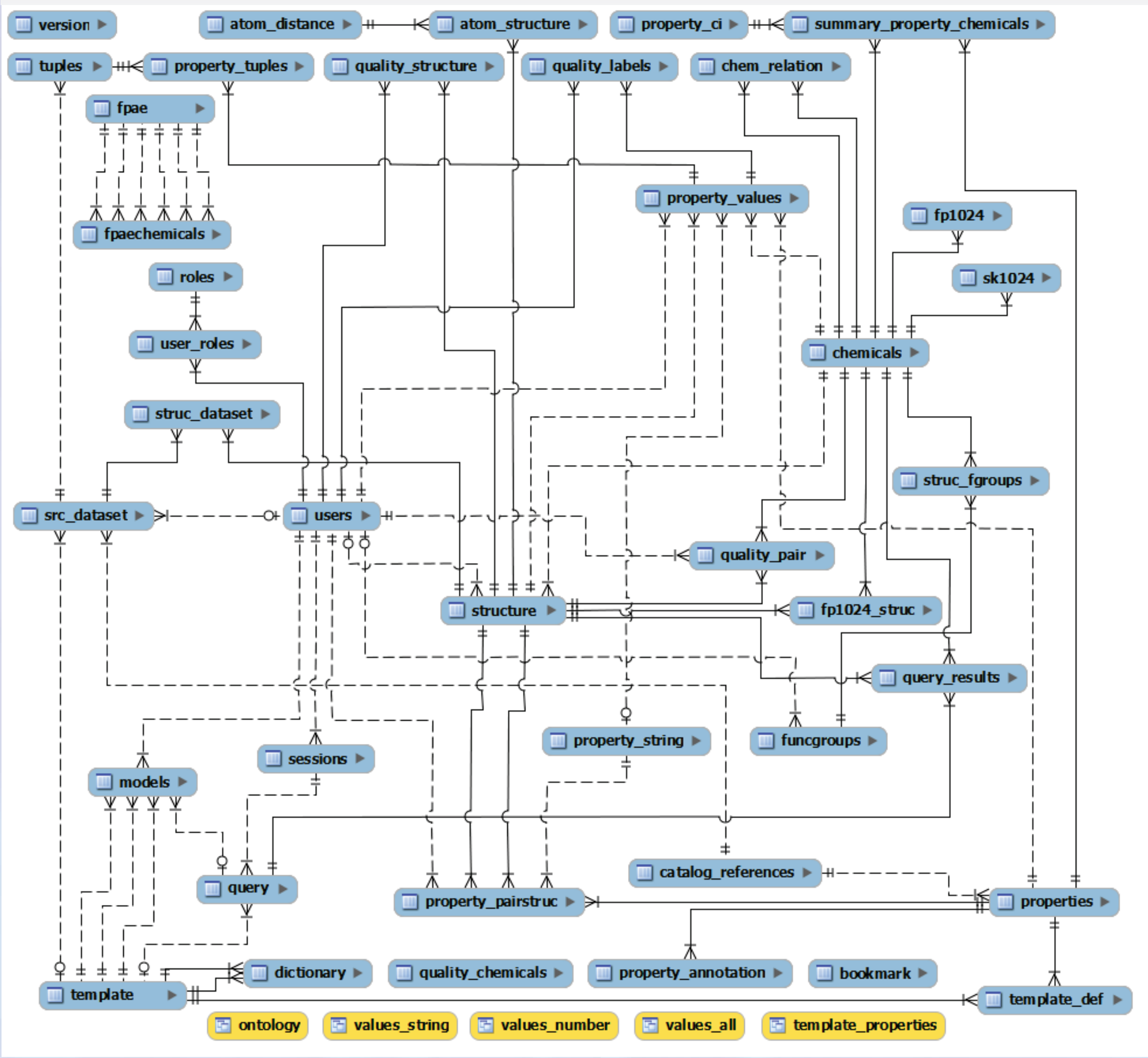
Tel. +359 886802011. Fax +359 29526265. E-mail: jeliazkova.nina@gmail.com

Introduction

An important pre-requisite for the successful implementation of the main principles of the 3Rs - Reduction, Refinement and Replacement alternatives - is the universal access to high quality experimental data on various chemical properties. Unfortunately, even today, the "state-of-the-art" is characterised by highly fragmented and unconnected life sciences data (both from a physical and ontological perspective), which is furthermore frequently inaccurate and/or difficult or even impossible to find or access. We present an AMBIT web services-based case study of integration and comparison of 67 datasets with physico-chemical and/or experimental toxicity data, originating from various public and commercial sources. The datasets can be retrieved as a whole or by submission of search queries on chemical identifiers, properties, structures, sub-structures or similarity. The content from different datasets can be easily collated, thanks to the universal database structure design and the ontology that establishes a shared terminology and meaning of the data fields. Additionally, the datasets are "model-ready", and can be used as an input to OpenTox compliant predictive models. Query and submission of new data or data modifications can be carried out through the web services interface, which implements the OpenTox framework API.

Chemical structures (including ECHA's list of pre-registered substances[1]) have been collected from various public sources and/or generated by name to structure conversion[2]. In this process inconsistencies between chemical structures have been discovered and flagged automatically through built-in heuristics. We report the overlap between the datasets in terms of number of common compounds, as well as mutual similarity or coverage of the "chemical domain", calculated by structure-based and descriptor based methods. Several computational resources and predictive methods are seamlessly integrated by the uniform OpenTox application programming interface. The similarity between a user supplied set of chemicals and the existing datasets can be conveniently accessed online, either programmatically or via a web browser.

AMBIT Database



The AMBIT database is a relational database, consisting of several repositories for compounds, properties, QSAR models, users, references, as well as several tables containing pre-processed information which allows speeding up substructure and similarity queries. The current implementation is based on MySQL[3]. An overview of the entity-relationship diagram of the database is provided in the above figure.

Datasets Included in this Study

Dataset Name	Number of Compounds	Uniform Resource Identifier (URI)
ECHA list of pre-registered substances	143835	https://ambit.uni-plovdiv.bg:8443/ambit2/dataset/1?pagesize=10&page=0
Chemical Identifier Resolver[4]	72985	https://ambit.uni-plovdiv.bg:8443/ambit2/dataset/5?pagesize=10&page=0
ChemIDplus[5]	80468	https://ambit.uni-plovdiv.bg:8443/ambit2/dataset/3?pagesize=10&page=0
ChemDraw[6]	22519	https://ambit.uni-plovdiv.bg:8443/ambit2/dataset/7?pagesize=10&page=0
ECBPRS[7]	80410	https://ambit.uni-plovdiv.bg:8443/ambit2/dataset/6483?pagesize=10&page=0
ISSCAN[8]	1150	https://ambit.uni-plovdiv.bg:8443/ambit2/dataset/31?pagesize=10&page=0
ISSMIC	151	https://ambit.uni-plovdiv.bg:8443/ambit2/dataset/33?pagesize=10&page=0
ISSSTY	223	https://ambit.uni-plovdiv.bg:8443/ambit2/dataset/35?pagesize=10&page=0
CPDBAS[9]	1547	https://ambit.uni-plovdiv.bg:8443/ambit2/dataset/9?pagesize=10&page=0
DBPCAN	209	https://ambit.uni-plovdiv.bg:8443/ambit2/dataset/11?pagesize=10&page=0
EPAFHM	617	https://ambit.uni-plovdiv.bg:8443/ambit2/dataset/13?pagesize=10&page=0
FDAMDD	1216	https://ambit.uni-plovdiv.bg:8443/ambit2/dataset/15?pagesize=10&page=0
HPVCSI	3548	https://ambit.uni-plovdiv.bg:8443/ambit2/dataset/17?pagesize=10&page=0
HPVISD	1006	https://ambit.uni-plovdiv.bg:8443/ambit2/dataset/19?pagesize=10&page=0
IRISTR	544	https://ambit.uni-plovdiv.bg:8443/ambit2/dataset/21?pagesize=10&page=0
KIERBL	278	https://ambit.uni-plovdiv.bg:8443/ambit2/dataset/23?pagesize=10&page=0
NCTRER	232	https://ambit.uni-plovdiv.bg:8443/ambit2/dataset/25?pagesize=10&page=0
NTPBSI	2330	https://ambit.uni-plovdiv.bg:8443/ambit2/dataset/27?pagesize=10&page=0
NTPHTS	1408	https://ambit.uni-plovdiv.bg:8443/ambit2/dataset/29?pagesize=10&page=0
TOXCST[10]	320	https://ambit.uni-plovdiv.bg:8443/ambit2/dataset/37?pagesize=10&page=0
TOXCST_ACEA	320	https://ambit.uni-plovdiv.bg:8443/ambit2/dataset/83?pagesize=10&page=0
TOXCST_Attagene	320	https://ambit.uni-plovdiv.bg:8443/ambit2/dataset/723?pagesize=10&page=0
TOXCST_BioSeek	320	https://ambit.uni-plovdiv.bg:8443/ambit2/dataset/1363?pagesize=10&page=0
TOXCST_Cellumen	320	https://ambit.uni-plovdiv.bg:8443/ambit2/dataset/2003?pagesize=10&page=0
TOXCST_CellzDirect	320	https://ambit.uni-plovdiv.bg:8443/ambit2/dataset/2643?pagesize=10&page=0
TOXCST_Gentronix	320	https://ambit.uni-plovdiv.bg:8443/ambit2/dataset/3283?pagesize=10&page=0
TOXCST_NCGC	320	https://ambit.uni-plovdiv.bg:8443/ambit2/dataset/3923?pagesize=10&page=0
TOXCST_Novascreen	320	https://ambit.uni-plovdiv.bg:8443/ambit2/dataset/4563?pagesize=10&page=0
TOXCST_Solidus	320	https://ambit.uni-plovdiv.bg:8443/ambit2/dataset/5203?pagesize=10&page=0
TOXCST_ToxRefDB	320	https://ambit.uni-plovdiv.bg:8443/ambit2/dataset/5843?pagesize=10&page=0
TXCST2	960	https://ambit.uni-plovdiv.bg:8443/ambit2/dataset/39?pagesize=10&page=0
ECETOC Technical Report No. 66 Skin irritation and corrosion Reference Chemicals data base (1995)	176	https://ambit.uni-plovdiv.bg:8443/ambit2/dataset/41?pagesize=10&page=0
Local Lymph Node Data for the Evaluation of Skin Sensitization - Compilation of historical data (2005)	209	https://ambit.uni-plovdiv.bg:8443/ambit2/dataset/43?pagesize=10&page=0
Local Lymph Node Data for the Evaluation of Skin Sensitization - Second compilation (2010)	108	https://ambit.uni-plovdiv.bg:8443/ambit2/dataset/45?pagesize=10&page=0
Bioconcentration factor (BCF) Gold Standard Database	1130	https://ambit.uni-plovdiv.bg:8443/ambit2/dataset/47?pagesize=10&page=0
Benchmark Data Set for pKa Prediction of Monoprotic Small Molecules the SMARTS Way	185	https://ambit.uni-plovdiv.bg:8443/ambit2/dataset/49?pagesize=10&page=0
Benchmark Data Set for In Silico Prediction of Ames Mutagenicity	6197	https://ambit.uni-plovdiv.bg:8443/ambit2/dataset/51?pagesize=10&page=0
Bursi AMES Toxicity Dataset	4337	https://ambit.uni-plovdiv.bg:8443/ambit2/dataset/53?pagesize=10&page=0
EPI_AOP[11]	818	https://ambit.uni-plovdiv.bg:8443/ambit2/dataset/55?pagesize=10&page=0
EPI_BCF	685	https://ambit.uni-plovdiv.bg:8443/ambit2/dataset/57?pagesize=10&page=0
EPI_BioHC	175	https://ambit.uni-plovdiv.bg:8443/ambit2/dataset/59?pagesize=10&page=0
EPI_Biowin	1263	https://ambit.uni-plovdiv.bg:8443/ambit2/dataset/61?pagesize=10&page=0
EPI_Boil_Pt	5890	https://ambit.uni-plovdiv.bg:8443/ambit2/dataset/63?pagesize=10&page=0
EPI_Henry	1829	https://ambit.uni-plovdiv.bg:8443/ambit2/dataset/65?pagesize=10&page=0
EPI_KM	631	https://ambit.uni-plovdiv.bg:8443/ambit2/dataset/67?pagesize=10&page=0
EPI_KOA	308	https://ambit.uni-plovdiv.bg:8443/ambit2/dataset/69?pagesize=10&page=0
EPI_Kowwin	15809	https://ambit.uni-plovdiv.bg:8443/ambit2/dataset/71?pagesize=10&page=0
EPI_Melt_Pt	10051	https://ambit.uni-plovdiv.bg:8443/ambit2/dataset/73?pagesize=10&page=0
EPI_PCKOC	788	https://ambit.uni-plovdiv.bg:8443/ambit2/dataset/75?pagesize=10&page=0
EPI_VP	3037	https://ambit.uni-plovdiv.bg:8443/ambit2/dataset/77?pagesize=10&page=0
EPI_WaterFrag	5764	https://ambit.uni-plovdiv.bg:8443/ambit2/dataset/79?pagesize=10&page=0
EPI_Wskowwin	2348	https://ambit.uni-plovdiv.bg:8443/ambit2/dataset/81?pagesize=10&page=0
NAME2STRUCTURE (OPSIN)	70646	https://ambit.uni-plovdiv.bg:8443/ambit2/dataset/6487?pagesize=10&page=0
PubChem Structures + Assays[12]	473965	https://ambit.uni-plovdiv.bg:8443/ambit2/dataset/6489?pagesize=10&page=0
Leadscope_carc_level_2*	2988	https://ambit.uni-plovdiv.bg:8443/ambit2/dataset/6490?pagesize=10&page=0
Leadscope_ccris_genetox*	8001	https://ambit.uni-plovdiv.bg:8443/ambit2/dataset/6492?pagesize=10&page=0
Leadscope__cder_chronic*	121	https://ambit.uni-plovdiv.bg:8443/ambit2/dataset/6494?pagesize=10&page=0
Leadscope_cder_genetox*	336	https://ambit.uni-plovdiv.bg:8443/ambit2/dataset/6496?pagesize=10&page=0
Leadscope_cder_repro_dev*	58	https://ambit.uni-plovdiv.bg:8443/ambit2/dataset/6498?pagesize=10&page=0
Leadscope_cfsan_acute*	1070	https://ambit.uni-plovdiv.bg:8443/ambit2/dataset/6500?pagesize=10&page=0
Leadscope_cfsan_chronic*	655	https://ambit.uni-plovdiv.bg:8443/ambit2/dataset/6502?pagesize=10&page=0
Leadscope_cfsan_genetox*	696	https://ambit.uni-plovdiv.bg:8443/ambit2/dataset/6504?pagesize=10&page=0
Leadscope_cfsan_repro_dev*	312	https://ambit.uni-plovdiv.bg:8443/ambit2/dataset/6506?pagesize=10&page=0
Leadscope_fda_marketed_drugs*	6637	https://ambit.uni-plovdiv.bg:8443/ambit2/dataset/6508?pagesize=10&page=0
Leadscope_genetox_level_2*	10155	https://ambit.uni-plovdiv.bg:8443/ambit2/dataset/6510?pagesize=10&page=0
Leadscope_ntp_genetox*	2128	https://ambit.uni-plovdiv.bg:8443/ambit2/dataset/6512?pagesize=10&page=0
Pharmatrope_AERS_hepatobiliary_system*	1274	https://ambit.uni-plovdiv.bg:8443/ambit2/dataset/6514?pagesize=10&page=0

(*) datasets accessible solely to OpenTox partners, due to restrictions imposed by the corresponding license agreements

Results and Discussion

The overlap between datasets with experimental data and ECHA's list of pre-registered substances is in the range from 0.01% to 5.2% of ECHA's entries. The EPI datasets with physicochemical properties have the largest overlap with ECHA's list (Melting point 5.2%, Boiling point 3.4%, Kow 3.2%), followed by PubChem (4.3%), commercial Leadscope genetox datasets (2.3-2.84%) and various carcinogenicity and mutagenicity datasets (0.4-1.7%). The overlap of ToxCast phase I data is as small as 0.2% of ECHA's entries. Interestingly, from the 473965 PubChem structures with associated assay data, used in this study, only 1.3% (6198) are part of ECHA's list of pre-registered substances. Data compilation and fusion, combined with the OpenTox web service API, enables flexible automated retrieval and collation of properties for a given query compound.

[1] ECHA's list of pre-registered substances http://apps.echa.europa.eu/preregistered/prsDownload.aspx
[2] Chemical Name to Structure: OPSIN, an Open Source Solution Daniel M. Lowe, Peter T. Corbett, Peter Murray-Rust, Robert C. Glen *Journal of Chemical Information and Modeling* 2011 *51* (3), 739-753
[3] MySQL http://www.mysql.com/
[4] Chemical Identifier Resolver http://cactus.nci.nih.gov/chemical/structure
[5] ChemIDplus http://chem.sis.nlm.nih.gov/chemidplus/
[6] ChemDraw http://www.cambridgesoft.com/software/ChemDraw
[7] Ex-European Chemicals Bureau pre-registered substances list
[8] ISSTOX Chemical Toxicity Databases http://www.iss.it/meca/dati/cont.php?id=199&lang=1&tipo=25
[9] DSSTox http://www.epa.gov/ncct/dsstox/index.html
[10] ToxCast http://www.epa.gov/ncct/toxcast/
[11] EPI Suite data http://esc.syrres.com/interkow/EpiSuiteData.htm
[12] Downloadable Structure Files of PubChem Compounds http://cactus.nci.nih.gov/download/roadmap/

www.OpenTox.org

OpenTox - An Open Source Predictive Toxicology Framework, funded under the EU Seventh Framework Program: HEALTH-2007-1.3-3 Promotion, development, validation, acceptance and implementation of QSARs (Quantitative Structure-Activity Relationships) for toxicology, Project Reference Number Health-F5-2008-200787 (2008-2011).