Conceptual Model Enhancing Accessibility of Data from Cancer–Related Environmental Risk Assessment Studies

Ladislav Dušek, Jiří Hřebíček, Miroslav Kubásek, Jiří Jarkovský, Jiří Kalina, Roman Baroš, Zdeňka Bednářová, Jana Klánová, Ivan Holoubek

Conceptual model enhancing accessibility of data from cancer–related environmental risk assessment studies

Ladislav Dušek[1,2], Jiří Hřebíček[1], Miroslav Kubásek[1], Jiří Jarkovský[1], Jiří Kalina[1], Roman Baroš[2], Zdeňka Bednářová[2], Jana Klánová[2], Ivan Holoubek[2]

[1] Institute of Biostatistics and Analyses, Masaryk University, Kamenice 128/3, 625 00 Brno, Czech Republic
[2] Research Centre for Toxic Compounds in the Environment, Masaryk University, Kamenice 128/3, 625 00 Brno, Czech Republic
{dusek,hrebicek,kubasek,jarkovsky,kalina}@iba.muni.cz,
{baros, bednarova, klanova, holoubek}@recetox.muni.cz

Abstract. This paper proposes conceptual model which can be used to facilitate the discovery, integration and analysis of environmental data in cancer-related risk studies. Persistent organic pollutants were chosen as a model because of their persistence, bioaccumulation potential and genotoxicity. Part dealing with cancer risk is primarily focused on population-based observations encompassing a wide range of epidemiologic studies, from local investigations to national cancer registries. The proposed model adopted multilayer hierarchy working with characteristics of given entities (POPs, cancer diseases as *nomenclature classes*) and couples "*observation – measurement*" as content defining classes. The proposal extends formally used taxonomy applying multidimensional set of descriptors including scores of measurement validity and precision. This solution has the potential to aid multidisciplinary data discovery and knowledge mining. The same structure of descriptors used for environmental and cancer part enables the users to integrate different data sources recognizing their methodical origin, time & space coordinates and validity.

Keywords: Persistent organic pollutants, cancer risk, data model, data discovery

1 Introducing problems with data accessibility

"Data rich – information poor" is becoming obligatory phrase or accepted "professional dialect" associated with environmental monitoring. It also extends to the cancer risk assessment which has recently attracted increasing attention. Most problems can be explained by the heterogeneity of input data ranging from laboratory bio-tests to multilevel epidemiologic observations. Progress increasingly requires standardized access to multi-disciplinary information resources, including chemical, geological, meteorological, epidemiologic and demographic data. Each broadly ranged ecological or human risk study must adopt both following scenarios [1,2]:

1. retrospective exploitation of data sources and their description in discovery process
2. prospective arrangement enabling effective electronic data capture in future

From the viewpoint of informatics, environmental risk assessment can be characterized as processing of heterogeneous data leading to probabilistic estimation of some uncertain (prospective approach) or on the other hand relatively certain (retrospective approach) risk event. Main complications that hamper progress in this field are highlighted in the following list:

1. Extremely wide range of data types and structures in environmental studies
2. Insufficient metadata description and standardization
3. Lack of well established repositories based on standardized protocols which is in strong contrast to methodical progress in environmental and medical sciences
4. Variability of technologies, coding and reporting systems used by different research groups
5. Growing number of small and not adequately published and described studies, which however produce valuable and important data.

Especially last point deserves special attention. Growing number of studies is not accompanied with adequate progress in information technologies and in practical implementations of SW tools [3,4]. It inevitably results in publishing of non-consistent outcomes with ad hoc data management support. To discover such broadly heterogeneous data we need consensus on data and metadata standardization, but it itself is not enough. We need sufficiently complex conceptual models, advancing development of formal ontologies over environmental and epidemiological data capture systems.

Although there are some usable standardizing concepts already published (Ecological Metadata Language)[1] [5,6], they are not extensively used in practice or there is a lack in support in data capture systems. Environmental data collection is still subjected to research in the informatics field [7,8]. In cancer research, we can take the advantage of accessible nomenclature standards like Clinical Terms (SNOMED-CT)[2], the Unified Medical Language System (UMLS)[3], and the National Cancer Institute Thesaurus (NCIT)[4]. Several consistent attempts to design ontologies functional for cancer treatment and management have also been published recently [9].

Nevertheless, widespread support for ontology-based approaches is not implemented in the field ecological risk assessment [10]. However, interest in developing ontologies is growing, because new synthetic environmental analyses increasingly rely on access to a broad range of cross-disciplinary data sources and monitoring studies. The effective system should encompass not only structure and content of such data repositories, but also hierarchical architecture and mutual relationships among components [11].

That is why we try to propose multidisciplinary conceptual model with ambition to discover and process data on environmental pollution and cancer risk using the same

[1] http://knb.ecoinformatics.org/software/eml/
[2] http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html
[3] http://www.nlm.nih.gov/research/umls/
[4] http://www.mindswap.org/2003/CancerOntology/

methodical template. Our principal aim is to support data integration across the geographical borders, disciplines and professional terminologies; as well as integration of newly gathered data with data already collected and archived.

2 Persistent organic pollutants as model for cancer risk studies

Nowadays, many environmental factors, both chemical and physical, are proved to be involved in causing cancer. In our proposal of conceptual data model for cancer risk studies, we took persistent organic pollutants (POPs) as a proper model namely due to the following reasons:

- POPs have become recently intensively studied due to their properties which represent remarkable risk for ecosystems and human population (persistence, bioaccumulation, carcinogenicity, genotoxicity). The ability to accumulate in human tissues together with persistence enable long term exposure of an individual, the effect of which is further enhanced by genotoxic impact.
- Clear nomenclature of POPs which facilitates development of formal concepts based on well defined entities.
- There are important international initiatives to control distribution and associated impacts of POPs; participating countries and institutions form growing family of potential users of developed informatics solutions. The decisions by the Conferences of the Parties to the Basel, Rotterdam and Stockholm conventions on enhancing cooperation and coordination among the three conventions invite Parties to coordinate their efforts when implementing the three conventions to ensure close cooperation among relevant ministries and programmes at the national levels [12].

Not all POPs are proved as direct carcinogens or co-carcinogens. There are several widely accepted systems reporting data on carcinogenicity (IARC: International Agency for Research on Cancer; US EPA: United States Environmental Protection Agency; ACGIH: American Conference of Governmental Industrial Hygienists, see also Table 1). Therefore, this property is an important classifier extending information content given by nomenclature itself.

3 Cancer epidemiology as endpoint in risk studies

Growing cancer burden affects visible proportion of worldwide population and logically attracts research interest. Furthermore, the era of personalized medicine put the cancer diseases to the position of primary target of IT support [13,14]. Studies focused on cancer epidemiology address the virtual top of the system of cancer research. Cancer is however entirely multifactorial problem with its roots in molecular mechanisms inside the cells. Therefore, the main added value of IT is to facilitate integration of data acquired from multiple levels of investigation. The knowledge of mechanisms explaining cancer origin can mostly explain changes observed in the population level, or at least minimize the bias in interpretation.

Cancer–related environmental risk studies require very comprehensive data background. We typically evaluate probability of increased risk for a given population, typically exposed to some dangerous factor. Prospective studies work namely with laboratory data and estimate intensity of probable exposure pathways. Retrospective studies utilize accessible epidemiologic observations performed in a target area. Estimates of incidence, mortality or prevalence are employed as frequent en-points.

Most of the investigations studying harmful effect of chemicals on population cancer burden are designed as case control studies, which often determine their frequently inconclusive results. We can mention problems with mutual variations in outcomes of similar studies, insufficient power of accessible retrospective records on individual environmental exposures, lack of biomarkers reflecting stages of the carcinogenic process or problems with sufficiently long time series of observation [15,16]. Origin of analyzed data, precision and validity of measurements are therefore very important attributes to be followed in these studies.

4 Proposed conceptual model and reasoning of its structure

Here we proposed simplified conceptual model that should broaden our capability for understanding the validity, content and relevance of the data coming from environmental and epidemiologic monitoring (Table 1). The adopted concept should support scientists in mapping of cancer-risk. The proposed model is based on hierarchically layered architecture providing different levels of classifiers or properties of homologous entities as well as scoring of data origin. The model works with three principle layers:

1. **Entities** (POPs or cancer diagnoses) defined on the basis of internationally standardized nomenclature systems. The level is linked with classifiers, i.e. given properties extending the nomenclature and filtering homologous groups of entities.
2. **Observation – measurement** level and its descriptors, focused namely on time & space coordinates, methodical attributes, measured endpoints, reference benchmarking of their value and validity scoring.
3. **Content** identification describing employed measures, units and precision estimates.

Important principles applied in conceptual model construction are further summarized here:

1. Reduction of the number of object properties is important and practical; when necessary, the set of attributes can be expanded before specifically designed data discovery. Too many object properties cannot be utilized efficiently in retrospective exploitation of the resources.
2. Any relevant data discovery must reflect heterogeneity of experimental and methodical approaches at ecosystem and population level. That is why type of the study or data resource is obligatory attribute among the observation descriptors.

3. A measured value cannot be interpreted without reference to a defined/known measurement standard or reference benchmarks. Both internal reference norms (e.g. self-benchmarking of time series data) and external benchmarks (e.g. background concentration levels or limits, hygienic norms, detection limit of applied method, international epidemiologic burden) are used.
4. Descriptors of measures must fulfill obligatory Measurement Standard, i.e. the units, scales and lists of attributes defining origin of the measures (e.g. examined matrix, sampling methods, investigated population, cohort, demographic selection etc.)

We proposed the same discovery template for POPs and cancer resources as it is summarized in Table 1. From the first step, we must categorize the key subjects, i.e. POPs and cancer diagnoses. At this obligatory tier, internationally validated nomenclature is recommended and summarized in Tables 2 and 3. The subsequent tiers gradually unravel attributes important for interpretation of cancer risk studies (e.g. carcinogenicity of POPs, malignant/benign classification of cancers, data origin in terms of study design, etc.). These layers form multidimensional descriptive space which is significantly more robust than any single formal classification. This minimizes the probability of missing or omitting of some important facts and protects the solution against selection bias or misinterpretation.

The same template used for POPs and cancer risk enables the IT tools to integrate these data resources. The relevance of the integration process relies on the ability to determine if two values (studies) are compatible, not only in time and space coordinates. Description of model levels in table 1 implies interdisciplinary interactions of classifiers extending the nomenclature (e.g. “carcinogenic compound x malignant neoplasm”), observation – measurement validity identifiers (e.g. long-term national monitoring of POPs x national cancer registry) and mutually related observation-measurement (e.g. trend in POPs concentration in food chains vs. disease specific mortality due to GIT cancers). The most important added value of the model is the capability to determine if two data sets can be either fully or partially merged or mutually related once they are discovered. To decide it, the system undertakes important steps in all levels of proposed architecture:

1. The system must control relevance and compatibility at the taxonomic level (nomenclature) and in space & time localization of data resources
2. Identified data resources must be assessed if, and at what semantic resolution, the data are compatible (level of classifiers and/or extending descriptors like type of the study, etc.)
3. Finally, the measurement standards for the mutually related environmental and epidemiologic endpoints have to be controlled for compatibility (units, scales, reference benchmarking, methodical origin).

Obviously, not all descriptors must be necessarily fulfilled in all data discovery sessions, sometimes the uncertainty is too high. Different situations give to different tiers different weights. For example, the situation is thoroughly different if someone needs well designed retrospective or prospective assessment than if it is sudden catastrophic situation like exposure due to industrial accident where we must in first line mitigate the immediate effects. Moreover environmental factors cannot necessarily impact upon human population is some isolated system, highly probably

they interact with other harmful effects associated with life style, occupational factors and relating exposures causing probable carcinogenic synergies. It complicates the interpretation of population risk studies where precise and standardized description of input data is becoming a key step limiting the relevance of reached outcome.

5 Projection of proposed conceptual model to data standards

The quality of conceptual model determines its utility for assisting in data discovery and information searching. However, the applicability of any such model strongly depends on quality of description and content of processed data sets. That is why, we should insist on minimized, but obligatory database components, i.e. limited number of entities and their descriptors. Minimized data model as standard can be used both retrospectively (scoring of validity and usefulness of discovered resources) and prospectively (when designing new data capture systems). Proposed conceptual model intrinsically encompasses these obligatory items:

1. POPs data resources
   - institution (origin of data), time & space coordinates, type of resource (study design), examined entities (compounds), measures and methodical descriptors (experimental units, values & units, matrices, methods)
2. Cancer risk data resources
   - institution (origin of data), time & space coordinates, type of resource (study design, examined effects), examined entities (tumors, cancer diseases), measures and methodical descriptors (experimental units, values & units, cohorts, methods)

The system allows any type of reasonable extension; additional properties may be added on demand. However, minimized data standard ensures accessibility of key information namely in Measurement level of the model; i.e. when and where measurements were recorded, who recorded each measurement, the methodology of measurements, study design and aim. In this way the model can improve data visibility to search engines and enables greater levels of automation of common data transformation, summarization and integration.

Proposed conceptual model also contributes to widely recommended discovery of data based on the concepts they really represent [10,17,18]. In contrast to formal frameworks usually published with focus on one discipline [11,19] our model presumes search which exploits relationships between classes within environmental and human data sets as well as interdisciplinary relationships between the two areas. The concept supports development and formalization of ontologies, relevant for both environmental sciences and cancer epidemiology.

Ontology should represent the knowledge in a domain of interest, defined via the terminology (concepts, nomenclature) used within the domain and the properties and relationships among domain objects [20]. This concept is fully implemented in the model proposed here; the nomenclature baseline is extended by selected descriptors with defined dependencies. It is a formal framework for observational studies where we adopted structured approach recognizing key entities (nomenclature classes) in the 1st level and their characteristics (classifiers) important for the cancer risk studies.

Second level consists of measurements and their characteristics, i.e. validity criteria, origin of data, etc.. Third level covers content identification, namely values and units, scales.

6 Impact of proposed data model on data processing and analysis

Population studies focused on cancer risk are complex and require processing of highly diverse data. Even if we can get adequate data sources accessible for analyses, it is often difficult to select the best approach how to mutually relate measured factors; mostly our later analytic steps assume some specific input or data aggregation from the preceding measurements. That is why the data structure must be well defined but at the same time, flexible enough to reflect a wide range of possible hypotheses. Regarding heterogeneity of environmental problems, no unique, definitely the best model can be recommended. Of course, such system cannot be constructed retrospectively, on demand of running analyses. Baseline standardization proposed here in conceptual data model positively impacts upon analytical procedures, namely in the following three fields.

- **Hierarchical structure advances the data analysis.** The proposed conceptual model intrinsically distinguishes hierarchy of levels and descriptors which facilitates implementation of tools focused on data analysis and knowledge mining. The position of nomenclature entities and measurements can be used to denote a wide range of entity characteristics (nominal or ordinal measures of existence, prevalence). Using the hierarchy of descriptors we can easily decide whether the data are useful for a particular analysis.
- **Conceptual model supports robust reference comparison of values.** Regarding data analysis, very important attribute of the proposed model is incorporation of measurement level and its characteristics. Validity criteria reflect some precision measures as well as reference values or protocol standards. A measured value cannot be interpreted and analyzed without reference to a defined measurement standard.
- **Stratified analyses and integration of different data sources.** Hierarchical relationships among nomenclature classes and descriptors also potentiate development of automated SW tools for comparison of values using different strata. For example we can summarize prevalence of some cancer according to site locations because the sites and their population provide context for observation of cancer load. Similarly, the sites and matrices provide a context for measurement of POPs exposure. Both summaries can be then interlinked using various time frames. The concept thus facilitates evidence-based data integration, reasoned by compatibility of interlinked values.

7 Examples of practical implementation

Proposed model has already been used and implemented in SW toolkit focused on data discovery over Czech National Cancer Registry (system SVOD[5], [21]) and on processing of data from various POPs monitoring networks (system GENASIS[6]). Both information systems distinguish object entities (nomenclature items) and enable users to stratify accessible measures (content of resources) across a set o classifiers and methodically important attributes.

8 Conclusions and future challenges

In this study we proposed interdisciplinary conceptual model as a support of cancer-related environmental risk studies. The model can be useful in basic characterization of data standards and context of observations, as well as for information search. The model intrinsically defines dependency of obligatory descriptors and hierarchy in nomenclature attributes and supports establishment of data repositories with respect to other meaningful dimensions like cancer-related properties of chemical compounds, origin of data, coding of extreme or unusual values, etc. Such repositories allow the scientists to work with functional properties of nomenclature entities and related measurements. Moreover, measurements are linked to internal and external reference benchmarks which subsequently facilitate data integration or summaries.

Of course, many barriers that limit interdisciplinary data discovery still remain. The problems refer intrinsically to the information reachable in observation studies and cannot be easily solved by informatics. In cancer risk assessment, it is hard to collect representative data in relatively short period of time. Timescales here are long and the ability to switch a system to more complex level is limited by cost and organizational constraints. It is mostly not possible to carry out adequate assessment of the large scale systems with techniques that have been successful in smaller systems with limited heterogeneity.

Therefore, completely new methodical and experimental approaches are needed, especially those introducing novel, more sensitive and specific indicators. Population monitoring using methods of molecular epidemiology combined with reliable data on exposure offers such new powerful approach to determine the effect of genotoxic agents on human populations [16]. Study on the genetic polymorphism that can be a risk indicator for cancer development is a newly occurring stream in environmental sciences [22]. This methodical progress is making possible the collection and organization of biological informatics at an unprecedented level of detail and in extremely large quantities.

Acknowledgements. This research received financial support from the CETOCOEN project of the European Structural Funds (CZ.1.05/2.1.00/01.0001) and project FP7

[5] http://www.svod.cz
[6] http://www.genasis.cz

No. 247893 TaToo – Tagging Tool based on a Semantic Discovery Framework) granted by European Commission.

References

1. Michener, W.K., Brunt, J.W. : Ecological Data: Design, Management and Processing. Blackwell Science, Oxford (2000)
2. Jones, M.B., Schildhauer, M., Reichman, O.J., Bowers, S.: The New Bioinformatics: integrating ecological data from the gene to the biosphere. Ann. Rev. Ecol. Evol. Syst. 37, 519–544 (2006)
3. Elmagarmid, A., Rusinkiewicz,M., Sheth, A.: Management of Heterogeneous and Autonomous Database Systems, vol. 4. Morgan Kaufmann, San Francisco (1999)
4. Grossman, D.A., Frieder, O.: Information Retrieval: Algorithms and Heuristics. Springer, Heidelberg (2004)
5. Darwin Core: Darwin Core Schema (version 1.3), a draft standard of the Taxonomic Database Working Group (TDWG). http://wiki.tdwg.org/DarwinCore
6. DCMI. DCMI Metadata Terms. http://www.dublincore.org/ documents/dcmi-terms.
7. Athanasiadis, I. N., Mitkas, P. A.: A methodology for developing environmental information systems with software agents. In: Cortés U., Poch, M. (eds.) Whitestein Series in Software Agent Technologies and Autonomic Computing: Advanced Agent-Based Environmental Management Systems. Springer, Heidelberg, pp. 119–137 (2009)
8. Huang, P. S., Shih, L. H.: Effective environmental management through environmental knowledge management. Int. J. Environ. Sci. Tech. 6, 35-50 (2009)
9. Brochhausen, M., Spear, A.D., Cocos, C., et al.: The ACGT Master Ontology and its applications – Towards an ontology-driven cancer research and management system. J. Biomed. Inform. 44, 8–25 (2011)
10. Madin, J., Bowers S., Schildhauera M., et al.: An ontology for describing and synthesizing ecological observation data. Int. J. Ecol. Informatics 2, 279–296 (2007)
11. Williams, R.J., Martinez, N.D., Golbeck, J.: Ontologies for ecoinformatics. J. Web Semant. 4, 237–242 (2006)
12. UNEP Report of the First Expert Meeting to update the Guidance on the Global Monitoring Plan for Persistent Organic Pollutants (2010) http://chm.pops.int/Programmes/GlobalMonitoringPlan/Meetings/GMP1stExpertMeeting2010/tabid/760/ctl/Download/mid/3261/language/en-US/Default.aspx
13. Sotiriou, C., Pickard, M.J.: Taking gene-expression profiling to the clinic: when will molecular signatures become relevant to patient care? Nat. Rev. 7,545–53 (2007)
14. Tsiknakis, M., Brochhausen, M., Nabrzyski, J., et al.: A semantic grid infrastructure enabling integrated access and analysis of multilevel biomedical data in support of post-genomic clinical trials on Cancer. IEEE Trans. Inform. Technol. Biomed., Special issue on Bio-Grids 12, 191–204 (2008)
15. Sram, R.J.: Future research directions to characterize environmental mutagens in highly polluted area. Environ. Health Perspect. 104 (Suppl 3), 603–607 (1996)
16. Kyrtopoulos, S.A., Georgiadis, P., Autrup, H., et al.: Biomarkers of genotoxicity of urban air pollution. Overview and descriptive data from a molecular epidemiology study on populations exposed to moderate-to-low levels of polycyclic aromatic hydrocarbons: the AULIS project. Mutat. Res. 496, 207–228 (2001)
17. Berkley, C., Jones, M.B., Bojilova, J., Higgins, D.: Metacat: a schema-independent XML database system. Proc. of the 13th Intl. Conf. on Scientific and Statistical Database Management. IEEE Computer Society (2001)

18. Borgida, A.: Description logics in data management. IEEE Trans. Knowl. Data Eng. 7, 671–682 (1995)
19. Bard, J.R.L., Rhee, S.Y.: Ontologies in biology: design, applications and future challenges. Nat. Rev. Genet. 5, 213–222 (2004)
20. Baader, F., Calvanese, D.,McGuinness,D., Nardi, D., Patel-Schneider, P.: The Description Logic Handbook: Theory, Implementation, and Applications. Cambridge University Press (2003)
21. Dušek, L., Mužík, J., Kubásek, M., Koptíková, J., Žaloudík, J., Vyzula, R. Epidemiology of malignant tumours in the Czech Republic [online], http://www.svod.cz
22. Knudsen, L.E., Loft, S.H., Autrup, H. Risk assessment: the importance of genetic polymorphisms in man. Mutat. Res. 482, 83–88 (2001)

Appendix

Table 1. Conceptual model proposed for environmental cancer-related data management

1a. Resources of persistent organic pollutants (POPs)

CONCEPTUAL MODEL - LEVELS	DEFINITION & COMMENT
RESOURCE IDENTIFICATORS	
	Obligatory descriptors identifying institution (project) which guarantees the data (mostly also as owner of the resource content). In already closed resources, the identification is supplied with overall time/ space description.
OBJECTS – KEY ENTITIES	
NOMENCLATURE	Internationally used nomenclature of POP compounds (UNEP, 2010) – see Table 2. System allows selection of individual compounds and their groups.
OBJECT CLASSIFIERS	Categorized classifiers derived from external (encyclopedic) sources of information. Classifiers define groups but can be used also for scoring of individuals. Classifying criterion is linked to the individual compounds and/or to their groups.
Carcinogenicity	Attribute extending recognition of nomenclature classes, coded as no/yes/suspected. Code is directly interlinked with individual compounds. There are several international database sources of this information (IARC/US EPA/ACGIH) – see table 1.
Reference concentration values	Internal (time series analysis, background values) and/or external reference benchmarks. The classifier is coupled with given entity (compound), typically with direct link to matrix sampled and method used.
OBSERVATION – MEASUREMENT (OM)	
TIME & SPACE COORDINATES	Obligatory attributes, also proposed as inescapable items of any data standard.
STUDY TYPE	Study type (list): Long-term environmental monitoring / Short-

(design)	term environmental monitoring / Case studies / Screening.
PROBLEM STUDIED (exposure)	Problem studied (list): Accident, short-term exposure / Long-term exposure / Random inspection (survey of some area) / Examination of background (reference) site
METHODICAL ATTRIBUTES Measured entities Matrix Experimental unit Sampling methods Analytical methods	Obligatory identification of observation – measurement, necessary for interpretation of measured values. Measured entities select nomenclature items which are examined in given environmental matrix (soil, sediment, water, air, biota). Experimental unit identifies context of measured values (micro-samples within site, site – single sample, site – mixed sample, sample mixed across sites). Sampling and analytical methods fulfill minimized list of items which follows standardized norms and guidelines.
CONTENT	
Measures Units Precision measures	Content of the resource, in case of POPs mostly concentration levels in internationally standardized unit scales. Precision measures include sample variability (in concentration units) or detection limits of performed analytical methods.

1b. Resources of cancer epidemiology and risk

CONCEPTUAL MODEL - LEVELS	**DEFINITION & COMMENT**
RESOURCE IDENTIFICATORS	
	Obligatory descriptors identifying institution (project) which guarantees the data (mostly also as owner of the resource content). In already closed resources, the identification is supplied with overall time/ space /population description.
OBJECTS – KEY ENTITIES	
NOMENCLATURE	Internationally guaranteed system of classification of diseases – see Table 3. "Cancer" is used in many synonymous terms: tumor, neoplasm, metastasis.
OBJECT CLASSIFIERS	Categorized classifiers derived from external (encyclopedic) sources of information. Classifying criterion is linked to the individual cancer diagnosis and extent its information value
Nomenclature subsystems (TNM classification)	Internationally standardized nomenclature of cancer diseases, based on ICD-O-3 as key system and ICD-10 as multi-component subsystem for identification of malignant neoplasms.
Tumor type	Classifier important for risk studies focused on some type of harmful exposure (list): malignant / benign / unknown behavior
Reference values	Internal (time series analysis, background values) and/or external reference benchmarks (internationally reported values; reference epidemiological characteristics). The classifier is coupled with given entity (cancer type) and epidemiological measure, typically with relation to type of population observed.
OBSERVATION – MEASUREMENT (OM)	
TIME & SPACE	Obligatory attributes, also proposed as inescapable items of any

COORDINATES	data standard.
STUDY TYPE (design) PROBLEM STUDIED (exposure)	Study type (list): National epidemiological registry / Local (regional) registry / Hospital-based project / Cancer screening / Clinical trial / Cohort study / Case-control study / Descriptive epidemiologic observation Problem studied (list): Genetic factors, hereditary syndromes / Life style factors / Demography, ageing, gender studies / Occupational factors / Environmental factors / toxic exposures
METHODICAL ATTRIBUTES Measured entities Matrix Experimental unit Sampling/measurement methods	Obligatory identification of observation – measurement, necessary for interpretation of measured values. Measured entities select nomenclature items which are examined in given population. Experimental unit identifies context of measured values (representative population, selected cohort). Sampling/measurement methods fulfill minimized list of items which follows standardized guidelines for epidemiologic observation studies.
CONTENT	
Measures Units Precision measures	Content of the resource, typically recognized epidemiological measure (incidence, mortality, prevalence) in internationally standardized unit scales (crude estimate, ASR, etc.). Precision measures include population representativeness (coverage) of the data resource.

Table 2. List of POPs from annexes A, B and C of the Stockholm convention and their congeners according to recommendation for monitoring from the first workshop that considered the 2nd revision of the Guidance document for the GMP, held 12-14 April 2010 in Geneva (UNEP 2010) .

CAS	ES	name	level	state	carcinogenity[1]
309-00-2	**206-215-8**	**aldrin**	**1**	**substance**	3/B2/A3
57-74-9	**200-349-0**	**chlordane**	**1**	**group**	2B/B2/A3
5103-71-9	225-825-5	*cis*-chlordan	2	substance	-/-/-
5103-74-2	225-826-0	*trans*-chlordan	2	substance	-/-/-
5103-73-1		*cis*-nonachlor	2	substance	-/-/-
39765-80-5		*trans*-nonachlor	2	substance	-/-/-
26880-48-8		oxychlordane	2	mixture	-/-/-
8017-34-3		**DDT**	**1**	**group**	2B/B2/A3
50-29-3	200-024-3	4,4'-DDT	2	substance	2B/B2/A3
789-02-6	212-332-5	2,4'-DDT	2	substance	-/-/-
72-55-9	200-784-6	4,4'-DDE	2	substance	2B/B2/-
3424-82-6	222-318-0	2,4'-DDE	2	substance	-/-/-
72-54-8	200-783-0	4,4'-DDD	2	substance	2B/B2/-
53-19-0	200-166-6	2,4'-DDD (mitotane)	2	substance	-/-/-
60-57-1	**200-484-5**	**dieldrin**	**1**	**substance**	3/B2/A4
72-20-8	**200-775-7**	**endrine**	**1**	**substance**	3/D/A4
118-74-1	**204-273-9**	**hexachlorbenzene (HCB)**	**1**	**substance**	2B/B2/A3
76-44-8	**200-962-3**	**heptachlor**	**1**	**substance**	2B/B2/A3
1024-57-3	213-831-0	heptachlor epoxide	2	substance	3/B2/A3
2385-85-5	**219-196-6**	**mirex**	**1**	**substance**	2B/-/-
1336-36-3	**215-648-1**	**polychlorinated biphenyls (PCB)**	**1**	**group**	2A/B2/
7012-37-5	230-293-2	2,4,4'-trichlorobiphenyl (PCB 28)	2	substance	-/-/-
35693-99-3		2,2',5,5'-tetrachlorobiphenyl (PCB 52)	2	substance	-/-/-
37680-73-2		2,2',4,5,5'-pentachlorobiphenyl (PCB 101)	2	substance	-/-/-
31508-00-6		2,3',4,4',5-pentachlorobiphenyl (PCB 118)	2	substance	-/-/-
35065-28-2		2,2',3,4,4',5'-hexachlorobiphenyl (PCB 138)	2	substance	-/-/-
35065-27-1		2,2',4,4',5,5'-hexachlorobiphenyl (PCB 153)	2	substance	-/-/-
35065-		2,2',3,4,4',5,5'-heptachlorobiphenyl	2	substance	-/-/-

29-3		(PCB 180)			
32598-13-3		3,3',4,4'-tetrachlorobiphenyl (PCB 77)	2	substance	-/-/-
70362-50-4		3,4,4',5-tetrachlorobiphenyl (PCB 81)	2	substance	-/-/-
32598-14-4		2,3,3',4,4'-pentachlorobiphenyl (PCB 105)	2	substance	3/-/-
74472-37-0		2,3,4,4',5-pentachlorobiphenyl (PCB 114)	2	substance	-/-/-
31508-00-6		2,3',4,4',5-pentachlorobiphenyl (PCB 118)	2	substance	-/-/-
65510-44-3		2,3',4,4',5'-pentachlorobiphenyl (PCB 123)	2	substance	-/-/-
57465-28-8		3,3',4,4',5-pentachlorobiphenyl (PCB 126)	2	substance	-/-/-
38380-08-4		2,3,3',4,4',5-hexachlorobiphenyl (PCB 156)	2	substance	-/-/-
69782-90-7		2,3,3',4,4',5'-hexachlorobiphenyl (PCB 157)	2	substance	-/-/-
52663-72-6		2,3',4,4',5,5'-hexachlorobiphenyl (PCB 167)	2	substance	-/-/-
32774-16-6		3,3',4,4',5,5'-hexachlorobiphenyl (PCB 169)	2	substance	-/-/-
39635-31-9		2,3,3',4,4',5,5'-heptachlorobiphenyl (PCB 189)	2	substance	-/-/-
		polychlorinated dibenzo-p-dioxins (PCDD)	**1**	**group**	-/-/-
1746-01-6	217-122-7	2,3,7,8-tetrachlorodibenzo[b,e][1,4]dioxin (2378 TCDD)	2	substance	1/-/-
40321-76-4		1,2,3,7,8-Pentachlorodibenzo-p-dioxin (12378-PeCDD)	2	substance	3/-/-
39227-28-6		1,2,3,4,7,8-Hexachlorodibenzo-p-dioxin (123478-HxCDD)	2	substance	3/-/-
57653-85-7		1,2,3,6,7,8-Hexachlorodibenzo-p-dioxin (123678-HxCDD)	2	substance	3/-/-
19408-74-3		1,2,3,7,8,9-Hexachlorodibenzo-p-dioxin (123789-HxCDD)	2	substance	3/B2/-
35822-46-9		1,2,3,4,6,7,8-Heptachlorodibenzo-p-dioxin (1234678-HpCDD)	2	substance	3/-/-
3268-87-9		Octachlorodibenzo-p-dioxin (OCDD)	2	substance	3/-/-
		polychlorinated dibenzofurans (PCDF)	**1**	**group**	-/-/-
51207-31-9		2,3,7,8-tetrachlorodibenzofuran (2378-TCDF)	2	substance	3/-/-
57117-41-6		1,2,3,7,8-pentachlorodibenzofuran (12378-PeCDF)	2	substance	3/-/-
57117-31-4		2,3,4,7,8-pentachlorodibenzofuran (23478-PeCDF)	2	substance	3/-/-
70648-26-9		1,2,3,4,7,8-hexachlorodibenzofuran (123478-HxCDF)	2	substance	3/-/-
57117-44-9		1,2,3,6,7,8-hexachlorodibenzofuran (123678-HxCDF)	2	substance	3/-/-
72918-21-9		1,2,3,7,8,9-hexachlorodibenzofuran (1,2,3,7,8,9-HxCDF)	2	substance	3/-/-
60851-34-5		2,3,4,6,7,8-hexachlorodibenzofuran (234678-HxCDF)	2	substance	3/-/-

67562-39-4		1,2,3,4,6,7,8-heptachlorodibenzofuran (1234678-HpCDF)	2	substance	3/-/-
55673-89-7		1,2,3,4,7,8,9-heptachlorodibenzofuran (1234789-HpCDF)	2	substance	3/-/-
39001-02-0		Octachlorodibenzofuran (OCDF)	2	substance	3/-/-
8001-35-2	**232-283-3**	**toxaphene**	**1**	**mixture**	2B/B2/A3
142534-71-2		2-endo,3-exo,5-endo,6-exo,8,8,10,10-octachlorobornan (P26)	2	substance	-/-/-
6680-80-8		2-endo,3-exo,5-endo,6-exo,8,8,9,10,10-nonachlorobornan (P50)	2	substance	-/-/-
154159-06-5		2,2,5,5,8,9,9,10,10-nonachlorobornan (P62)	2	substance	-/-/-
143-50-0	**205-601-3**	**chlordecone**	**1**	**substance**	2B/-/-
319-84-6	**206-270-8**	**α-hexachlorcyclohexane**	**1**	**substance**	-/B2/-
319-85-7	**206-271-3**	**β-hexachlorcyclohexane**	**1**	**substance**	-/C/-
58-89-9	**200-401-2**	**γ-hexachlorcyclohexane**	**1**	**substance**	-/-/A3
36355-01-8	**252-994-2**	**hexabromobiphenyl (HBB)**	**1**	**group**	-/-/-
59080-40-9		PBB153	2	substance	-/-/-
67888-98-6		PBB138	2	substance	-/-/-
608-93-5	**210-172-0**	**pentachlorbenzene (PeCBz)**	**1**	**substance**	-/D/-
40088-47-9		**tetrabromodiphenyl ethers (TBDE)**	**1**	**group**	-/-/-
147217-75-2		2,2',4 (BDE 17)	2	substance	-/-/-
41318-75-6		2,4,4' (BDE 28)	2	substance	-/-/-
5436-43-1		2,2',4,4'-tetrabromdiphenylether (BDE 47)	2	substance	-/-/-
32534-81-9		**pentabromodiphenyl ethers (PeBDE)**	**1**	**group**	-/-/-
60348-60-9		2,2',4,4',5-pentabromodiphenyl ether (BDE 99)	2	substance	-/-/-
189084-64-8		2,2',4,4',6-pentabromodiphenyl ether (BDE 100)	2	substance	-/-/-
36483-60-0		**hexabromodiphenyl ethers (HxBDE)**	**1**	**group**	-/-/-
68631-49-2		2,2',4,4',5,5'-hexabromodiphenyl ether (BDE 153)	2	substance	-/-/-
207122-15-4		2,2',4,4',5,6'-hexabromodiphenyl ether (BDE 154)	2	substance	-/-/-
68928-80-3		**heptabromodiphenyl ethers (HpBDE)**	**1**	**group**	-/-/-
446255-22-7		2,2',3,3',4,5',6-heptabromodiphenyl ether (BDE 175)	2	substance	-/-/-
207122-16-5		2,2',3,4,4',5',6-heptabromodiphenyl ether (BDE183)	2	substance	-/-/-
32536-52-0		**oktabromodiphenyl ether (OBDE)**	**1**	**group**	-/-/-
1763-23-1	**217-179-8**	**perfluorooctane sulfonic acid (PFOS) and derivates**	**1**	**group**	-/-/-

31506-32-8	N-methyl heptadecafluorooctane sulfonamide (NMeFOSA)	2	substance	-/-/-
4151-50-2	N-ethyl heptadecafluorooctane sulfonamide (NEtFOSA)	2	substance	-/-/-
24448-09-7	N-methyl heptadecafluorooctane sulfonamidoethanol (NMeFOSE)	2	substance	-/-/-
1691-99-2	N-ethyl heptadecafluorooctane sulfonamidoethanol (NEtFOSE)	2	substance	-/-/-

[1] Carcinogenity groups IARC/US EPA/ACGIH, - means, that the substance is not listed. Toxned database was used: http://toxnet.nlm.nih.gov/cgi-bin/sis/search.

IARC: International Agency for Research on Cancer (http://monographs.iarc.fr/index.php)

Group 1	Carcinogenic to humans (107 agents)
Group 2A	Probably carcinogenic to humans (59 agents)
Group 2B	Possibly carcinogenic to humans (266 agents)
Group 3	Not classifiable as to its carcinogenicity to humans (508 agents)
Group 4	Probably not carcinogenic to humans (1 agent)

US EPA: United States Environmental Protection Agency (http://www.epa.gov/iris/)

Group A	Human carcinogen
Group B1	Probable human carcinogen (limited evidence of carcinogenity from epidemiological studies)
Group B2	Probable human carcinogen (sufficient evidence of carcinogenicity in animals and others)
Group C	Possible human carcinogen
Group D	Not classifiable as to human carcinogenity
Group E	Evidence of non-carcinogenity for humans

ACGIH: American Conference of Governmental Industrial Hygienists (http://www.acgih.org/SiteSearch/index.cfm)

Group A1	Confirmed human carcinogen
Group A2	Suspected human carcinogen
Group A3	Confirmed animal carcinogen with unknown relevance to humans
Group A4	Not classifiable as a human carcinogen
Group A5	Not suspected as a human carcinogen

Table 3. International classification systems of cancer diagnoses[7]

International Classification of Diseases, 10th edition (ICD-10)
(C00–C14) Malignant neoplasms, lip, oral cavity and pharynx
(C15–C26) Malignant neoplasms, digestive organs
(C30–C39) Malignant neoplasms, respiratory system and intrathoracic organs
(C40–C41) Malignant neoplasms, bone and articular cartilage
(C43–C44) Malignant neoplasms, skin
(C45–C49) Malignant neoplasms, connective and soft tissue
(C50–C58) Malignant neoplasms, breast and female genital organs
(C60–C63) Malignant neoplasms, male genital organs
(C64–C68) Malignant neoplasms, urinary organs
(C69–C72) Malignant neoplasms, eye, brain and central nervous system
(C73–C75) Malignant neoplasms, endocrine glands and related structures
(C76–C80) Malignant neoplasms, secondary and ill-defined
(C81–C96) Malignant neoplasms, stated or presumed to be primary, of lymphoid, haematopoietic and related tissue
(C97) Malignant neoplasms of independent (primary) multiple sites
(D00–D09) In situ neoplasms
(D10–D36) Benign neoplasms
(D37–D48) Neoplasms of uncertain or unknown behavior
International Classification of Diseases for Oncology, 3rd Edition (ICD-O-3) - overview
Topography. The topography of the tumor is described by topographical code. Topographical code corresponds to the C section of ICD-10 (exceptions are listed in ICD-O-3).
Morphology. The morphology provides five-digit codes ranging from M-8000/0 to M-9989/3. The first four digits indicate the specific histological term. The fifth digit after the slash (/) is the code, which indicates whether a tumor is malignant, benign, in situ, or uncertain.
Grade. A separate one-digit code is also provided for histological grading (differentiation).
(8000–8009) Not otherwise specified
(8000–8004) Neoplasms, NOS
(8010–8790) Epithelial
(8010–8040) Epithelial neoplasms, NOS
(8050–8080) Squamous cell neoplasms
(8090–8110) Basal cell neoplasms
(8120–8130) Transitional cell Papillomas And Carcinomas
(8140–8380) Adenomas And Adenocarcinomas (glands)
(8390–8420) Adnexal And Skin appendage Neoplasms
(8430–8439) Mucoepidermoid Neoplasms
(8440–8490) Cystic, Mucinous And Serous Neoplasms
(8500–8540) Ductal, Lobular And Medullary Neoplasms
(8550–8559) Acinar cell neoplasms
(8560–8580) Complex epithelial neoplasms
(8590–8670) Specialized gonadal neoplasms
(8680–8710) Paragangliomas And Glomus tumors

[7] WHO, International Classification of Diseases (ICD); http://www.who.int/classifications;
U.S. National Institute of Health, National Cancer Institute, http://www.seer.cancer.gov/iccc

(8720–8790) Nevi And Melanomas

(8800–9370) Connective tissue

(8800–8809) Soft tissue Tumors And Sarcomas, Nos
(8810–8830) Fibromatous neoplasms
(8840–8849) Myxomatous neoplasms
(8850–8880) Lipomatous neoplasms
(8890–8920) Myomatous neoplasms
(8930–8990) Complex Mixed And Stromal Neoplasms
(9000–9030) Fibroepithelial Neoplasms
(9040–9049) Synovial-Like Neoplasms
(9050–9059) Mesothelial Neoplasms
(9060–9090) Germ cell Neoplasms
(9100–9109) Trophoblastic neoplasms
(9110–9119) Mesonephromas
(9120–9160) Blood vessel tumors
(9170–9179) Lymphatic vessel tumors
(9180–9240) Osseous And Chondromatous neoplasms
(9250–9259) Giant cell tumors
(9260–9269) Miscellaneous bone tumors
(9270–9340) Odontogenic tumors
(9350–9370) Miscellaneous tumors

(9380–9589) Nervous system

(9380–9480) Gliomas
(9421/3) Pilocytic astrocytoma
(9440/3) Glioblastoma multiforme
(9490–9520) Neuroepitheliomatous neoplasms
(9530–9539) Meningiomas
(9540–9570) Nerve sheath tumors
(9580–9589) Granular cell tumors and Alveolar soft part sarcoma

(9590–9999) Hematologic (Leukemias, Lymphomas and related disorders)

ICD-O-3 classification of Hematologic malignances according to WHO Classification of Tumors of Haematopoietic and Lymphoid Tissues (ICD-10 diagnoses C81–C96) - simplified

Lymphomas and related disorders

(9590–9599) Malignant lymphoma, NOS, Or diffuse
(9650–9660) Hodgkin's disease
(9670–9680) Malignant lymphoma Specified Type, Diffuse Or Nos
(9690–9699) Malignant lymphoma, Follicular Or Nodular, With Or Without diffuse areas
(9700–9709) Specified Cutaneous And Peripheral T-Cell Lymphomas
(9710–9719) Other Specified Non-Hodgkin's lymphomas
(9720–9729) Other Lymphoreticular neoplasms
(9730–9739) Plasma cell tumors
(9740–9749) Mast cell Tumors
(9760–9769) Immunoproliferative diseases

Lymphoid leukemias, and related conditions

(9800–9809) Leukemias, NOS

(9820–9829) Lymphoid leukemias

(9830–9839) Plasma cell leukemia

Myeloid leukemias, and related conditions

(9840–9849) Erythroleukemias (FAB-M6)

(9850–9859) Lymphosarcoma cell leukemia

(9860–9869) Myeloid (Granulocytic) Leukemias

(9870–9889) Basophilic leukemia and Eosinophilic leukemia

(9890–9899) Monocytic leukemias

(9900–9948) Other Leukemias

Other

(9950–9970) Miscellaneous Myeloproliferative And Lymphoproliferative disorders

(9980–9989) Myelodysplastic syndrome

International Classification of Childhood Cancer (ICCC) based on International Classification of Diseases for Oncology, 3rd Edition (ICD-O-3)

I Leukemias, myeloproliferative diseases, and myelodysplastic diseases

[011] (a) Lymphoid leukemias

[012] (b) Acute myeloid leukemias

[013] (c) Chronic myeloproliferative diseases

[014] (d) Myelodysplastic syndrome and other myeloproliferative diseases

[015] (e) Unspecified and other specified leukemias

II Lymphomas and reticuloendothelial neoplazma

[021] (a) Hodgkin lymphomas

[022] (b) Non-Hodgkin lymphomas (except Burkitt lymphoma)

[023] (c) Burkitt lymphoma

[024] (d) Miscellaneous lymphoreticular neoplasms

[025] (e) Unspecified lymphomas

III CNS and miscellaneous intracranial and intraspinal neoplazma

[031] (a) Ependymomas and choroid plexus tumor

[032] (b) Astrocytomas

[033] (c) Intracranial and intraspinal embryonal tumors

[034] (d) Other gliomas

[035] (e) Other specified intracranial and intraspinal neoplasms

[036] (f) Unspecified intracranial and intraspinal neoplasms

IV Neuroblastoma and other peripheral nervous cell tumors

[041] (a) Neuroblastoma and ganglioneuroblastoma

[042] (b) Other peripheral nervous cell tumors

[050] V Retinoblastoma

VI Renal tumors

[061] (a) Nephroblastoma and other nonepithelial renal tumors

[062] (b) Renal carcinomas

[063] (c) Unspecified malignant renal tumors

VII Hepatic tumors

[071] (a) Hepatoblastoma

[072] (b) Hepatic carcinomas

[073] (c) Unspecified malignant hepatic tumors

VIII Malignant bone tumors

[081] (a) Osteosarcomas

[082] (b) Chondrosarcomas

[083] (c) Ewing tumor and related sarcomas of bone

[084] (d) Other specified malignant bone tumors

[085] (e) Unspecified malignant bone tumors

IX Soft tissue and other extraosseous sarcomas

[091] (a)Rhabdomyosarcomas

[092] (b) Fibrosarcomas, peripheral nerve sheath tumors, and other fibrous neoplasms

[093] (c) Kaposi sarcoma

[094] (d) Other specified soft tissue sarcomas

[095] (e) Unspecified soft tissue sarcomas

X Germ cell tumors, trophoblastic tumors, and neoplasms of gonads

[101] (a) Intracranial and intraspinal germ cell tumors

[102] (b) Malignant extracranial and extragonadal germ cell tumors

[103] (c) Malignant gonadal germ cell tumors

[104] (d) Gonadal carcinomas

[105] (e) Other and unspecified malignant gonadal tumors

XI Other malignant epithelial neoplasms and malignant melanomas

[111] (a) Adrenocortical carcinomas

[112] (b) Thyroid carcinomas

[113] (c) Nasopharyngeal carcinomas

[114] (d) Malignant melanomas

[115] (e) Skin carcinomas

[116] (f) Other and unspecified carcinomas

XII Other and unspecified malignant neoplasms

[121] (a) Other specified malignant tumors

[122] (b) Other unspecified malignant tumors

[999] Not Classified by ICCC or in situ