Data Management in Medicine: The EPIweb Information System, a Case Study and some Open Issues

Pierpaolo Vittorini and Ferdinando di Orio

University of L'Aquila
Department of Internal Medicine and Public Health
67100 L'Aquila (AQ) - Via S. Sisto
ITALY

pierpaolo.vittorini@cc.univaq.it

Abstract. Data management in medicine usually takes place through the usage of heterogeneous and legacy systems. So far, these information systems were rarely reusable into other investigations and, expecially in multicenter studies, their development absorbed part of the given funding. To overcome these limitation, we developed the EPIweb system, i.e. a totally configurable web-based information system which helps the epidemiologists/public health practitioners to conduct their (distributed) researches by following a workflow made up of the selection of the remote centers, the creation of the questionnaires, the data entry, the statistical processing of the data, and the generation of the technical reports. The most important EPIweb features are shown in the paper by means of a recent investigation conducted with the system. Finally, we discuss some open issues regarding the improvement of the EPIweb system in the direction of scientific workflow management.

1 Introduction

To support their decisions, clinicians and public health practitioners use both information technology, statistical and operational research methods [1–3], usually applied to massive amount of data stored on databases (e.g., [4]).

Recently, the need for "[...] collecting and analyzing morbidity, mortality, and other relevant data and facilitate the timely dissemination of results to appropriate decision makers [...]" [5] has lead to the development of several Internet-based information systems, among the many we mention PHIN [6] which proposes an automated exchange of data between public health partners through ebXML compliant web services [7], ESSENCE system [8] which uses a secure file transfer protocol to send the data over the Internet, RSVP [9] which uses a combination of web and Java technologies to collect the data, and a long list of information systems developed to support their related multicenter studies (e.g., [10, 11]).

So far, these sophisticated information systems were rarely reusable and, expecially in multicenter studies, their development absorbed part of the given funding. Furthermore, some of these systems focused only on the data collection activity, and the tasks of analyse, display, report and map the collected data were demanded to specialized software. To overcome these limitations, we present the EPIweb system [12], i.e. a web-based information system [13] which helps the epidemiologists to conduct their studies by adopting a clear and guided workflow made up of the following phases: the selection of the remote centers (also through randomization), the creation of the questionnaires, the data entry, the statistical processing of the data, and the generation of the technical reports. All these phases are configurable, therefore the system is both flexible and responsive to changing requirements, and also reusable and "tailorable" to different investigations. Furthermore, the underlying database supports information integration since data coming from different sources can be integrated in the form of answers given to different questionnaires.

To show these features, an investigation entirely performed through the EPIweb system regarding the middle school students' nutritional habits in the Municipality of L'Aquila is discussed. The "lesson learned" from this study suggests several implementation improvements and particularly in the direction of scientific workflow management.

2 EPIweb features and architecture

EPIweb implements the workflow highlighted in figure 1: an epidemiological study starts with the selection of the remote centers, continues with the questionnaires development, the data entry, the data analysis, the report generation, and ends with a discussion of the achieved results. The following three typologies of users "interact" with the workflow: the epidemiologists (hereafter called study administrators) which organize and manage their studies, the remote centers which participate in the data entry, and the users which read the reports and discuss the published results.

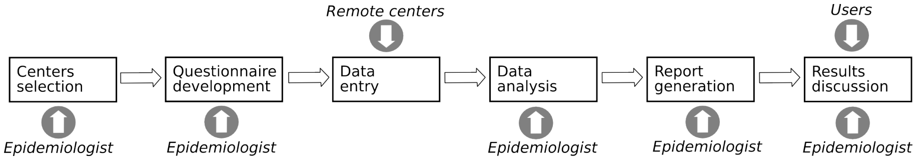


Fig. 1. The EPIweb workflow of an epidemiological study

The EPIweb system is a web-based application following the architecture depicted in figure 2. At the highest level of abstraction, the apache web server [14] takes care of both the communication with the clients – when necessary through a secure connection – and the execution of the PHP scripts [15]. The scripts implement all the EPIweb features, by relying on the MySQL database [16] to store the data, on STATA™ [17] to execute the statistical analyses, and on the LaTeX typesetting system [18] to produce the technical reports. Since the proposed architecture incorporates LaTeX and STATA™ like two software components reused in a wider project, a reduction of the overall development efforts was achieved [19]. This architecture also reduces the deployment efforts, i.e. the costs connected with the set of activities which follow the initial release, since the software updates or bug fixes are applied only in the EPIweb system.

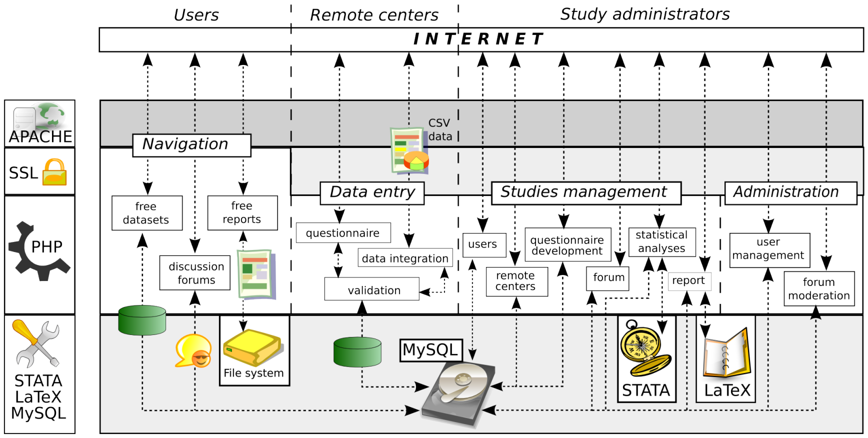


Fig. 2. The EPIweb architecture

In detail, the architecture is divided into four sections, namely: the navigation, the data entry, the studies management and the administration sections.

The navigation section is opened to the users, and offers the possibility to download free datasets and reports, and to debate the results of the available investigations through a discussion forum. The free datasets are generated on-the-fly by a PHP script which extracts from the underlying database the proper tuples. The free reports are available as PDF files stored in the file system.

In the data entry section, the remote centers can fill in the questionnaire and/or integrate into the underlying database any old dataset if available as CSV data. We remark that the entered data is validated towards several constraints defined by the epidemiologist. The constraints may regard:

- that a question must be considered as a primary key;
- an allowed range for a numeric answer;
- when a missing value can be accepted;
- the definition of relationships between questions, i.e. a question can be placed if a certain statement is true. The interface of the EPIweb system currently allows the development of statements in which a question is placed only in connection with the value of a previous answer.

The studies management section allows an epidemiologist to organize its study in terms of the following areas: (i) users, (ii) centers, (iii) questionnaires, (iv) statistical analyses, (v) reports, and (vi) forums. The users area offers the possibility to involve other users in the investigation. In the centers area, remote centers can be included/excluded from the data entry. The questionnaires are organized as a list of questions created through a visual interface, and can be developed to honor the above mentioned constraints. The following kind of data types can be used:

- Free text (limited to 80 characters);
- Integer values (the range is -2147483648 to 2147483647);
- Floating values (the allowable values are $-3.402823466 \cdot 10^{+38}$ to $-1.175494351 \cdot 10^{-38}$, 0, and $1.175494351 \cdot 10^{-38}$ to $3.402823466 \cdot 10^{+38}$);
- Sets of values (stored as integers).

To process the data, the administrators define a list of analyses by selecting the desired one among a set of predefined statistical analyses and graphs, or by creating new variables. Then, STATA™ is invoked in background by providing it with a dataset and a do-file containing the needed commands. As output, STATA™ produces (i) a log file with the results of the analyses and (ii) a collection of EPS files containing the graphs. In case, the collected data can also be exported in the CSV format – readable by most programs like Excel, Epi Info™, STATA™ and SAS™ – and analysed through an external software. The technical reports are created from a list of comments regarding the results of the most significant statistical analyses: the EPIweb system automatically "assembles" all the given comments/results,automatically creates a `tex` file, then invokes LaTeX to produce the PDF file, and finally displays such a file as the resulting technical report. The report contains: (i) a title page, (ii) a page with a note and the list of the participants of the study, (iii) the table of contents, (iv) a chapter with a description of the epidemiological study and a note which foregoes the results, (v) a central chapter with all comments/results regarding the analysis, (vi) the description of the questionnaire and of the dataset, (vii) the list of tables, and (viii) the list of figures.

Finally, in the administration section, a discussion forum regarding the results of a study can be opened and moderated. Furthermore, a user management facility is also provided.

The database was conceptually modelled (see E/R diagram depicted in figure 3) to allow the storage (i) of the epidemiological data organized in terms of answers given to questionnaires, (ii) of the statistical analysis as a collection of processings taken from certain statistical models, and (iii) of the reports from comments given to selected processings. In detail, a dataset comes from a *questionnaire*, whose `name` is stored in the database. A questionnaire is made up of a list of *question*s. The questions are organized as a double-linked list, where the `next`/`prev` attributes point to the next/previous questions, respectively. The `type` attribute indicates the datatype of the expected answer and the `opt` attribute stores its possible options. The `null` attribute indicates whether a missings value might be accepted, and the `text` attribute contains the text of the question. The `ic` attribute is used to store an eventual internal consistency check. Actually, it is under consideration the possibility to model constrainst using the results coming from business rules [20]. To a certain question, an *answer* is given, whose value is stored in the proper attribute (either `int`, `float` or `str80`). The flag `missing` is obviously used to represent a missing value. Furthermore, the time in which the answer was entered is stored in the `time_stamp` attribute. A statistical *processing* is the actual application of a certain *model* in a given analysis. Therefore, a *model* has a `name`, a

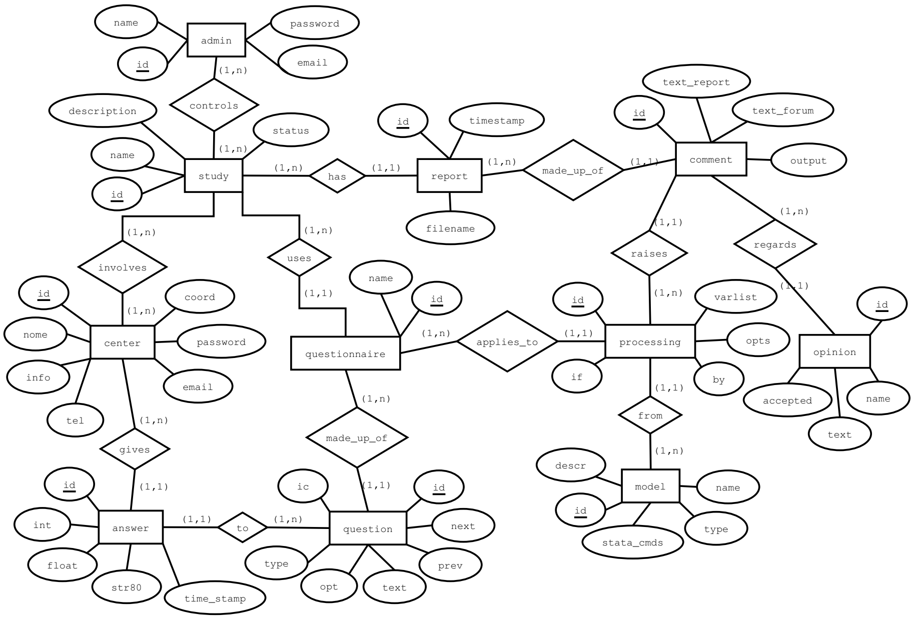


Fig. 3. The E/R diagram of the underlying database

type which specifies whether it is a statistical analysis or a graph, and a stata_cmds attribute which contains its abstract STATA implementation. A *model* becomes "tangible" by means of the proper varlist, by prefix, if qualifier and opt options, stored in the *processing* entity. Hence, the *model*s represent all the available statistical analyses/graphs, while the *processing*s are the actual analyses executed by an epidemiologist in his/her investigation.

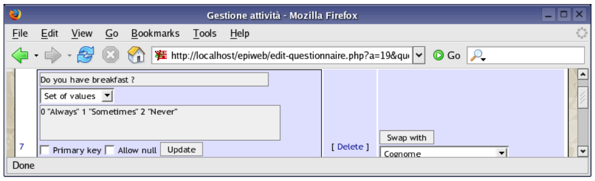
This conceptual model supports information integration since data coming from different sources can be stored in the database as answers given to ad-hoc developed questionnaires. In other terms, each different data source is represented by a questionnaire, whose questions must reflect the metadata structure of the original data source.

3 The case study

Hereafter, a sample system run which shows how EPIweb supported an investigation regarding the middle school students's nutritional habits in the Municipality of L'Aquila (Italy) [21] is reported.

The epidemiologist connects to the server, logs in as an EPIweb administrator, and creates a new study by giving it a name (i.e. "Middle school students' nutritional habits in the Municipality of L'Aquila") and a short description. Hence, the epidemiologist decides to select the remote centers through simple randomization. By clicking on the proper link, a pop-up window is opened, and the epidemiologist (i) adds all the possible remote centers, (ii) randomly selects a subset of a chosen size, and (iii) stores the resulting centers into the database. Then, the epidemiologist develops the questionnaire by sequentially adding the needed questions. In this phase, the epidemiologist has to take attention in specifying e.g. when a missing value might be accepted, the list of the allowed answers (by using the associative sets), the range of acceptance, when a question should be asked, etc.

Figure 4 shows the question regarding if the student has breakfast. The expected answer belongs to a set of choices: 0=Always, 1=Sometimes, and 2=Never. Therefore, the epidemiologist creates the "Do you have breakfast?" question as a "Set of values" with a related associative set specified with the string `0 "Always" 1 "Sometimes" 2 "Never"` (see figure).


Gestione attività - Mozilla Firefox

http://localhost/epiweb/edit-questionnaire.php?a=19&qu

Do you have breakfast ?

Set of values

0 "Always" 1 "Sometimes" 2 "Never"

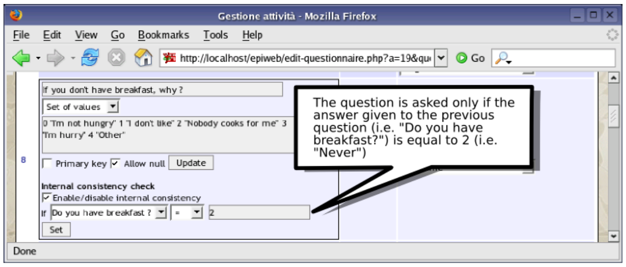
7 □ Primary key □ Allow null

Swap with Cognome

Done

Fig. 4. The questionnaire creation area (first question)

Figure 5 shows the subsequent question, which regards the reason because the student does not have breakfast. Since this question must be asked only to the students who answered "Never" to the previous one, an internal consistency check must be applied: the question is asked only if the answer given to the "Do you have breakfast?" question is equal to 2 (i.e. "Never").


Gestione attività - Mozilla Firefox

http://localhost/epiweb/edit-questionnaire.php?a=19&qu

If you don't have breakfast, why ?

Set of values

0 'I'm not hungry' 1 'I don't like" 2 'Nobody cooks for me" 3 'I'm hurry' 4 'Other"

8 □ Primary key ☑ Allow null

Internal consistency check

☑ Enable/disable internal consistency

If Do you have breakfast ? = 2

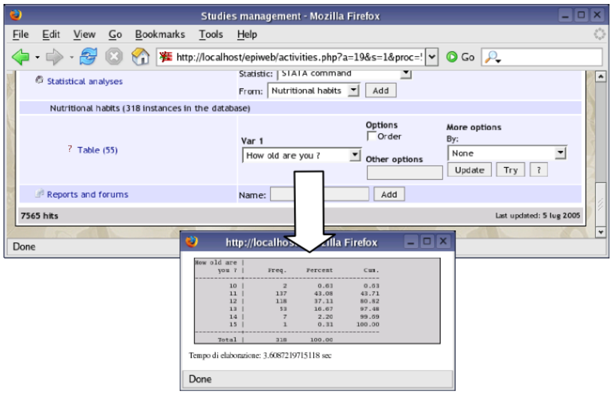
Done

The question is asked only if the answer given to the previous question (i.e. "Do you have breakfast?") is equal to 2 (i.e. "Never")

Fig. 5. The questionnaire creation area (subsequent, constrained question)

When the questionnaire is ready, the epidemiologist enables the data entry. The remote centers log in the EPIweb system and fill in their questionnaires question by question. Depending on the constraints given by the epidemiologist, the system either accepts or rejects the questionnaire.

When the data entry is over, the epidemiologist begins the statistical processing of the data. Initially, he decides to investigate the age of the students. To this aim, he creates a table and a pie chart regarding the question "How old are you?". Figure 6 shows the interface used to create the statistical analysis, and the corresponding result. With similar operations, the epidemiologists continues the analysis and adds all the needed statistics, tests and graphs.



Studies management - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://localhost/epiweb/activities.php?a=19&s=1&proc=!
Go
Statistical analyses
Add
Table (55)
Options
□ Order
More options
Update
Try
?
Reports and forums
Add
7565 hits
Last updated: 5 lug 2005
Done
http://localhos... ...zilla Firefox
Done

Statistic: STATA command
From: Nutritional habits
Nutritional habits (318 instances in the database)
Var 1: How old are you ?
Other options:
By: None
Name:

```
How old are |
      you ? |      Freq.     Percent        Cum.
------------+-----------------------------------
         10 |          2        0.63        0.63
         11 |        137       43.08       43.71
         12 |        118       37.11       80.82
         13 |         53       16.67       97.48
         14 |          7        2.20       99.69
         15 |          1        0.31      100.00
------------+-----------------------------------
      Total |        318      100.00
```

Tempo di elaborazione: 3.6087219715118 sec

Fig. 6. The age distribution

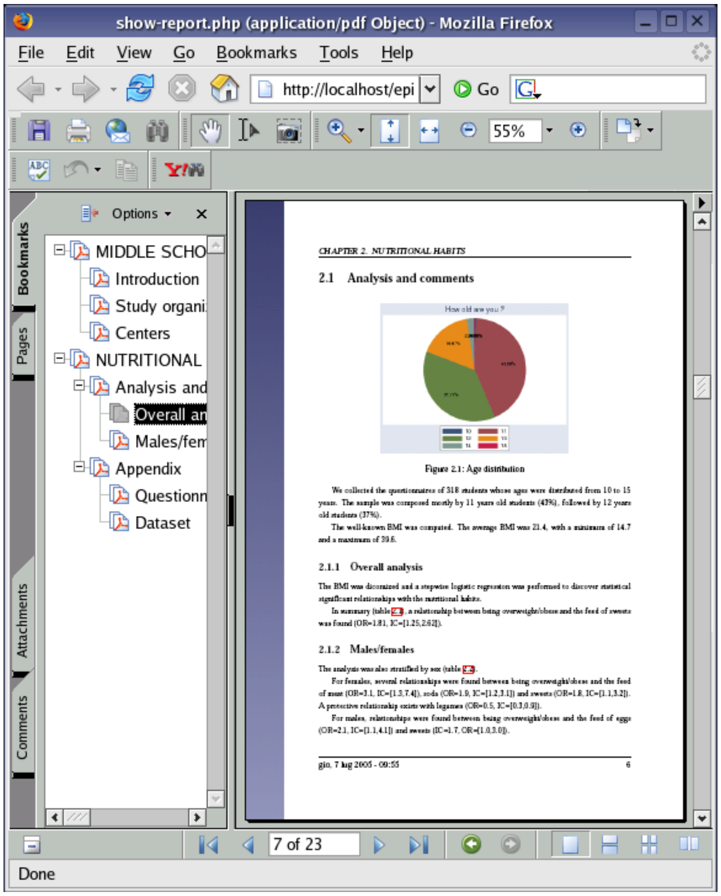
Finally, he starts the development of the technical report. We recall that the report is created by the EPIweb system by "assembling" the comments given to the analyses/graphs that the epidemiologist considers relevant. Figure 7 finally depicts the produced report.

We remark that a study can be also repeated periodically, since the timestamp of each answer is stored in the dataset. Also the center which entered the answer is stored in the dataset, therefore, comparisons regarding e.g. centers, repeated studies can be performed.

4 Discussion

The paper discussed the need for an integrated and flexible approach to effectively collect, validate, analyse, display and report epidemiological data. We have briefly shown in the sample system run that the EPIweb system can be easily "tailored" to the specificities of an epidemiological study, and that it can handle a large variety of investigations by properly changing the questionnaire and the statistical analyses. It is worth remarking that the EPIweb system implements a centralized flow of information, which makes available both the statistical investigations and/or the technical reports regarding the most important research findings timely.

The study regarding the nutritional habits was also used to assess the advantages of the EPIweb system, throughout its comparison with similar investigations conducted



CHAPTER 2. NUTRITIONAL HABITS

2.1 Analysis and comments

Figure 2.1: Age distribution

We collected the questionnaires of 318 students whose ages were distributed from 10 to 15 years. The sample was composed mostly by 11 years old students (43%), followed by 12 years old students (37%).

The well-known BMI was computed. The average BMI was 21.4, with a minimum of 14.7 and a maximum of 39.6.

2.1.1 Overall analysis

The BMI was dicotomized and a stepwise logistic regression was performed to discover statistical significant relationships with the nutritional habits.

In summary (table 2.1), a relationship between being overweight/obese and the feed of sweets was found (OR=1.81, IC=[1.25,2.62]).

2.1.2 Males/females

The analysis was also stratified by sex (table 2.2).

For females, several relationships were found between being overweight/obese and the feed of meat (OR=3.1, IC=[1.3,7.4]), soda (OR=1.9, IC=[1.2,3.1]) and sweets (OR=1.8, IC=[1.1,3.2]). A protective relationship exists with legumes (OR=0.5, IC=[0.3,0.9]).

For males, relationships were found between being overweight/obese and the feed of eggs (OR=2.1, IC=[1.1,4.1]) and sweets (IC=1.7, OR=[1.0,3.0]).

gio, 7 lug 2005 - 09:55 6

Fig. 7. The technical report

with traditional methods. The physician who supervised the data entry, who also weighed and measured the students, reported that the data collection activity was greatly improved by the adoption of the EPIweb system: the data was collected timely and directly in a digital format, and the entering of typos was prevented. Nevertheless, the epidemiologist who managed this study, even if satisfied of the processing capabilities, asked for improving the report generation interface which allows only a stereotyped organization.

Further open issues arised during the investigation. The workflow embedded in the EPIweb system which "drives" an epidemiological study was not clearly perceived by the epidemiologist, which asked for a more clear and intuitive interface. Furthermore, we state that a fundamental support could be given in the selection of the best statistical analysis useful to reach a certain research objective. We point out that this aim can be achieved by implementing in the next release of the EPIweb system a scientific workflow management based on a hierarchical statistical analyses organization like that

- Which is the central tendency?
    - Variable: numerical
        - Measure
            - Mean
        - Graph
            - Histogram
            - Steam-leaf plot
            - Box graph
            - Frequency polygon
    - Variable: ranks
        - Measure
            - Median
            - Mode
        - Graph
            - Bar graph
            - Box graph
    - ...
- Which is the dispersion?
    - Variable: numerical
        - Variance
        - Standard deviation
    - Variable: ranks
        - Max-min
        - IRQ
    - ...
- Do you want to investigate into the normality deviation?
    - Variable: numerical
        - Skewness
        - Kurtosis
- Is there a difference? [22]
- Is there an association? [22]
- Are there two or more independent variables? [22]

Fig. 8. A possible hierarchical organization of (a portion of) the statistical analyses useful for medical research

highlighted in figure 8, which extends the flowcharts for relating reseach questions to statistical methods available in [22]. In the figure, the research objective is the main question, then further “refinements” are required (e.g., the variable type, if a graph or a measure is needed), until the proper statistics is selected. For instance, to measure the dispersion of a numerical variable (i.e., not expressed in terms of ranks, neither frequencies), the system could propose to evaluate either its variance or standard deviation.

References

1. Raghupathi, W., Tan, J.: Strategic IT applications in health care. Communications of the ACM **45** (2002) 56–61
2. Hauck, K., Smith, P.C., Goddard, M.: The Economics of Priority Setting for Health Care – A Literature Review. The World Bank (2003)
3. Eubank, S., Guclu, H., Kumar, A., Marathe, M., Srinivasan, A., Toroczkal, Z., Wang, N.: Modelling disease outbreaks in realistic urban social networks. Science **429** (2004) 180–4
4. Capocaccia, R., Gatta, G., Roazzi, P., Carrani, E., Santaquilani, M., De Angelis, R., Tavilla, A., the EUROCARE Working Group: The EUROCARE-3 database: methodology of data collection, standardisation, quality control and statistical analysis. Annals of Oncology **14** (2003) 14–27

5. Sosin, D.M.: Draft framework for evaluating syndromic surveillance systems. Journal of Urban Health **2 Suppl 1** (2003) 8–13
6. Loonsk, J.: PHIN overview. In: 1st Annual PHIN Conference. (2003)
7. OASIS: ebXML. Available on-line (2005) http://www.ebxml.org/.
8. Lombardo, J.S., Burkom, H., Pavlin, J.: ESSENCE II and the framework for evaluating syndromic surveillance systems. Morbidity and Mortality Weekly Report **53** (2004) 159–165
9. Zelicoff, A., Brillman, J., Forslund, D.W., George, J.E., Zink, S., Koening, S.: The rapid syndrome validation project (RSVP). Sandia National Laboratories (2001)
10. Sanson, R.L., Morris, R.S., Stern, M.W.: EpiMAN-FMD: a decision support system for managing epidemics of vesicular disease. Revue scientifique et technique (International Office of Epizootics) **18** (1999) 593–605
11. Silva, S., Gouveia-Oliveira, R., Maretzek, A., Carriço, J., Gudnason, T., Kristinsson, K.G., Ekdahl, K., Brito-Avô, A., Tomasz, A., Sanches, I.S., de Lencastre, H., Almeida, J.: EU-RISWEB – web-based epidemiological surveillance of antibiotic-resistant pneumococci in day care centers. BMC Medical Informatics and Decision Making **3** (2003)
12. Vittorini, P., Necozione, S., di Orio, F.: The information technology for the management of health care data: the EPIweb project. Epidemiologia e Prevenzione (2005) . In press.
13. Deshpande, Y., Hansen, S.: Web engineering: creating a discipline among disciplines. IEEE Multimedia **8** (2001) 82–87
14. Laurie, B., Laurie, P.: Apache: The Definitive Guide (3rd Edition). O'Reilly (2002)
15. Sklar, D.: Learning PHP 5. O'Reilly (2004)
16. DuBois, P.: MySQL, Second Edition. Sams (2003)
17. Stata Corporation: Stata Base Reference Manual (4 volumes). Stata Press (2003)
18. Lamport, L.: LaTeX: A Document Preparation System. Addison Wesley (1994)
19. Mili, H., Mili, A., Yacoub, S., Addy, E.: Reuse Based Software Engineering: Techniques, Organizations, and Measurement. Wiley (2002)
20. Hay, D.C.: Modeling business rules: the business constraint metamodel. The Data Administration Newsletter (TDAN.com) (2004)
21. Cesare, B., Vittorini, P., Sallusti, E., Bontempo, V., Graziani, M., Necozione, S., di Orio, F.: Le abitudini alimentari dei ragazzi di scuola media nel Comune di L'Aquila: risultati di uno studio descrittivo. In: IX Conferenza Nazionale di Sanità Pubblica. (2005) in Italian.
22. Dawson-Saunders, B., Trapp, R.G.: Appendix C: Flowcharts for relating reseach questions to statistical methods. In: Basic & Clinical Biostatistics. Appleton & Lange (1994)