Modern Information Technologies in Environmental Health Surveillance: an Overview and Analysis

Yvan Bédard, PhD [1]

William D. Henriques, PhD, MSPH [2]

[1] Centre for Research in Geomatics, Université Laval, Quebec City, Canada G1K7P4, yvan.bedard@scg.ulaval.ca

[2] GIS Coordinator, Agency for Toxic Substances and Disease Registry, Division of Health Assessment and Consultation, 1600 Clifton Road NE, MS E56, Atlanta GA 30333, whenriques@cdc.gov

Introduction

Recently, the world has seen the introduction of numerous new information technologies (IT) that are having significant impacts on society. Many authors speak of a communication revolution as important as those following the invention of the printing press, radio and television. As a result of this revolution, new expressions such as 'Global Village', 'Information Society', and 'Digital Earth' have been introduced to describe this new connectedness we experience. These new technologies are being implemented extensively in developed countries and in isolated pockets of developing nations and are having deep impacts on things such as environmental health surveillance. The aim of the present paper is to give an overview of technologies that are having or will likely have the most important impacts in environmental health research and practice.

Overview Of Modern Information Technology For Environmental Health Surveillance

World Wide Web

The massive penetration of the internet and the **World Wide Web** (W3) in today's society has created new opportunities to provide/access information and services. For the first time, a massive amount of information and services are available immediately worldwide 24 hours a day. The most important opportunities offered today include:

a) **E-mail** (electronic mail): The most widely used application on the internet, used to send/receive messages, documents and other files from any location.

b) **Web sites:** Where an organization or individual gives access to "web surfers" to static and limited information or massive and dynamic information. The best user interfaces are designed for easy navigation using hyperlinks (clickable links to other parts of the site or

other web sites) and other formats. When an organization restricts a part of its web site to internal employees, it becomes an **intranet,** and when they add access to selected external clients or partners, it becomes an **extranet**.

c) **Portals:** Web sites that offer a large array of well organized and indexed information, search engines and customizable services such as weather forecasts, user-selected sport news, etc. (e.g. Yahoo). When a portal focuses on a precise field of information (e.g. environmental health), it is called a vertical portal, or **vortal**.

d) **E-commerce:** Some web sites offer electronic commerce allowing an organization, such as a retail store, to sell products directly via the web. Among these sites, one finds **digital libraries** that provide users with searchable and downloadable catalogs of digital documents (e.g. reports, datasets, maps and satellite images).

e) The web also offers technologies for **distance learning** and **workgroups**. These usually involve static or interactive multi-point communication capabilities such as textual communication, group chatting, whiteboarding, group calendaring, etc. Specialized software is required for the host organization only (except for a facultative on-line web camera).

Additionally, there are a few general-purpose web sites dedicated to facilitate searching on the web and thousands of specialized **search sites**. **Metasearch sites** are sites that make simultaneous use of several **search sites** and present the results to the user in an organized format. To benefit from the internet and W3 technologies, a user needs access to an ISP (Internet Service Provider), an electronic address and an internet connection. To offer such services to others one must add a web server and specialized software (firewall) to these technical requirements.

DBMS and Universal Servers:

DBMS (Data Base Management Systems) include tools such as Oracle DBMS, SQL-Server, Informix, Sybase, DB2, Access, etc. This family of tools is 30 years old and has attained a high level of commercial maturity, especially with the market lead of the relational approach developed over the last 20 years. **Relational DBMS** allows one to define a database structure, feed it with simple data (a string of characters, numbers, dates or boolean values), verify its integrity, manipulate the data, query them and build automatic reports (Date 2000). They can be

accessed simultaneously by several, or even thousands, of users without crashing or corrupting the data. These data can be stored in a unique site or distributed over several sites. Access to the data can be direct, via application-built graphical user interfaces or via the web.

With the demand-driven influence of the media-rich web as well as the push-driven influence of the multimedia-capable object-oriented DBMS appearing in the 1990s, there has been an evolution of Relational DBMS into hybrid Object-Relational DBMS, or **Universal Servers** (e.g. Oracle 8i and Cartridges, Informix with Datablades). These are called "universal" because they are not restricted to the traditional types of data found in DBMS and also have the added capability of storing, manipulating and querying multimedia information. Thus, today's DBMS are well adapted to the web revolution.

Data Warehouses and the latest Decision Support Tools:

While DBMS were created to bring coherence among previously disparate, independent, redundant and application-specific data files (Figure 1), most organizations have implemented databases in a way that has created isolated database islands. There has been an evolution from having independent and redundant files to independent and overlapping databases. This is considered an improvement though as the overlap and coherence problems have become easier to manage. Nevertheless, the situation still lacks the unified view of a system where data coming from different databases are organized and ready to rapidly provide strategic, synthesized and aggregated information for high-level decision making (Inmon et al 1996). **Data warehouses** which "provide a unified view of dispersed heterogeneous databases in order to efficiently feed the decision-support tools used for strategic decision making" have been designed to address this issue (Bédard, Merrett and Han 2000). To achieve this, the warehouse must import, in read-only and batch modes, subsets of the source database (called legacy systems) and process/integrate them so that the resulting information to be stored in the warehouse is consistent and properly aggregated (Poe 1995).

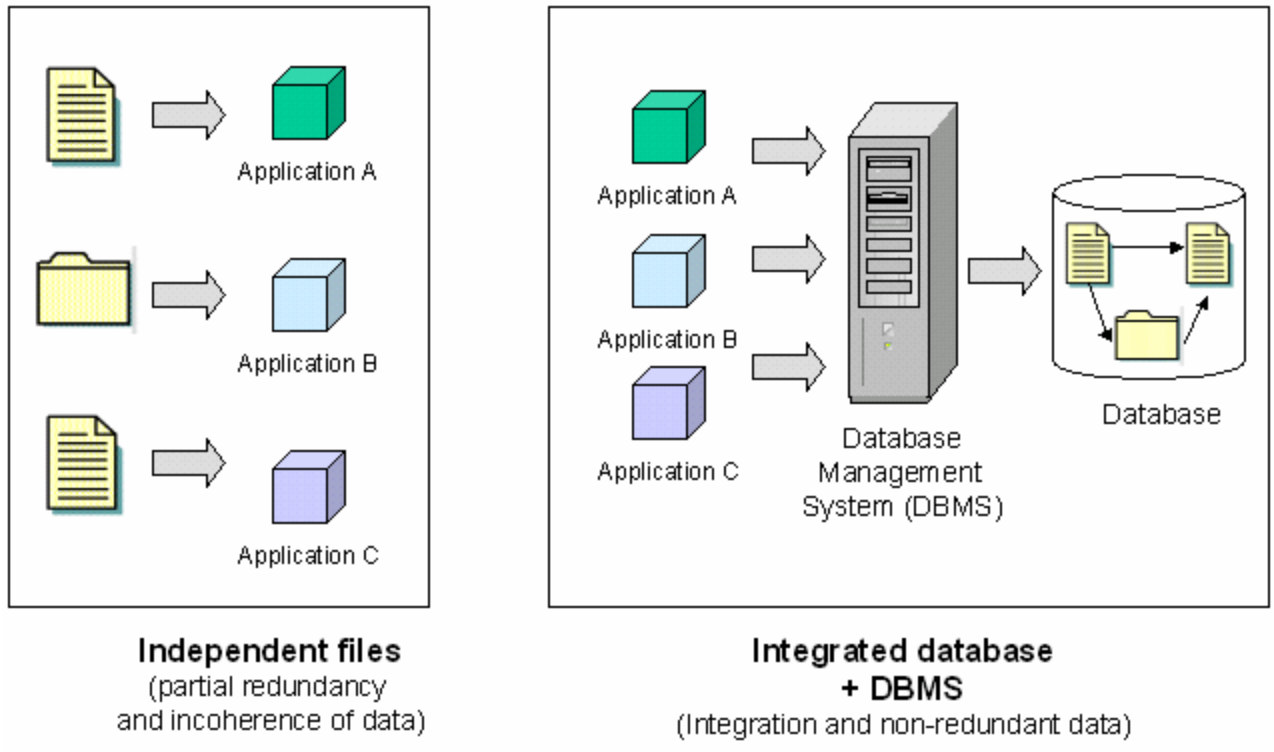

Figure 1: independent files vs integrated database + DBMS

When data are updated in the legacy systems, the new information is added to the warehouse without replacing previous data allowing them to support the analysis of trends over time, a key element needed for decision making (Brackett 1996). Consequently, data warehouses are considered the main source of information for knowledge discovery and business intelligence.

Data warehouses differ from traditional databases in that they are designed to support small volumes of long aggregation-oriented strategic-level transactions involving large volumes of data. To achieve these opposing objectives that cannot be met in a single database when the volume of data becomes large, two different database designs (and sometimes technologies) are used: the object-relational design of traditional DBMS for transaction-oriented operations and the multidimensional design of data warehouses for analysis-oriented operations and knowledge discovery (i.e. decision-support; e.g. Red Brick, Essbase and Oracle Express).

When an organization-wide data warehouse is not needed, one may use the same technology to build a focused, specialized, domain-specific mini-warehouse extracting data from a subset of legacy systems to develop more summarized information in a mini-warehouse called a **datamart**. We regularly find several datamarts in an organization built on top of unique enterprise-wide data warehouses to avoid to avoid another level of isolated information island.

In such systems, the datamarts and data warehouse are designed so that the datamart offers subject-oriented, more highly aggregated information for a specialized view with faster data access than in the warehouse approach.

In order to support the extraction of useful knowledge from the data warehouse/datamart, one needs a decision-support tool such as Query and Report builders, On-Line Analytical Processing software (OLAP) and Data Mining tools.

a) **Query and report builders**: these tools (e.g. Impromptu, Crystal Report) facilitate the creation of queries and reports by replacing the standard technically-driven SQL interface (Structured Query Language) by a more intuitive user-interface, usually based on natural language (e.g. plain English) or better query/report-driven graphical interfaces.

b) **OLAP**: the most popular category of decision-support tools providing unique capabilities to explore massive amounts of data in a rapid, intuitive and interactive way. Such ad hoc discovery-driven exploration of data relies on the multidimensional nature of the warehouse data structure (called data cube) where the user can go directly from detailed levels of information to more aggregated/summarized levels of information (drilling down and rolling up) (*cf. Figure 2A)* as well as navigating from one category of information to another category (*cf.Figure 2B*) correlating, filtering, slicing them, etc. (e.g.'s are Powerplay and Bussiness Objects).

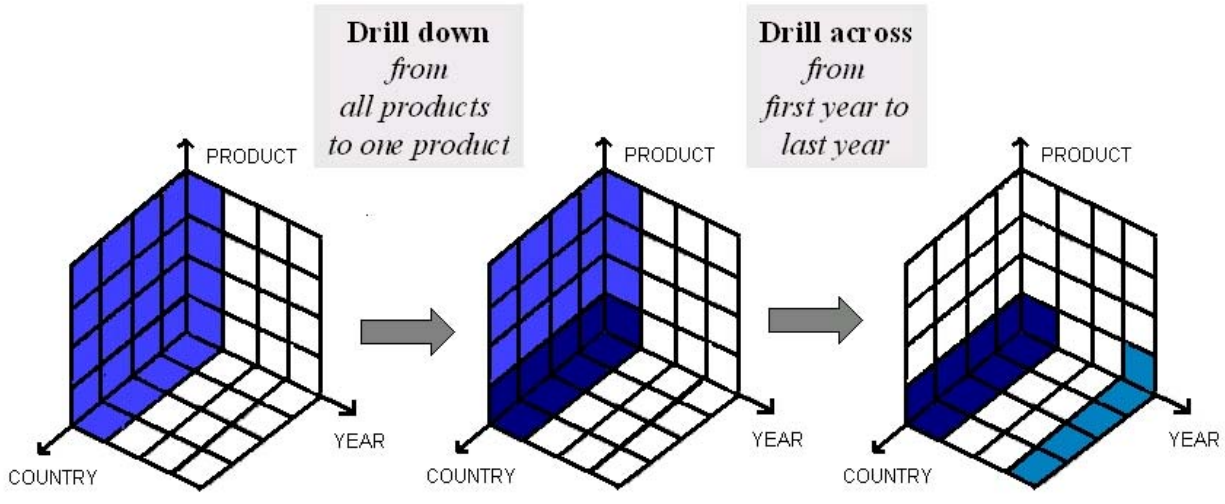

Figure 2: OLAP navigation

c) **Data Mining**: this category of knowledge discovery packages aims at automating the search for hidden patterns, correlations or trends in larges data cubes and to automatically make predictions based on historical data. The "automatic" nature of the exploration methods lead to the discovery of unexpected and complex patterns and accelerate the exploration of larges warehouses. To achieve this they use complex techniques such as neural networks, decision trees, genetic algorithms, rule induction and nearest neighbor calculations.

It is becoming more common to find these technologies embedded in statistical packages and DBMS thus improving their decision-support qualities. Although several decision-support products are third-party add-ons to DBMS or to specialized warehouse/marts engines, today's trend is to find the multidimensional capabilities as well as the decision-support front-ends built into major universal servers. Finally, it is possible to make data warehouses and datamarts accessible through the web with browser-based query and report builders as well as EIS (Executive Information Systems, i.e. dashboard-like read-only reactive reporting tool).

Geographic Information Systems (GIS) and related technologies:

Much of the information that organizations maintain includes geographic elements such as a street address, postal code, county, province/state, or map location specified by geographic coordinates. In the early 1980s, digital mapping converged with database management systems (DBMS) giving rise to the first commercial **Geographic Information Systems** (GIS) such as ArcInfo and Intergraph MGE. This has allowed organizations to relate tabular information to locations on digital maps and produce thematic maps (Figure 3). Rapidly, spatial analysis functions have been added and from the mid-80s to the mid-90s capabilities such as buffering of spatial data layers to result in demographic profiles within a distance from a feature of interest (Heitgerd, 1994), spatial intersection to identify areas suitable for the proliferation of a disease vector (Byron Wood, JPL CA), and network analysis (e.g. shortest path, routing, defining the zones covered within a given timeframe of traveling) were developed. Coupled with location gathering technologies such as global positioning systems (GPS), computer mapping and spatial analysis will be the next revolution in computational technologies affecting the way in which we examine data for environmental and heath surveillance.

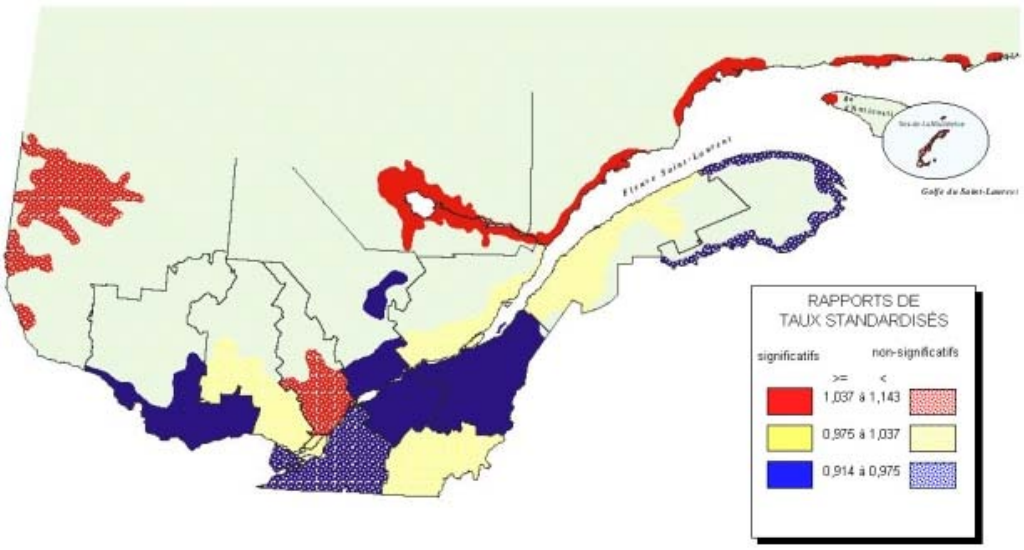

Figure 3: example of a screen copy displaying a thematic map (cancer for the province of Quebec)

The mass market penetration among larger organizations and upper-end products, in relation to the capability of **universal servers to manage geographic data** and to display maps with **low-cost viewers** in a client-server architecture is having a major impact on the market (e.g. ArcView, Geomedia and MapInfo).

A similar phenomenon is taking place with **geographic web servers** offering web-mapping capabilities (Figure 4). Servers (e.g. Geomedia WebMap, MapXtreme, MapGuide, etc.) are used for various types of applications (address locating, distance learning, trip planning, etc.) and to enrich traditional web technologies. Two examples of the latter are (1) **geographic digital libraries** which allow one to access, obtain and download digital maps, aerial photographs and satellite images from a government or e-commerce web site, and (2) **location-commerce** (or l-commerce) which provides custom maps showing locations of user-requested services and directional information.

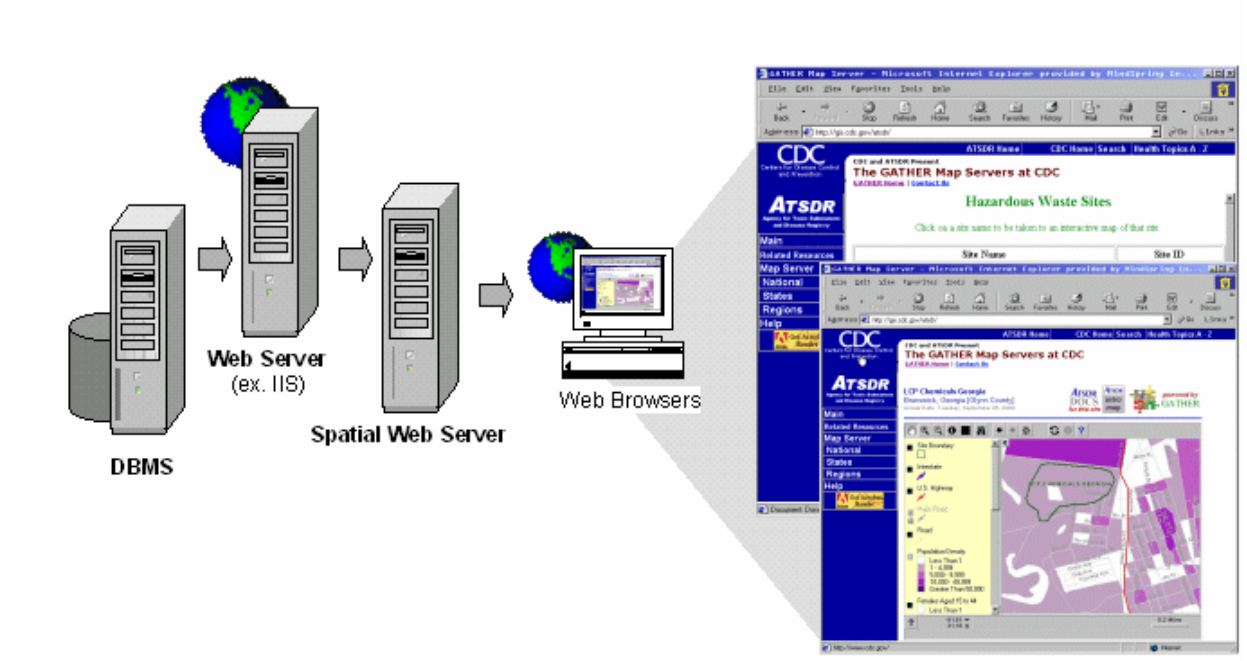

Figure 4: Example of a spatial web server, the GATHER map server at CDC

Finally, the GIS community is moving towards internationally accepted open-standards (cf. ISO and OpenGIS Consortium), interoperability solutions (e.g. OGDI, the Open Geospatial Data Store) and the development of very efficient geographic data fusion tools (e.g. FME, the Feature Manipulation Engine from Safe Software) allowing one to integrate/process geographic data from diverse sources. As a result of these advances, the first projects of spatial data warehouses, spatial OLAP and spatial data mining are moving out of research labs and into the applications market. The GIS community and its technologies have become part of the information technology mainstream.

Overview Of Environmental Health Surveillance Needs Regarding InformationTechnology

A drastic improvement in computer technology has occurred over the past 20 years. More information is available to more people eager to learn of the impacts of environmental pressures on public health. Databases allow access to annual summaries of such things as chemical emmissions information compiled through the US Environmental Protection Agency's Toxic Release Inventory.

While this information is useful, the reporting of sheer volumes of chemicals is not enough. A more suitable measure of the impact of chemical volumes is needed utilizing knowledge of compound toxicity. An interface that links time- and volume-based information on chemical

emissions to an algorithm that considers the weighted risk to adverse health effects of each substance over a geographic area is a tool that is technologically feasible, but requires the convergence of disciplines. The technological challenge is to develop tools that convert extensive databases containing existing environmental data in relational database systems (RDBMS) into maps that depict risk of adverse health impacts to people in a specific geographic area (e.g. Figure 5). Linking toxicity information with concentration data, we can develop maps that more closely depict geographic areas of potential environmental health concern. While this simplistic example uses an approach of querying source information (ATSDR's HazDat database) to map only those sites that fulfill specific criteria, it fails to consider additional sources of human exposure to environmental contamination. As additional information is added to gain a more comprehensive view of the environmental impact of chemicals, the picture becomes more complex. Reporting volumes of chemicals dispersed into the environment is useful, but tools are still needed to summarize and categorize (e.g. heavy metals) this information over a geographic area and by target organ system (e.g. neurotoxins). The emergence of computational tools that take what we know about individual chemicals and convert this information into an interactive map describing areas of risk by summarizing and weighting data will provide a new view that assists the lay person in determining the impact of the industrial world on the health of their community. User-friendly databases are being constructed in many countries and many examples are given in a detailed study by Catelan et al (2000). Leading-edge applications and user interfaces based on spatial data warehousing and spatial OLAP, such as the SPHINX (Alberta) and ICEM-SE (Quebec) projects, will provide a tool that gives access to aggregated and detailed information as well as both outside and in-house information (given proper access rights) in both aggregated and single datasets.

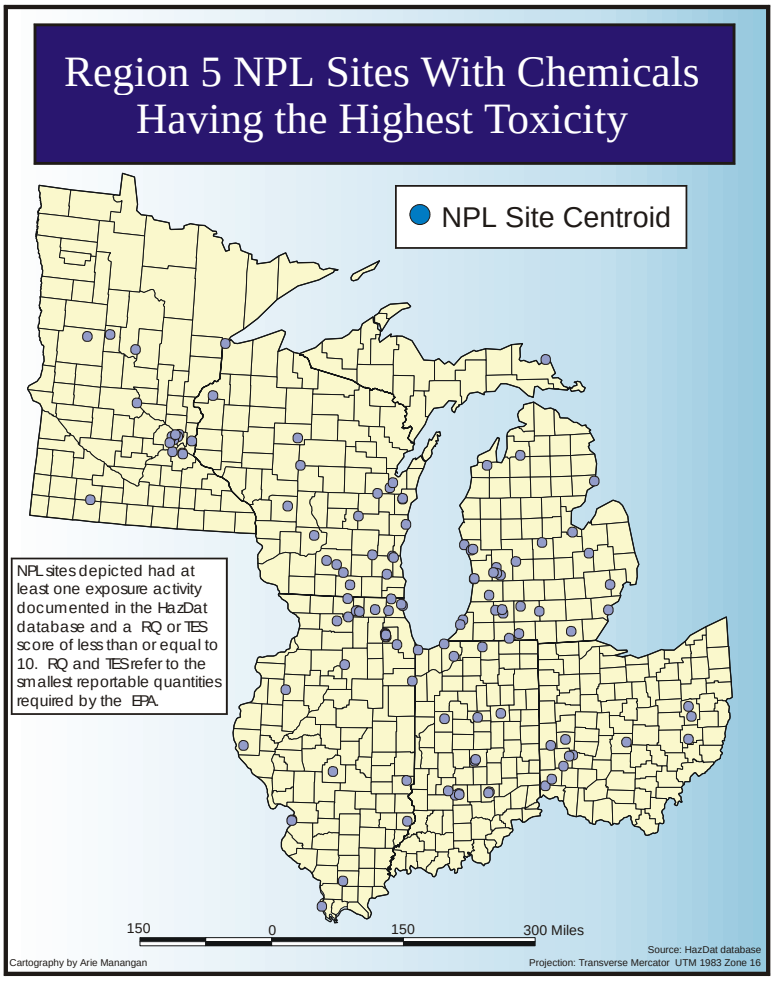

Figure 5: ATSDR's HazDat database was queried to identify only those sites with documented off-site exposure activities and containing substances considered most toxic.

Applications, Benefits and Resources Required

As our world becomes increasingly industrialized, populations that are more susceptible to adverse effects of environmental degradation will need refuge. The elderly, women of reproductive age and youth have a right to know where levels of pollution exceed what is considered to be "safe". Once identified, interim health protection measures must be put in place for these populations, as chemicals will persist in the environment long after the beginning of clean up efforts. A more practical approach to health protection can be taken once we examine all regional contributions to chemical exposure, not only those of industrial origin. Resulting information could influence urban planning by considering the relationships between environmental pollution sources and behaviour and geographic formations.

Challenges For Successful Implementation and Suggestions For Future Orientations

Assessing such things as the dose of a compound that may result in an adverse health effect to humans is rarely an easy task. Computer networks and legacy mainframe systems contain billions of records of analytical samples collected at hazardous waste sites that document the levels of priority pollutants found in air, water, and soil. Rather than considering the individual toxicity of compounds at the parts-per-billion level, it may be more prudent to consider the relative toxicity of compounds (possibly to specific target organs) as an approach to weighing toxicity of chemicals over a broad geographic area. We can enhance the traditional epidemiological studies by taking into account the possibilities offered by modern GIS such as spatial statistics and geographic overlays with other datasets (e.g. exposure modeling results, sociodemographics, land use, topography) to better qualify exposure and risk levels. These areas can then be investigated in greater detail using more detailed health and environmental information in legacy systems.

However, when one wants to identify the biggest challenges to integrate such modern IT in the day-to-day work of environmental health specialists, one must look at the activities related to data access, data usability assessment and data standardisation (Gosselin et al 2000). Knowing what data exists where remains a challenge as well as obtaining these data (costs, confidentiality) and transforming them in a usable format (restructuring, recoding, validating, aggregating, geocoding, etc.). Other major challenges identified by Gosselin et al (2000) include the provision of more training (both in formal and informal settings, including having access to technical support) and finding adequate funding to sustain and build capacity for the use of evidence-based tools built with modern information technologies. The real challenges ahead of us are more administrative/political than technological, more driven by data finding/access difficulties than by the technology and more about the usability of systems and adequate training/support than about the technology.

Conclusions

Technology is rapidly changing the type and amount of information that is accessible to specialists and the public. New technologies are providing the necessary tools for decision making and analysis. In the near future, we will be able to find simple responses to simple

questions such as the best place to raise a family, report on local and regional health statistics for this location, and provide directions to surrounding health services. This will be done with information that exists today, but using modes of interaction that we are just beginning to develop.

References

Bédard Y., P. Gosselin, M. Jerrett, S. Elliott, D. Mowat, J. Moore, M. Goddard, R. Catelan, A. Gingras and P. Poitras 2000. Recent Technological Trends vs Users' Needs in Health Surveillance, a Canadian Study. International Health Information Conference "Infocus 2000", co-organized by the Canadian Institute for Health Information and Canada's Health Informatics Association, Vancouver, Canada, June 24-27, 9 pages.

Bédard Y., T. Merrett and J. Han 2000. Fundamentals of Spatial Data Warehousing for Geographic Knowledge Discovery. Chapter of the book "Geographic Data Mining and Knowledge Discovery" edited by H. Miller and J. Han, Research Monographs in GIS series edited by Peter Fisher and Jonathan Raper, Taylor & Francis; to be published.

Brackett M.H. 1996. The Data Warehouse Challenge: Taming Data Chaos. John Wiley & Sons, 579 pages.

Catelan R., P. Gosselin et Y. Bédard 2000. Évolution récente de la pénétration de la géomatique dans le domaine de la surveillance en santé : revue bibliographique. Article soumis à la Revue Internationale de Géomatique, Éditions Hermès, Paris.

Date C.J. 2000. An Introduction to Database Systems, 7th Edition. Addison-Wesley, 938 pages.

Gisler W. 1986. The Uses of Spatial Analysis in Medical Geography: A Review. Social Science in Medicine 23:963–73.

Gosselin P., Y. Bédard, M. Jerrett, S.J. Elliott, R. Catelan, P. Poitras and A. Gingras 2000. GIS and OLAP in Health Surveillance: Needs Analysis for Successful Integration. Final report prepared for Health Canada, Ottawa, February 10th, 72 pages.

Inmon W.H., D. Richard, and D. Hackathorn 1996. Using the Data Warehouse. John Wiley & Sons, 285 pages.

Kulldorff M. 1998. Statistical Methods for Spatial Epidemiology: Tests for Randomness. In: GIS and Health in Europe. Ed. A Gatrell, M Löytönen. London: Taylor & Francis

Longley P.A., M.F. Goodchild, D.J. Maguire and D.W. Rhind 1999. Geographical Information Systems: Principles, Techniques, Applications, and Management, 2nd Edition. John Wiley & Sons, New York, 1101 pages.

Petridou E, Revinthi K, Alexander FE, Haidas S, Koliouskas D, Kosmidis H, Piperopoulou F, Tzortzatou F, Trichopoulos D. 1996. Space-Time Clustering of Childhood Leukemia in Greece: Evidence Supporting a Viral Etiology. British Journal of Cancer 73:1278–83.

Poe V. 1995. Building a Data Warehouse for Decision Support. Prentice Hall, 210 pages.

Roney N, Henriques WD, Fay M, Holler J, Susten S. 1998. Determining Priority Hazardous Substances Related to Hazardous Waste Sites. Toxicology and Industrial Health Vol 14, No.4, 521-531.

Williams-Johnson MM, Henriques WD and Fay RM. 1996. Investigating Ratios of Health Effect Levels Using ATSDR's HazDat Database: Extrapolation Methodologies in Quantitative Risk Assessment. J. Clean Technol., Environ. Toxicol. & Occup. Med. Vol. 5(4): 347-360.