On the Challenge of Creating and Communicating Air Quality Information: A Case for Environmental Engineers*

Leo Wanner[1,3], Bernd Bohnet[2], Nadjet Bouayad-Agha[3],
François Lareau[3], Achim Lohmeyer[4], Daniel Nicklaß[2]

[1]Institució Catalana de Recerca i Estudis Avançats (ICREA), [2]University of Stuttgart,
[3]Universitat Pompeu Fabra, Barcelona, [4]Ingenieurbüro Lohmeyer, Karlsruhe

Abstract

Building environmental information systems is a challenging task. It involves a detailed analysis of the addressee's needs, development of air pollutant concentration assessment and interpretation techniques, which must be tailored to the addressee, as well as the development of information planning and production strategies. Environmental engineers are increasingly often in charge of setting up environmental services. They must be thus knowledgeable in a number of issues they did not face before. We present an air quality information system MARQUIS, which is usable, on the one hand, as a base service that can be extended and adapted to a new environmental information scenario, and, on the other hand, as a source of teaching and training material for environmental engineers.

1. Introduction

Creation and communication of air quality information (AQI) is a challenging task. Environmental engineers still too often think of it in terms of tables and graphics, with a possible reference to general background information on the nature of the individual air pollutant substances and their relevance to human health. However, we must be aware that

(i) non-specialists (in particular, general public, but also decision makers in politics and industry) are not able to assess and interpret "raw" pollutant concentrations in order to draw proper conclusions for their own behaviour, respectively decisions;
(ii) the kind and amount of the information relevant to the information addressee depends on his health, prior knowledge and cultural characteristics; often, it is essential to go beyond the mere pollutant concentrations;
(iii) the mode in which the information is to be communicated depends on the nature of the data and, again, on the characteristics of the addressee; while tables and graphics may be adequate for specialists, text presentation is certainly more adequate for general public.

In order to be able to adequately address these aspects of creation and presentation of air quality information, environmental specialists need to acquire skills which go considerably beyond the skills traditionally required in the field. Firstly, they must learn that AQI is not uniform for all its users. As in the case of any information service, different types of potential addressees exist – with all consequences this implies with respect to need studies, user typology construction, etc. Secondly, they must learn how models for assessment and interpretation of air quality data are being built. Such models can be conceived either as rule-based *expert system shells* or as *ma-*

* Corresponding authors: Leo Wanner (leo.wanner@upf.edu) and Daniel Nicklaß (DN@ier.uni-stuttgart.de)

chine learning applications. In other words, they must be knowledgeable in fundamentals of Artificial Intelligence and Machine Learning. Thirdly they must be able to evaluate (if not develop) the information production programs (natural language text generators, graphic generators, and table generators) available for use and understand the nature of input structures they require. That is, they must be knowledgeable in the fundamentals of certain fields of Applied Natural Language Processing.

The *Multimodal Air Quality Information Service for General Public* (MARQUIS)[†] can serve as a case study of the areas involved in creation and communication of air quality information, as a base system that can be extended to a new environmental scenario, and also as a showcase for environmental engineers.

2. The Case Study: General Overview of the MARQUIS-Service

The need for new generation, user-tailored AQI services is increasingly acknowledged. Some proposals for such services have already been made. The most well-known among them is, without doubt, APNEE; cf., e.g., (Bøhler et al., 2000; Peinel et al., 2000; Johansen et al., 2001). MARQUIS shares with APNEE the basic assumptions such as the importance of the individualization of air quality information, the need to offer the service via all major modern communication channels, etc. However, unlike APNEE, MARQUIS furthermore provides a contextual interpretation of the measured and forecasted air pollutant concentrations and it uses advanced natural language processing techniques for the production of AQI. In short, the characteristic features of the MARQUIS-Service are:

(i) coverage of the major air pollutant substances monitored in five European regions;[‡]
(ii) reference to a default user profile typology, with the option of a flexible individualization of each profile by the users;
(iii) coverage of the major modern communication channels: web, email, mobile phone services (SMS, WAP, and MMS), TV, and printed media;
(iv) use of advanced air quality assessment and interpretation models;
(v) use of natural language processing techniques for planning and generation of multimodal and multilingual material.[§]

Technically, MARQUIS is a distributed service in that the raw AQ data verification and selected forecasting modules are run remotely on the servers of the data providers, while the common

---

[†] MARQUIS was funded by the EC in the framework of the eContent Programme under the contract number EDC-11258. Its consortium was composed of Barcelona Media Fundació Universitat Pompeu Fabra as coordinator, University of Stuttgart, Televisió Catalunya, Servei Meteorològic de Catalunya, Landesanstalt für Umwelt Baden-Württemberg, Ingenieur Büro Lohmeyer, Finnish Meteorological Institute, Departamento de Ciências e Engenharia do Ambiente, Universidade Nova de Lisboa, Institute for Ecology of Industrial Areas.

[‡] The following regions are covered in MARQUIS: Baden-Württemberg (Germany), Catalonia (Spain), Finland, Portugal and Upper Silesia (Poland).

[§] The information is offered in terms of texts, tables and graphics (including pictograms). Textual information is provided in eight languages: in the primary languages of the regions (German, Catalan and Spanish, Finnish, Portuguese, Polish) as well as in English (the "lingua franca of the EU") and French (the language of the neighbour region of Baden-Württemberg and Catalonia).

data base, the assessment and interpretation and information production modules are run on the central MARQUIS-service. Given that nearly all involved tasks are time consuming, the realization of a most efficient architecture possible was an issue. A two-pipe architecture with the following distribution of the tasks among the pipes proved to be most adequate:

Pipe 1:

1. Monitoring of air pollutant concentrations and of meteorological conditions in the five MARQUIS-regions and execution of data quality assurance and air pollution forecasting models.[**]
2. Delivery of the measured and forecasted data from the local DBs to the MARQUIS-server, where they are kept in the central DB in a uniform format.
3. Assessment and interpretation of the delivered data with respect to their relevance to any of the EU and regional environmental legislation issues and to any of the MARQUIS users with a profile from the user profile typology; determining the primary meteorological and contextual influence on the measured and forecasted air quality (for explaining / justifying them). Air quality forecast models may also be run in this stage when required.

Pipe 2:

1. Receiving an information request from a user via the MARQUIS-Client interface (this can be an automated periodic or a singular request).
2. Selecting the content that is relevant for the user in question from the structure produced in Step 3 of Pipe 1.
3. Generating the discourse structure of the content to be conveyed to the user, determining the appropriate mode for the individual chunks of the content, and starting the corresponding information generators.
4. Generating the information by table, graphic and multilingual text generators.
5. Conveying the generated information to the user using his/her preferred communication channel.

To account for these tasks, the architecture contains the following modules: central MARQUIS DB, AQ assessment and interpretation module, document planning module, and information generation module. Each of the modules consists of several submodules. For a more detailed presentation of MARQUIS, see, e.g., (Wanner et al., 2007) as well as the publicly available deliverables of the MARQUIS-project.

Given its modular architecture and the coverage of all major tasks that need to be addressed in the framework of an environmental information service, MARQUIS can be exploited as a showcase for such services. More precisely, it can be used, as already mentioned above, as a case study of the areas involved in creation and communication of air quality information, as a base system that can be extended to a new environmental scenario, and also as a training device. Only when environmental information specialists get a chance to analyze the advantages and shortcomings of existing systems *in situ* will they be able to build more adequate services.

[**] The meteorological data are provided by the Finnish Meteorological Institute for Finland, Baden-Württemberg and Upper Silesia; by the Portuguese Meteorology Institute for Portugal, and by the Catalan Meteorological Service for Catalonia.

In what follows, we first discuss the possibility to use MARQUIS as a multifunctional environmental service beyond its current application and then how it could be used as a teaching and case study material.

3. MARQUIS as Multifunctional Environmental Information Service

The most obvious exploitation of MARQUIS beyond its scope defined by the project is its extension to a not yet incorporated region, covering the same range of pollutant substances as already treated by the Service. However, it can be also used for the production of information on different pollutant substances (such as pollen) and/or different environmental conditions (such as weather, snow and or water quality, etc.). Let us discuss in turn the tasks involved in both kinds of extension – and thus the tasks faced by environmental engineers.

3.1 Extending MARQUIS to Further Regions

AQI is regional. Therefore, the option to use MARQUIS for the production of AQI for regions not yet covered by the Service is obvious. For this purpose, two alternative running modi of the Service can be pictured: as a multiregional service or as a mono-regional service.
Originally, MARQUIS-Service has been conceived as a multiregional centralized service; see above. Therefore, its extension region by region has already been practiced and is thus a known procedure.
When the Service is operated in the multiregional mode, the following actions must be performed to add a region:

(1) ensure the delivery of the air quality data monitored and forecasted in the region in question to the central MARQUIS-DB in the appropriate format
(2) if the Service is to address a user group not captured by the MARQUIS user profile typology, this user group must be introduced and default types of information to be offered to members of this group must be determined
(3) introduction of the air quality index scale information and all region-specific health warnings, advices, etc.; if any information is user group-specific, this must be specified accordingly
(4) extension of the interpretation and assessment module to cover the regional index and all user related information
(5) if the language of the region is not covered yet, the lexical and grammatical resources for this language must be developed.

For the use of the full potential of the MARUIS-service, an air quality forecast model or manual forecast of the major pollutant substances of the region in question is essential.
The basic end user profile typology as identified in MARQUIS is likely to be similar for any European region. At the global level, this typology distinguishes between environmental specialists, medical professionals, patients with diseases aggravated by negative AQ-conditions, and general public – with more fine-grained distinctions introduced at subsequent levels. However, depending on the target of the Service, the introduction of new user profiles (such as joggers or

senior citizens) might be desirable. The necessity of the extension of the user typology should be derived from a qualitative addressee study in the region.
Furthermore, if the region in question is in a country not yet covered by MARQUIS, it is likely that it will have an own AQ-index, own health warnings and correlations between AQ- and meteorological conditions. These peculiarities must be incorporated into the assessment and interpretation module.
The content selection criteria and the way the information is presented is assumed in MARQUIS to be the same for users with a given profile across the regions.
As a consequence of the market study in a given region, the Service can be also operated as a mono-regional service in which the AQI is provided to one region only. In this case, there is no need for a central DB and a central server; the data delivery stage can be thus omitted.

3.2 Adapting the MARQUIS-Service to other Environmental topics

The kernel of the MARQUIS-Service can also be used to report on other environmental issues such as meteorological conditions, air quality, snow quality, etc. Obviously, the cost for getting the service running on one of these topics is higher than for air pollution. Thus, the assessment and module and parts of the document planning module need to be newly developed. In the case of the assessment module this is clear: the interpretation of the raw data is strictly domain-specific. As far as the document planning module is concerned, the content selection strategy and communication criteria for the determination of the discourse structure of the information presentation are likely to be different since they are, again domain specific (Kittredge et al., 1991). In contrast, the procedure for the derivation of the discourse structure based on the selected content and the communication criteria is universal and does not need thus to be reimplemented.
Furthermore, at least the generation lexica must be extended in order to cover the vocabulary of the new topic. An empirical study must ensure that the grammatical constructions encountered in the textual information on the topic are covered by the language generators suffice.

4. MARQUIS as a Showcase

MARQUIS covers all primary topics involved in the production of user tailored air quality information. It can be thus used as a showcase and source of inspiration for environmental engineers. In this section, we describe how each of the topics has been realized in MARQUIS and how it can be addressed in the general scenario of environmental information services.

4.1 Addressing the User

User-oriented topics are reflected by the user typology and user interfaces.

4.1.1 User Typology

Air quality information must be tailored to specific user groups (Johansen et al., 2001). Often, as system developers, we are tempted to think that we know what the different user groups are and what kind of information each of the groups needs. However, interviews with users may turn out to be a surprise. Empirical studies of the user market and interviews with a representative number of users with different backgrounds are absolutely indispensable. In MARQUIS, a qualitative

study has been carried out first in Catalonia and then adapted to the other test regions. The study consisted in structured and semi-structured personal interviews with representatives of target groups considered as potential users of the service. About 100 individuals have been interviewed in each region. All results were analyzed to define the user profiles and the preferences of the users with respect to the type of content and presentation mode. The user study resulted in a definition of an initial user profile typology which predefines the content to be provided to a user who accesses the service for the first time and identifies him-/herself as member of one of the user groups of the typology. Figure 1 displays the user profile typology.

1. Consultant/AQ-Technician
2. Medical professional
    - 2.1 Asthma specialist
    - 2.2 Specialist for other respiratory diseases
    - 2.3 Specialist for non-respiratory diseases
3. Patient
    - 3.1 Asthma patient
    - 3.2 Patient- Other respiratory diseases
4. General Public
    - 4.1 Printed Media Reader
    - 4.2 Television Viewer
    - 4.3 Mobile Phone User
    - 4.4 Internet User
5. Information Broker
    - 5.1 Printed Media
    - 5.2 Television
    - 5.3 Mobile Phone Services
    - 5.4 Internet Portals

Figure 1: User profile typology in MARQUIS

The initial user profiles can be customized with respect to content, language, mode of presentation, etc. via a web-based interface. The option of customization is important, e.g., in the case of cross-border users, of people who travel from one region to another, or users who are end users and information brokers at the same time.
The user typology itself can also be extended in order to account for new user groups.

4.1.2 Service-User Interfaces

It is needless to underline the importance of well-designed interfaces in a user-oriented information service. In accordance with its coverage, MARQUIS offers interfaces to web, WAP, SMS, and email. Tests were also performed with a telephone-based service interface. For printed media and television, the layout and the script of the presentation, respectively, are elaborated individually with each information broker client.
The design of the interfaces was carried out in several iterations; after each iteration, user feedback was sought and subsequently evaluated.

4.2 Air Quality Assessment and Interpretation

The assessment of pollutant concentrations for purposes of forecasting is generally recognized to be an important aspect of air quality information services. Much work has been done in this area, Here, it suffices to say that in MARQUIS, regional multiple regression and machine learning based models are used (Lohmeyer et al., submitted). Machine learning (ML) techniques provide an alternative to the more common multiple regression models. It is worth to evaluate the whole range of ML techniques, i.e., not only neural networks, which are already rather widely used for air quality forecasting, but also, e.g., *Nearest Neighbour*,[††] *Support Vector Machines*, etc.

The interpretation of the air pollutant concentrations with respect to their relevance to the users summarized in the user typology must form the central part of any air quality information service. The corresponding module within the architecture of the service can be considered an "expert system shell" since it contains the know-how of the air quality specialist in a distilled form. Different ways to realize the interpretation are possible. In MARQUIS, a rule-based implementation has been chosen; machine learning based techniques are also suitable for this task.
MARQUIS's interpretation module processes the following topics:

(i) European and region-specific AQ-indices,
(ii) health relevance for specific user profiles,
(iii) significant air pollution changes over a certain time period and exceedance of European and national thresholds,
(iv) justification of air quality conditions with meteorological conditions exploiting the correlations derived from the assessment models or provided by specialists.

The interpretation procedure is a multistage procedure. In each stage, one of the above topics (i)-(iv) is addressed. The output of the procedure is an XML-assessment structure. The structure is used by the document planning module to select the content to be communicated to a specified addressee as well as its form.
The first step in the assessment procedure is the computation of the regional AQ indices. The indices are computed for each region. That is, the FIN-index is computed not only for FIN, but also for the other four regions; the same applies to the indices of the other regions. This supports the cross-border view on air quality in a given region. The indices are computed hourly as well as for the whole day. This allows the service to provide air quality information of a specific time of the day as well as for the day as a time interval.
The health relevance of a given pollution is determined by matching the measured concentrations or the index values with the concentration–health effect tables compiled, again, for each region. Table 1 illustrates such a table for Finland. The table associates index value intervals with health effects. For a better comprehension by the user, each interval is associate a color.

[††] In MARQUIS, already a simple variant of the Nearest Neighbour technique showed a good performance; cf. (Lohmeyer, submitted).

Index	Color	Class	Health impacts	Other long-term impacts
151 -	Violet	Very poor	Adverse effects possible on sensitive subpopulation	Clear impacts on vegetation, material impacts
101 – 150	Red	Poor	Adverse effects possible on sensitive individuals	Clear impacts on vegetation, material impacts
76 -100	Orange	Fair	Unlikely effects	Clear impacts on vegetation, material impacts
51 - 75	Yellow	Satisfactory	Very unlikely effects	Mild environmental impacts
0 - 50	Green	Good	No health effects	Mild environmental impacts

Table 1: Concentration – health effect association in Finland

As not every measured index value or pollutant concentration is important or interesting for the users, in the second step of the assessment procedure, VIPs (very important points) are identified within the daily distribution curve. Standard VIPs are daily and local maxima, respectively minima. In addition, the first value of the day and the current value is a VIP. To determine the VIPs, a standard mathematical curve discussion is carried out. In order to avoid the identification of small (and thus irrelevant) local maxima/minima, certain smoothing strategies are applied. In Figure 2, the VIPs identified by the implemented assessment procedure are marked by red circles; the dashed line circles mark VIP-candidates eliminated during smoothing.

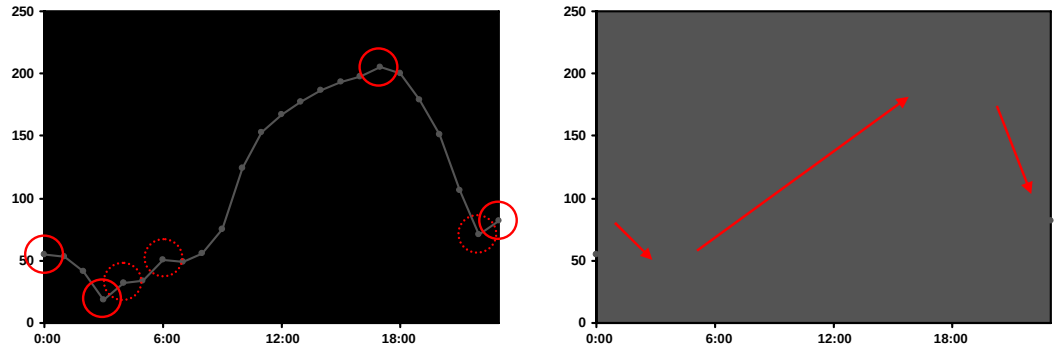

Figure 2: The detection of daily minima and maxima, decrease / increase of the AQI

National and European thresholds for the different pollutants are considered. To allow different region specific reports, the exceedance of these thresholds are provided in two ways. First, the threshold type and the start and end time of the exceedance are given, and, second, the degree of exceedance is provided in absolute and relative numbers.

An important feature of the MARQUIS-service is the availability of plausible explanations for measured and forecasted air pollutant concentrations. The correlation between air quality and meteorological conditions is the most important source for such explanations. A comprehensive range of potential correlations has been studied. The most relevant meteorological conditions include wind speed and direction, temperature, global radiation, precipitation, pressure difference between reference locations, relative humidity. External conditions such as day of the week (i.e., traffic), number of days since the last significant precipitation, etc. are also considered.

4.3 Content Selection and Information Planning

Content selection and planning of the "discourse structure" of the presentation is one of the two tasks related to computational linguistics that are the least familiar ones for environmental engineers. While the engineers do not need to become specialists in automatic information generation, they need to learn the basics in order to be able to choose the right off-the shelf modules and provide the input these modules require.

As the name suggests, content selection and information planning consists of two parts, which have often been treated in automatic information generation as one task, but which are rather different in their nature: content selection and discourse structure determination.

In multimodal information production, a third aspect comes into play: the selection of the adequate mode of representation (table, graphics, text or a combination of several modi) for each piece of information to be conveyed to the user.

Content selection deals with the choice of the relevant content for the user who asked for information. In the case of air quality information services this is most often straightforward: pick those content chunks from the output structure of the assessment and interpretation module that are relevant for the user according to his/her profile.

The question which content requires a textual form of presentation, which content is better presented as table, and which in terms of a graphics or a pictogram is to be addressed from the perspective of the user and from the perspective of the information presentation theory. Thus, the user might have a preferred mode in which he/she would like to have the information presented. In MARQUIS, he/she can specify this in the course of the personalization of the profile. These preferences should be followed as long as they are plausible from the information presentation theory point of view. When they contradict the principles of adequate information presentation, they must be overruled – as, e.g., the presentation of the concentration of a pollutant over a certain period of time.

The question of the discourse structure concerns first of all the presentation of textual information. It needs to be determined whether the order in which the information is to be presented is predetermined enough to use a text template or a more flexible strategy is needed because the order in which the information is presented depends on different contextual criteria. In MARQUIS, a hybrid strategy has been realized (Bouayad-Agha et al., 2006).

4.4 Information Generation

In accordance with the three modi in which AQI can be presented – text, table, and graphics – three different information generators must be integrated into an AQI service. Since table and graphics are more common in traditional AQI services, we address text generation only. In MARQUIS, the MATE-generator is used (Bohnet, 2006). MATE is a graph transducer based generator that is based on the multi-stratal Meaning-Text linguistic model, MTM (Mel'cuk, 1988). Six strata are relevant to us: conceptual, semantic, deep-syntactic, surface-syntactic, topological, and surface-morphological strata. Generation consists in mapping an input conceptual structure retrieved from the text plan provided by the document planning module to the surface-morphological structure (= text) via the structures of the intermediate strata listed above. The mapping of the structures between two adjacent strata is realized by a distinct language-specific grammar. In order to generate information in a specific language, thus five grammars and the

corresponding lexica must be developed. To facilitate the work of non-linguists, MATE offers a grammar development environment with a debugger, grammar inspector, interactive editors, etc. (Bohnet et al., 2000).

5. Summary and Conclusions

In order to be able to build adequate air quality (and, more generally, environmental) information services, environmental engineers must be knowledgeable in a number of areas that go well beyond their traditional core tasks – for instance, user modelling, machine learning techniques and automatic information generation using natural language processing techniques. However, they do not need to start from the scratch. Rather, services such as MARQUIS and APNEE already exist that can be used as base services and case study material.

References

Lohmeyer, A., I. Düring, M. Giereth, T. Hoffmann, D. Nicklaß, M. Klein, H. Scheu-Hachtel, C. Sörgel und L. Wanner. submitted. Kurzfrist-Feinstaub-Immissionsprognose für MARQUIS und LÜSA Sachsen-Anhalt. Gefahrstoffe und Reinhaltung der Luft.

Bøhler, T., Karatzas, K., Peinel G., Rose, T., San José, R. 2000. Providing Multi-Modal Access to Environmental Data – Customizable Information Services for Disseminating Urban Air Quality Information in APNEE. In *Proceedings of the 22nd Urban and Regional Data Management Symposium*.

Bernd Bohnet. 2006. Textgenerierung durch Transduktion linguistischer Strukturen. Köln: DISKI 298.

Bohnet, B.; Langjahr, A., Wanner, L. 2000. Development Environment for an MTT-Based Sentence Generator. In *Proceedings of the First International Conference on Natural Language Generation*. Israel, Mitzpe Ramon.

Bouayad-Agha N., Wanner, L., Nicklaß, D. 2006. Discourse Structuring of Dynamic Content. In: *Proceedings of SEPLN'2006.* Zarragoza, Spain.

Johansen, P.H., Karatzas, K., Lindberg, J.E., Peinel, G., Rose, T. 2001. Citizen-Centered Information Dissemination with Multimodal Information Channels. In Sustainability in the Information Society, Environmental Informatics 2001. 15th International Symposium, Informatics for Environmental Protection. Zurich, Switzerland.

Kittredge, R., Korelsky, T., Rambow, T. 1991. On the need for domain communication knowledge. *Computational Intelligence*. 16.3:408-445.

Mel'čuk, I. 1988. *Dependency Syntax.* Albany, NY: SUNY Press

Peinel, G., T. Rose and R. San José. 2000. Customized Information Services for Environmental Awareness in Urban Areas. In *Proceedings of the 7th World Congress on Intelligent Transport Systems*.

Wanner, L., Nicklaß, D., Bohnet. B., Bouayad-Agha, N., Bronder, J., Ferreira, F., Friedrich, R., Karppinen, A., Lareau, F., Lohmeyer, A., Panighi, A., Parisio, S., Scheu-Hachtel, H., Serpa, J. 2007. From Measurement Data to Environmental Information:MARQUIS - A Multimodal AiR QUality Information Service for the General Public. In *Proceedings of the International Symposium on Environmental Software Systems*.