D2.5 Data mediation

Project: ROADIDEA 215455
Document Number and Title: D2.5 Data mediation – How to convert data with different formats
Work-Package: WP2
Deliverable Type: Report
Contractual Date of Delivery: 30.09.2009
Actual Date of Delivery: 04.10.2009
Author/Responsible(s): Rene Kelpin (DLR)
Contributors: Gustaf Axelson, Antje Parnitzke (DLR)
Approval of this report: Coordinator, TC
Summary of this report: Report describes ROADIDEA data mediation architecture for input and output operations
Keyword List: Data, mediation, data archive, application, XML, format, request, response, Eprints, JAVA
Dissemination level: Public (PU)
Change History

<table>
<thead>
<tr>
<th>Version</th>
<th>Date</th>
<th>Status</th>
<th>Author (Partner)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Version0.1</td>
<td>27.08.2009</td>
<td>Draft</td>
<td>Gustaf Axelson (DLR)</td>
<td>First structural draft</td>
</tr>
<tr>
<td>Version0.2</td>
<td>09.09.2009</td>
<td>Draft</td>
<td>Gustaf Axelson (DLR)</td>
<td>Introduction and flow-charts created</td>
</tr>
<tr>
<td>Version0.3</td>
<td>16.09.2009</td>
<td>Draft</td>
<td>Rene Kelpin (DLR)</td>
<td>Output mediation chapter, application pictures added</td>
</tr>
<tr>
<td>Version0.4</td>
<td>25.09.2009</td>
<td>Draft</td>
<td>Rene Kelpin, Gustaf Axelson (DLR)</td>
<td>Executive summary and conclusion added</td>
</tr>
<tr>
<td>Version1.0</td>
<td>30.09.2009</td>
<td>Final</td>
<td>Rene Kelpin, Gustaf Axelson (DLR)</td>
<td>Final checking and internal DLR approval</td>
</tr>
<tr>
<td>Version1.1</td>
<td>04.10.2009</td>
<td>Final</td>
<td>Pirkko Saarikivi (FORC)</td>
<td>Quality checked</td>
</tr>
</tbody>
</table>

Distribution list

European Commission
Emilio Davila Gonzalez
Wolfgang Höfs

ROADIDEA partners
E-mail list
Executive summary

With the present report the ROADIDEA data mediation architecture is described and concluded. This mediation architecture combines the work of the preceding WP2 tasks in only one WP2 data handling system - the ROADIDEA data mediation.

After investigating information about available data sources in the project member states and storing corresponding data set descriptions in a preliminary data archive in the first work task 2.1, available data was analysed in terms of its usability (task 2.2) and interoperability with other data sets (task2.3). Taking main findings of those analyses into consideration, a common data model and a corresponding data format was developed (task 2.4). Finally both the archive and the format were implemented into the data mediation architecture as a base for project related data input and output relations. Right from the start all preceding WP2 efforts were aiming at the preparation and the definition of such a data mediation system.

The main goal of the data mediation approach, which was also followed within WP2 of the project ROADIDEA, is to mediate between data providers and data users on one hand and a data handling architecture on the other hand – as secure, reliable, and generic as possible. Direct access to the data-bases and repositories is not desired; new data sets need to be well described and analysed before being stored in a common data format; data availability requests have to be made and responded highly standardised. If these constraints are given, implementation of the data mediation architecture is the best choice. It has to be considered as the middleman between the data archive and the data applications such as transport services in the given project context.

This report describes the prerequisites and the implementation of the ROADIDEA data mediation application for both data streams: into the system and out of the system. For the input mediation demonstration three different data samples from the Gothenburg area are utilised. After the data preparations and the definition of corresponding data dictionaries the report describes the JAVA application and its usage.

The consideration of the output case distinguishes between the search and browsing functionality of the underlying Eprints platform and the JAVA output application. Both systems are described. The main difference between those two systems is that only the data mediation output application is able to process ‘machine-made’ data requests, while the Eprints system needs a human interaction. Thus, the mediation may be called by and answer to automatic generated request. However, with the described JAVA application this functionality could not be demonstrated.

A corresponding XML data request format was created and applied for the demonstrating application. Both the output mediation and the Eprints system are accessing the ROADIDEA data archive with all data sets described and analysed in the preceding work tasks. If data sets fitting to the selected choices and keywords are available in the data archive, a data request will be responded with a list of data set. The list comprises the most important data set information (ID, title, abstract and internal contact information). Search criteria and listed data set information can be adapted easily. ‘Machine-made’ requests will be answered with a data set list stored and submitted in a specific XML format.
With the described data mediation architecture a powerful tool for project related data access is given. This architecture has to be considered as the basis on which data related project applications may be premised. To provide an easy-to-use option for implementing new applications and utilisation was the main goal while developing the data mediation architecture. It merges all WP2 findings and results onto one system.

However, the underlying data archive determines the value of applicability of the system. That means that a system like the described ROADIDEA data mediation architecture depends strongly on investigated, described and stored data set descriptions. Thus, the data source investigation may never be considered as concluded. In the remaining runtime of the project ROADIDEA WP2 will continue with the data source investigation not only in the participating countries. Besides that, new utilisations of the provided architecture are to be defined in co-operation with data related work packages in ROADIDEA.
Table of contents

1. Introduction ........................................................................................................ 6
2. The ROADIDEA mediation approach .............................................................. 9
3. Input dataflow to mediation ........................................................................... 11
   3.1 Raw data cases ............................................................................................ 11
   3.2 Methodology ............................................................................................... 12
4. Output dataflow from mediation .................................................................... 16
   4.1. General considerations ................................................................................ 16
   4.2. Methodology .............................................................................................. 17
5. Conclusions ........................................................................................................ 21
1. Introduction

ROADIDEA WP2 is intended to support the projects ideas with data. The main idea with the project ROADIDEA is to create new ideas and to do innovations. In WP2 the data related prerequisites were considered in order to pave the way for reliable ROADIDEA services and innovations.

Data itself is no innovation! The innovation is the ideas and the creations of new services for ITS within the project. But to be able to fulfil the intentions with some of the ideas, data is needed. Thus, in this last task in WP2, the main findings and results of the preceding work tasks of WP2 were taking into account for following the ROADIDEA data mediation approach.

Since the beginning of the project ROADIDEA it was well known that there are difficulties in utilising and combining different data from different sources together on a common platform. This report describes the idea of and the work necessary for implementing the ROADIDEA data mediation project architecture.

In task 2.4 it was decided to use an XML-based ROADIDEA common format in the data platform aimed for supporting the pilots and upcoming data related ideas. The format is described with report D2.4 [ROADIDEA2009a]. This XML format is able to describe all conceivable data sources of value for the ROADIDEA project context - with all its included parameters and descriptors. But so far, this XML format is not more than an empty structure.

The data mediation in WP2 is aimed to transfer available raw data sets into this empty structure - into the specific ROADIDEA XML format. Both the data input by data providers and data output requests by users (e.g., service providers) is implemented as a website-embedded Java application called “The ROADIDEA data mediator”. The mediator needs to know how to sort and structure available information from the sources into the XML format. Therefore, to each dataset, a describing dictionary is created with all available (explicitly and implicitly) Meta information.
Figure 1 shows the input scheme of the ROADIDEA data mediation. The data providers are able to access the Mediator in order to store their data descriptions in the ROADIDEA data archive. For each data input case a specific dictionary file has to be created, which is to be considered as the translator between the raw data into the internal system language.

It has been well known during the previously made investigations that even if the project have access to data with similarities to other sources, it is not possible to use these sources together without some kind of treatment.

With chapter 3 of this report the data input user interface is described. It was implemented as a Java applet providing a web-based environment for establishing communication channels between data providers and the ROADIDEA data archive. It transfers raw data sources directly to the internal ROADIDEA XML format and stores the data in the archive.
Output

Figure 2, Data mediation - Output schema

Figure 2 visualises how the data output (data request) communication flow is working. In a JAVA implemented graphical user interface (GUI), the user is able to filter data sources by making specific selections. This produces a data request, which is to be translated into the internal language - ROADIDEA XML. This request will be processed and answered with a list of corresponding data sets. The output mediation is described in chapter 4 of this report.

To simplify the XML format generally; it can be described as a well structured book with all chapters created. But the book contains no useful information at this point. It is also a book without any kind of content of value as well as no interest for a further user. Therefore, the content in the book is needed to be added. This is meant with “mediation” and the process to perform this task is the “mediator”.

Further on, when the mediator should put the information into the book, the mediation program needs to know how to read the data and where to put the data into the book. At this point dictionaries are needed to understand the data. The mediator uses the dictionaries like a lexicon to understand the content in the data. The specific dictionary to each of the dataset also explicitly point out the name that is used for each parameter in the XML format. This is needed in order to put the desired information to a target with correct position within the XML format.

For demonstrating the data mediation approach data for the Gothenburg pilot was used. The Gothenburg pilot focus on the combination of weather related data and traffic data.

It has to be pointed out clearly that the demonstration is not more than one example of what the data mediation architecture is able to accomplish. Currently, it is not possible to use this specific mediation application on live data. But, for future development, all relevant and given information about the investigated data sources is available with the ROADIDEA data mediation application.
2. The ROADIDEA mediation approach

The basic approach of the project is to find a solution for the usage of different kind of data together. It was found that there often are difficulties using and combining data from different sources, even if the similarities between the sources are not that big.

The idea with the data mediation is to show that it is possible to read raw data files and put the information into the ROADIDEA XML format. For development purpose, historical data is used. The mediator, a website-embedded JAVA application (JAVA applet), consists of two parts:

- Data input mediation
- Data output mediation.

The Mediator uses dictionaries files. These dictionaries have to be considered as translators between the very diverse raw data environment and the internal, well structured ROADIDEA data archive. For the data input case the Mediator has to know for instance where to find specific information in the raw data file and where to put it in the XML archive file. The translation of data requests of any data users into an archive language is the focus of the output mediation case.

The demonstration tool will be available on a webpage as a GUI, Graphical User Interface. The Java tool is accessing the data archive with data stored in the XML format. Neither data providers nor data users have direct access to the ROADIDEA data archive. Users of the application have the possibility to trigger the generation of a data request by using the data output GUI and get a reply in terms of a list of corresponding data sets.

The Figure 3, Dataflow within WP2, shows an overview of the planned dataflow. The picture represents the architecture of WP2 with its included tasks. The picture aims to give an overview of proceeded dataflow through the structured data ontology. The dataflow is primary represented as input- and output dataflow.
Input dataflow

1. Information about data is stored
2. The data is collected from the provider, it can for example be an authority or a private company that can provide data
3. The input data is transferred from raw data into the XML format
4. The data is available in the data archive.

Output dataflow

1. A user wants to access data
2. The request is made with the user interface
3. The information request is transferred to access the Data Archive
4. Data is provided to the user via the user interface
3. Input dataflow to mediation

Generally, the input data for the mediation task is the data aimed for the four ROADIDEA pilots. In this feasibility demonstration case, data sets intended for the Gothenburg pilot are used. This chapter will clarify how the data is put into the mediation.

3.1 Raw data cases

The data sources from the Gothenburg pilot aimed for usage are:
- Traffic measurement
- Road Weather from stations
- Airport Weather, METAR

The following figures 4, 5 and 6 shows the applied sample from the different data sources aimed for the mediation. The data from all used sources are text based and also the same parameters, for example temperature and wind are available in the sources. All the referred data sets were described more detailed in previous WP2 D2.4 report.

Figure 4, Traffic measurements

Figure 5, Road Weather

Figure 6, Airport Weather
3.2 Methodology

The mediation consists of two different parts:

- **Data input - from provider**
- **Data output - on request to end user (next chapter)**

The data input case was to read the raw data files into the internal ROADIDEA XML format. This is done with the input Mediator. Input mediation dictionaries were designed and used for each data source in order to assign specific raw data content to a specific field of the ROADIDEA XML format.

Each dictionary includes two different types of information:

- **Header**: The first part is information about the dataset, such as contact information, data owner etcetera. This information is not explicitly given with the raw data set, but it is from great value for a later utilisation of the data for both users and future developers. The information was investigated and collected in task 2.1 of WP2.
- **Content**: The second part in the dictionary defines the content of the raw data set connected with the corresponding field’s names in the XML format.

The following figures 7, 8 and 9 show the dictionaries for the Gothenburg Pilot, that are created to make it possible to understand and to read the raw data files automatically. It can be seen in the files that the structure and the content is similar - except the number of columns in a specific raw data set.

The “header” explicitly shows additional information with describing background knowledge as well as among others position of the measurements. All three dictionaries for the mediation demonstration are included in the report in order to show that the structure of the files is always the same and even that all parameters are common. This is a prerequisite to a more generic approach, which might be followed with a further developed version of the described mediation system.

![Figure 7, Dictionary traffic data](image)
With the given dictionary files a mapping between the data set columns and the corresponding fields of the XML format was prepared. It has to be considered as an attempt to make the data sets with its entire description values machine readable.

The input mediation application

The mediation approach is implemented into the ROADIDEA website (WP2 / WP2 Results) as a web-enabled JAVA application. This application is the link between the raw data from the different sources and the ROADIDEA XML format. It performs the translation from arbitrary raw data sources into the well defined XML format.

To simplify the XML format generally; it can be described as a well structured book with all chapters created. But the book contains no useful information at this point. It is also a book without any kind of valuable content as well as no interest for a further user. Therefore, the content in the book is needed to be added. This is meant with “mediation” and the process to perform this task is done by the “mediator”.

Further on, when the mediator should put the information into the book, the mediation program needs to know how to read the data and where to put the data into the book. At this point the dictionaries are needed to understand the data. The mediator uses the dictionaries like a lexicon to understand the content in the data. The specific dictionary to each of the dataset also explicitly point out the name that is used for each parameter in the XML format. This is needed in order to put the desired information to a target with correct position within the XML format.
The figure 10 shows the structure of the input case for the data mediation application. It mainly focuses on three levels:

- Data input,
- Mediation processing
- Storage of data in the archive.

The data input part considers the raw data with its describing dictionaries. But also the ROADIDEA XML format is part of the input because the mediation uses the XML format as a structure to fill the data into. Raw data and complementing dictionaries to the data set are read into internal and temporary data storage before the mediation process is able to use this data. It is at this point, the data is ready to be used for its desired task. It has to be pointed out that it makes no difference whether the raw data sources are online or historically available.
Next part of the program is processing of the input data content using raw data content as well as dictionary’s descriptors to fill the ROADIDEA XML format with data. This is the main function for the mediation task. At first the header information given explicitly in the dictionary are filled into the corresponding fields of the XML format. Then, a loop controls the filling of data set content (data set columns) successively. If no more data set elements are available the loop terminates. At the end the XML file will be closed, named with connection to the raw data sources and stored in the ROADIDEA data archive.

Figure 11 shows the data input application. It is online available at the ROADIDEA website. As already mentioned above it is demonstrating the general feasibility of the ROADIDEA input mediation approach by using three test data set samples from the Gothenburg pilot only.

The processing of the input data is continued by the mediator until all available raw data is used and filled into the XML format and stored in the ROADIDEA archive. For this input mediation demonstration, only a limited amount of historical data is used.
4. Output dataflow from mediation

In the previous chapter it was described how data was merged into the data archive. Chapter 4 will focus on how to access that specific data. From a user point of view, the output is the interesting part, where it is possible - via the user interface - to request and get data descriptions from the archive. The described output mediation interface is a demonstration of the general archive accessibility, which uses historical data from the Gothenburg pilot as well.

- Data input – from provider (previous chapter)
- Data output – data on request to end user

4.1. General considerations

In an early phase of the data source investigation period in coordination with the Technical Committee and in consideration of

- the given European heterogeneous availability of data relevant for the ROADIDEA project purpose and
- the very large differences in terms of data set size, format and online accessibility (e.g. online text based road traffic observations vs. static Italian fog satellite radar images)

it was decided to store META information about relevant data sources in the ROADIDEA data archive only. Thus, even though it is possible (and demonstrated with the data input mediation described in chapter 3 of this report) to transfer arbitrary raw data sets into the ROADIDEA data archive format, the mediation output takes into account META information only.

WP2 offers two ways to access and browse the ROADIDEA data archive:

1. Eprints search and browsing functionality: Under http://www.roadidea.eu/packages/wp2/default.aspx the Eprints data archive architecture (set up in WP2 task 2.1) can be used for map visualisation of and content browsing through the ROADIDEA META data archive.

2. Output mediation demonstration: With a JAVA based output mediation application at http://www.roadidea.eu/packages/wp2/mediation.aspx, data set search requests can be created. This application demonstrates the translation of these requests into the internal XML language and gives back a list of corresponding data sets.

While describing the [Eprints] (open source; more information available at http://www.eprints.org/) browsing functionality briefly chapter 4.2 emphasises more on the methodology and the functionality of the output mediation demonstrator.

The most important difference between the described approaches (Eprints and output mediator) is that the output mediator may also be triggered by XML requests generated by service applications. Those requests could be sent to the ROADIDEA data mediation system and answered via common communication channels such as HTTP. No human
interaction with the ROADIDEA data archive is needed. Data availability requests on the ROADIDEA data archive could be implemented directly into applications or software solutions.

- **Data input** – from provider (previous chapter)
- **Data output** – data on request to end user

The main goal of the described output mediation task is to show available data sources, its usability and applicability in a user-friendly way. It allows users to browse the ROADIDEA data archive in order to get to know available road traffic, road weather, and supplementary data sources in Europe.

### 4.2. Methodology

**The Eprints server**

Eprints is an open sources software package and a very easy way to set up repositories for technical and non-technical data. The ROADIDEA data archive, set up within WP2 task 2.1, is based on the Eprints software package. The system has been adapted to the special needs of the ROADIDEA project. Its main strength is the included browse and search functionality. For the ROADIDEA data archive, this functionality is combined with a Google Map data visualisation, which allows a graphical and, thus, a very intuitive access to the data archive.

Figure 12 shows the entry page of the Eprints web user interface. In a European map the Eprints user interface marks the ROADIDEA project countries and shows the general availability of data sources for a selected data class. The visualisation is accessing all data set information stored in the Eprints data archive. The top level information is shown at the context taps, while the entire data set description may be accessed by following the ‘More information’ link.
The purpose of the ROADIDEA Eprints archive is to give an overview of available data presented in a graphical way. The presented overview depends strongly on the results of the performed data source investigation in task 2.1. The focus of that data source investigation was on the project member states. How useful such an archive could be, if information about data source availability all over Europe were known and well described?

Further more, the archive explicitly shows the differences in terms of format, availability, and reliability between – at first sight – similar data sources, and underpins the necessity of a:

- system, which mediates data between providers, users, and a central system on one hand and
- the definition of a common data archive, which was accomplished with the preceding work task 2.4 (see also [ROADIDEA2009a]) on the other hand.

It can clearly be seen in the data samples at the archive, how the format and the structure of the data sources diverge from one source to another.

But the Eprints system is only made for persons, not for applications to access. The mediator takes a step further to demonstrate this by providing a text based interface compared with the graphical Eprints interface.

**The output mediation application**

The output mediation application is a JAVA based web application (applet), which takes into account the data sets stored in the ROADIDEA data archive (Eprints). It translates
user requests into an XML language the system is speaking internally. These requests will be answered with a list of corresponding data sets and its main META elements.

Figure 13 show a flow chart of the implemented data output application. The main section of the flow chart describes the already implemented parts of the application, while processing and implementation of the dotted elements (automatic request generation and processing of data archive response) are prepared but not demonstrated with the online application.

![Flowchart for mediation output demonstrator](image)

The JAVA application was to be designed as intuitive as possible. For the given data output demonstration purpose user can select three kinds of attributes:

- Country,
- Data classification, and
- Keywords.

The country selection is mandatory. As long as no country (selection of one, several or all is possible) is selected, no further action can be performed. From the data class choice only one class can be selected. As for keywords, as many keywords can be typed in the ‘keywords’ text field as needed. In order to combine made selection and choices logically the ‘AND’ and ‘OR’ logics have to be used.
Figure 14 shows the output application. The application can be found and tested at http://www.roadidea.eu/packages/wp2/mediation.aspx. When at least one country is selected, the 'Create' button will be enabled. Pressing this button triggers the generation of a corresponding XML request file, which is shown in the text area of the application immediately. This request is well structured and it is including all made selections, choices, and keywords. (The very easy and intuitive structure has to be applied in order to enable user applications or services to create a data request automatically.)

The shown XML request can now be sent to the data archive by pressing the 'Send' button. A data archive search function is called and answered with a list of corresponding data sets from the ROADIDEA Eprints data archive. The list comprises the following META elements:

- Internal data set (Eprints) ID.
- Data set name,
- Abstract and
- Project partner contact information for accessing the data.

The 'Reset' button sets all selections on default and clears the response list in the text area and the program storage for a new request. In order to save a list from the text area it is recommended to copy the list and paste it into a text editor for later usage. For a response to an automatic request the applications sends the list as a structured XML file back to the calling application or function.

For available sections and actions (choices, selections and buttons) help context is available by pressing the '?' buttons.
5. Conclusions

With the given report the ROADIDEA data mediation architecture is described and finished. The applied Mediator has to be considered as a middleman between the ROADIDEA data archive and the outside world with its data sources and data users.

The WP2 data mediation application uses all previous results and knowledge from WP2. The result of this task demonstrates that it has been possible to create a very basic tool within the assignment of the investigation of how to merge different datasets together in the data mediation task. Even though the described mediation application uses historical data, the idea is that it should be possible to adapt the mediation approach on live data as well.

It has also been found that it is possible to combine data from different sources if they are provided in a well-structured format. For the mediation purpose, it was decided to use text-based data only. This standpoint limits this specific demonstration to not use other kind of data than text-based data, for example pictures.

One of the important conclusion and also recommendation from this work package is to have a definition of a common dataset with information close to traffic in Europe. The proposal is to have it as a call for a minimal dataset catalogue for ITS, Intelligent transportation system. This is necessary to have as a base for a better cooperation in the future within creating usable tools for a better information sharing and a safer traffic in Europe. The idea behind is that it is better to have a small selection of parameters available from a lot of places and sources in Europe, than having a lot of different sources separated from each other and difficult to use together.
References


