Oncology over Internet (O2I): information querying, filtering and retrieval by means of standard formats

P. Romano(1), P. Arrigo(2), G. Mauri(3), E. Merelli(4) e L. Milanesi(5)

(1) National Cancer Research Institute IST, Genova, Italy
(2) Electronics Circuits Institute ICE CNR, Genova, Italy
(3) Department of Systems and Computer Sciences and Communication DISCO, University of Milano Bicocca, Milano, Italy
(4) Department of Mathematics and Computer Sciences, University of Camerino, Camerino (MC), Italy
(5) Biomedical Technologies Institute ITB CNR, Milano, Italy

The huge amount of biomedical data that is currently available through Internet and the heterogeneity of the softwares that are used for their distribution make the task of searching, filtering and retrieving the information very difficult for the researchers. Often, data is manually analysed. Transfer between remote sites and the local computers is carried out by means of FTP, HTTP (web interface) or even by the well known "cut and paste" technique.

The need is felt for a system that is able to improve access to information, by making it more efficient and efficacious. Such a system could be represented by an improved user interface that should be able to answer the requests coming from end users, even the most complex ones, by autonomously accessing remote sites, databases and softwares, and by integrating the different results in a unique, filtered, integrated answer.

Such an integration can only be carried out by specialized softwares that are able to "understand" the information they have to cope with. These softwares will interoperate with "technologically advanced" remote information sources in order to retrieve the proper information and/or submit preliminar and intermediate data, until the desired result is achieved.

Among the tools that are currently available to this end, the most promising ones are XML based languages, CORBA and SOAP based servers and intelligent software agents platforms. XML (eXtensible Markup Language) allows for the definition of specialized DTDs (Document Types Definitions) that describe the contents of a piece of information by splitting it in subfields with associated reference names. CORBA (Common Object Request Broker Architecture) is an open standard for the interoperability of distributed objects, while SOAP (Simple Object Architecture Protocol) is a framework for the distribution of XML structured information by means of the HTTP protocol. Intelligent agents and multiagent systems are innovative softwares particularly convenient for distributed elaboration systems.

"Oncology over Internet - O2I" is a project that has recently been funded by the Italian Ministry of Education, University and Research (MIUR). Its goals are the design, the development and the setting up of an intelligent prototype system, accessible through Internet, that allows for the carrying out of a limited set of complex searches of oncological interest. These searches will normally involve access to and retrieval from a heterogeneous set of web sites and databases, by also using a varied set of innovative software tools. The system will be able to filter the results of the searches and integrate them in a unique answer.

The system will rely on a meta database (or knowledge base) including the defined complex searches and, for each of them, the involved databases, the ways for accessing, searching and extracting the information of interest, and of integrating results. The knowledge base contents will be finalized to the oncology field and will include three streams: basic research, clinics, general information.

The expected result of the O2I project is the availability of a prototype system, devoted to oncology, adequately tested and validated by oncologists and oncology researchers, exportable and publicly available on Internet.