❒   6045

Transformation of WSDL files using ETL in the E-orientation domain

Adib Jihad[1], Moutachaouik Hicham[2], Marzak Abdelaziz[3], Hain Mustapha[4]
[1,3]Informatics Department, LTIM Laboratory, Faculty of Sciences Casablanca,
University Hassan II Casablanca, Morocco
[2,4]Industrial Engineering Department, ENSAM Casablanca,
University of Hassan II Casablanca, Morocco

Article Info	ABSTRACT
	The E-orientations platforms become an essential space for students to base their choice on tangible elements, as they offer a macroscopic view of the different academic and professional fields. In this context the Center for Engineering Sciences and Applied Sciences "SISA" funded the MMS orientation project to focus on student orientation in Morocco. With respect to the peculiarity and the different characteristics of the E-orientation platforms, a comparative study is essential. In this article we will proceed by a comparative study of a sample of platforms in order to highlight the major functionalities that we will model through descriptive files. The work is divided into two parts: The first part will be a comparison and description of existing platforms using descriptive language (WSDL), the second part will use ETL as a transformation technology in order to highlight generic files that will serve as a basis for work. The expected meta model. **

Corresponding Author:

Adib Jihad,
Informatics Department,
LTIM Laboratory, Faculty of Sciences,
Hassan II University of Casablanca,
Ben M'sik, Cdt Driss El Harti, BP 7955 Sidi Othman Casablanca, Morocco
Email: adib.jihad@gmail.com

1. INTRODUCTION

Choosing the right career can be difficult for students because they need to consider many elements in order to make the right choices that determine their academic and professional life. This need makes sense at the level of E-orientation systems [1], and for the sake of standardization of a generic platform meeting the different expectations of students. Our ultimate goal is to have an instantiable meta-model at the base of existing E-orientation platforms [2], using the model driven architecture (MDA) [3]. In order to reach our goal, a comparative study of these platforms proves to be indispensable because of their multiplicity, their variety and the approaches used for their implementation.

In this work we propose a comparative and descriptive study of the existing E-orientation platforms, first we generated a descriptive file for each feature of these platforms using the web services description language (WSDL), later we process these files using the ETL technology which is based on three steps: the first step is to target the data and apply the necessary filters on the data, the second step ensures several tasks that guarantee the reliability, the consolidation and the correction of the data, which eliminates any ambiguity in redundant data, and completes missing values, the latter is an indispensable step which is responsible for power and loading and data, to get at the end of the generic files that encompass all the attributes.

This article will be organized as follows: in section 2 we present the existing E-orientation platforms in the Moroccan and French Community. Section 3 we process the descriptive files using ETL technology. In section 4, we conclude the main contributions and the works future.

2. THE EXISTING E-ORIENTATION PLATFORMS IN THE MOROCCAN AND FRENCH COMMUNITY

After the baccalaureate, the priority of young students is mainly their university orientation. Several platforms provide information to help them make the right decision, but each of them has advantages and disadvantages, we want to compare them, first to collect common information and then propose a descriptive file for each feature of these platforms.

2.1. Description of existing E-orientation platforms

E-orientation platforms give high school students who wish to continue their higher education, to learn, to register, to make training vows, etc. they aim to satisfy a maximum of candidates to make the right choice, by allowing them to access the training that interests them. To do this we offer some examples:
- Orientation.com [4]: A yearbook of diplomas, studies and trades well distributed according to the level of studies, the site offers sections specific to the alternation, to the outlets, and proposes numerous practical sheets.
- 9rayti.com [5]: the leading platform for orientation and education in Morocco, followed by more than 3 million users per year. And it is edited by Education Media Company.
- Onisep.com [6]: is a public institution under the Ministry of National Education, Higher Education and Research. As a public publisher, the Onisep produces and publish information on training and trades. It also offers services to students, parents and educational teams.
- Orientation-chabab.com [7]: a guide for high school graduates, high school students and Moroccan students for access to universities and private and public higher schools in Morocco.

After describing the platforms, we now compare their general characterization, their strengths and their weak points.

2.2. Comparison of existing E-orientation platforms

The Table 1 illustrates the common features of the selected E-orientation platforms (orientation, 9rayti, Onisep, Orientation-chabab), and shows the advantages and disadvantages of each of them [8]. The next step is to generate a descriptive file for each feature.

Table 1. Comparison of E-orientation platforms

Features	Description	Orientation	9rayti	Onisep	Orientation-chabab
Etablissements	Description of all universities and schools.	Yes	Yes	Yes	Yes
Formations	Formation List.	Yes	Yes	Yes	Yes
Diplômes	The most detailed Diploma Directory.	Yes	Yes	No	Yes
Concours	The most detailed Competition directory for access to Large schools.	Yes	Yes	No	Yes
Métiers	Information on possible jobs.	Yes	Yes	Yes	No
Actualités	Education News and Events.	Yes	Yes	No	Yes
International	The complete Country Study Guide.	Yes	Yes	No	No
Emploi	Discovery of the professional world concerning internships in companies, alternations, my job search.	No	No	Yes	Yes
JPO	Open Days.	Yes	Yes	No	No
Handicap	Information on the studies of disabled people.	No	No	Yes	No
Collèges/lycées	Presentation of the studies at the college concerning the organization, or the classes, Internships, Diploma....	No	Yes	Yes	No
Parent	Explanations of schools to parents, answers to their questions.	No	No	Yes	No
………..	………………………….	………..	………..	………..	………..

2.3. Generation of descriptive files (in WSDL) for each feature

We chose the Axis platform [9] (integrated in Eclipse [10]). The latter has a Java plugin called Java2WSDL, which generates a WSDL [11] file for a Web service using a Java class. So we have developed all the features of the four aforementioned E-orientation platforms into multiple Java classes containing attributes and functions, and thanks to this tool feature WSDL files are generated automatically [12].

As an illustration, the Figure 1 represents the example of a WSDL file generating the Establishment feature for the 9rayti platform. Finally, we have obtained all the WSDL files of the selected E-orientations platforms, in the next chapter these files are processed using ETL technology.

Figure 1. Extract from the WSDL file of the Establishment functionality

3. PROCESSING WSDL FILES USING ETL

This chapter deals with the descriptive files of the E-orientations platforms, ie. instead of having several WSDL files of the same functionality we will have only one standard WSDL file enriched by all possible attributes. This processing will be established later for all the functionalities of the existing platforms.

3.1. ETL tools

The main goal of ETL (Extract, Transform, and Load) processes is to facilitate moving and transformation of data. ETL is a collection of programs which adds significant value to data [13]. The ETL process is composed of the following:

- Reformatting: put in a standard format different resources (files, records from database...etc)
- Conciliation: during extraction ETL should detect duplicate data a process a merge to eliminate inconsistency
- Cleaning: at the analysis step, the ETL process let only relevant data and delete non consistent ones

3.1.1. Design of the extraction programs

There are two main methods to extract data in ETL: the first one consists to make a workspace which is an image of the origin data source, which allows to data analyst to query data with possibility of recovery, but its disadvantage is the space needed to store backup data. The second method consist of divide data source into subset of data and give the possibility to work directly on origin data, this method has the advantage of having the same overview to all data analyst but multi-access to data should be well managed.

3.1.2. Design of the transformation programs

In ETL process the main part is the transformation, in which several issues should be addressed like:
- Erroneous identifiers
- Erroneous data
- Multiple data sources
- Irrelevant data
- Synonyms and homonyms
- Embedded process logic
- The main steps used in transformation are:
- Apply a standard to rename data in the same category
- Make a merge of the duplicate data to a relevant one
- Translate data in the same language for a later comparison

3.1.3. Design of the load programs

After the two previous steps Data is ready to be loaded, which is the final step in ETL process, at the stage data can be added in a database table via a script like SQL [14] or can be done with DBMS framework using some utility, however a special attention should be done on integrity and indexing [15]. The ETL process was traditionally developed by programmers, which mean a lot of hours of effort and the risk of major issues and also the reusability of the code is not guaranteed. With the apparition of ETL tools a lot of tasks has been eliminated, the tools offer a fluency in all steps of ETL process and available in accordance with requirements (technical and commercial ones) [16]. The Table 2 illustrates a comparative study of the features of the extraction tools, in particular Apatar OS Data Integration, CloverETL Engine, KETL, Pentaho, Scriptella, Talend.

Table 2. Comparison of functionalities between different ETL tools [17]

	Integration	Interfaces	Graphical editor	Function-ality	Support	Document-ation	Up-to-dateness
Apatar OS Data Integration		✓	✓	✓	✓	✓	✓
CloverETL Engine	✓	✓		✓	✓	✓	✓
Enhydra Octopus	✓	✓		✓			
Jitterbit Integration Environment		✓	✓	✓	✓	✓	✓
KETL	✓	✓		✓	✓	✓	✓
Pentaho Data Integration	✓	✓	✓	✓	✓	✓	✓
Scriptella	✓	✓		✓	✓	✓	✓
Talend Open Studio	✓	✓	✓	✓	✓	✓	✓

By the comparing some of ETL tools it is concluded that Pentaho [18] and Talend [19] are good enough then other tools nd have wide vriety of products. It is proved by MySQL and many of companies by their case studies that Pentaho can handle small to large scale systems [17]. In the next part we will present in more detail the product we chose to set up by describing its features, as mentioned before our choice was for an open source tool, and by studying the best known on the market we opted for Pentaho, we noticed that it is much easier to handle.

3.2. Processing WSDL file using Pentaho
3.2.1. Penatho

Pentaho data integration (PDI), long known as the Kettle, is an open source ETL that allows to design and implement handling and data transformation [20], as Pentaho Data Integration is able correspondingly to be implemented for additional commitments:
- Transferring data among programs or databases
- Distributing data from databases to flat files
- Filling data vastly into databases

- Data debugging
- Incorporating components [21].

  Incorporating components of Pentaho Data Integration is informal to implement. Each procedure is formed through a graphical instrument where users can determine what to do deprived of creating an algorithm to specify how to do it; due to, it might be said that Pentaho Data Integration is focused on metadata. Pentaho Data Integration is able to be implemented as a separate artefact, or it might be implemented together with the Pentaho Suite. As an Extract Transform and Load instrument, this is the most widespread open source application offered. It supports a massive collection of input and output presentations, comprising text les, data pages, and different sets of database engines. Furthermore, the modification competences of Pentaho Data integration permit to operate data by means of almost zero restrictions.

  This component is one of the best and most valued ETL solutions on the market [22]. Its long history, strength and robustness make it a highly recommended tool. It allows for transformations and works in a very simple and intuitive way. Similarly, data integration projects are very easy to manage. The Figure 2 represents the Pentaho data integration interface and Workflow; the ETL component permits renovations and function in a very easy and instinctive method.
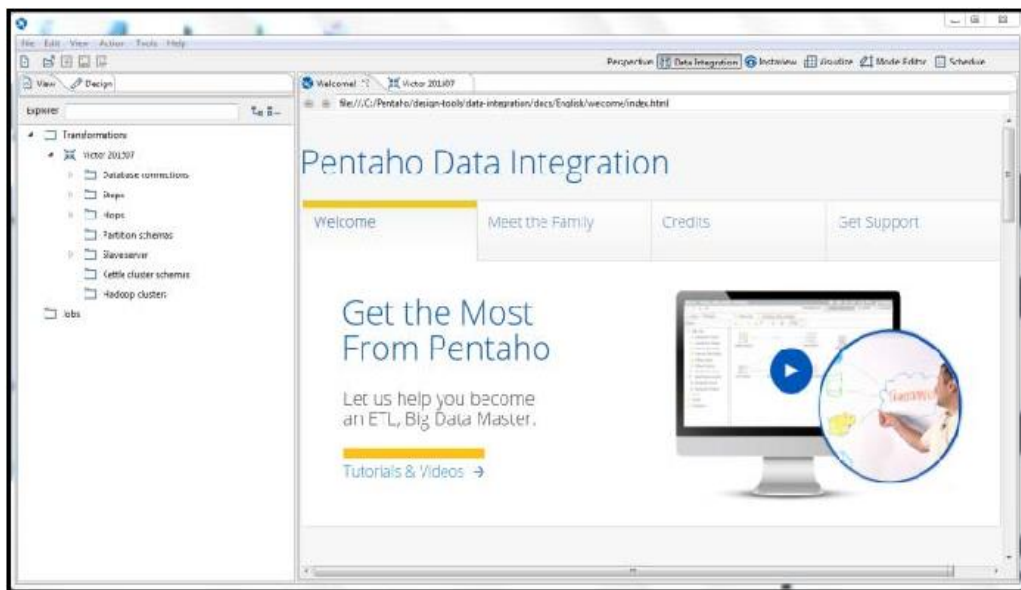

Figure2. Pentaho data integration interface and Workflow [23]

3.2.2. Design analysis

  The process of our work is unscrewed in two steps:
- The first step is to standardize the WSDL file for each feature of existing E-orientation platforms in order to respect a generic template in terms of structure and BOM. It will be iterated on each platform and for the same functionality to generate "standard" WSDL files.
- For the second step, the WSDL files of the same platform functionality are used as input, then the latter will be processed via XSLT [24] which contains the transformation rules mentioned below.

  At the end of the process we get a single standard WSDL file for this feature, and we set up the same operation for the rest of the features. Following this previous analysis, we will need to establish a dictionary to unite the names of the functionalities and attributes in order to avoid all the synonyms. Table 3 describes the management rules used throughout our treatment.

Table 2. Dictionary glossary

Former Functionality	New Functionality	New Attribute
Groupe des Etablissements au Maroc, Etablissements	Etablissements	Type_Ecole, Site_web, Statut, Description.
Guide des Métiers, Métiers	Métiers	Formation, Présentation.
International, Etudes Supérieures à l'étranger	International	-
Actualités&Evénement, Orientation	Actualités	-
Formations	Formations	Secteur

Transformation of WSDL files using ETL in the E-orientation domain (Adib Jihad)

Table 3. Management rules

Rule	Name	Description
R1	GEC	We keep the common elements between the functionalities of the platforms.
R2	AEC	We add elements that are not common (which belongs at least in a platform).
R3	AVE	The same value is assigned to elements that have the same meaning.
R4	GF	We keep all the functionality of the platforms.

3.2.3. Case study

In this section, we show through a case study, the transformation of WSDL files of E-orientation platforms via ETL. The extraction of the data, their transformations as well as their loadings is carried out as follows. The creation of a transformation with Pentaho Data Integration (spoon) according to three steps: the first step is reserved for extracting data from WSDL (XML-based) files of the same functionality, the second step one aims the XSL transformation from of an appropriate file containing the transformation rules mentioned above, then as for the last step, it consists of saving the result in a new WSDL document created on the basis of the contents of the XML file [25].

The phases of the ETL process are as follows:
- Extraction phase: As explained above, the first phase consists in acquiring scattered data. The tool pentaho has the ability to connect to WSDL files of the same functionality, it selectively reads the data of these files (it is the interest of the step), and thus filters the data in reading to extract only; relevant information.
- Transformation Phase: Data transformation is the main phase of our process. To process WSDL documents an XSL transformation is done. Our XSL file checks that data is consistent with existing data in WSDL files, assigns the same value for elements that have the same meaning, adds elements that are not common (which belongs at least in a platform).
- Integration phase or data feed: Once the processing ends, a phase of integration is made. The latter sends the resulting data to a WSDL document that is identical to the input data via a Pentaho Kettle data integration engine.

The Figure 3 illustrates the transformation of WSDL files retrieved from an existing E-orientation platform in order to standardize them. Here is an excerpt from the XSL file using the dictionary above to standardize a WSDL file of the same functionality as illustrated in Figure 4. The transformation of WSDL files of the same functionality can be done using this PDI. The Figure 5 represents an overview transformation design to use PDI-Kettle. Once the transformation is done, the WSDL file for this functionality is automatically generated.
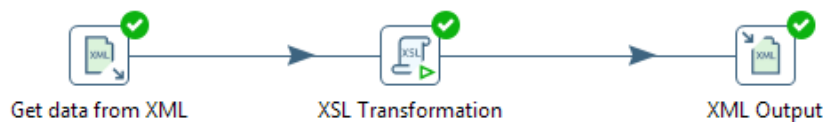

Figure 3. Transformation for the WSDL file of the Establishment functionality

```xml
<?xml version="1.0" encoding="UTF-8" ?>
<!-- l'élément racine -->
<xsl:stylesheet version="1.0" xmlns:xsl="http://www.w3.org/1999/XSL/Transform">
    <!-- l'élément output -->
    <xsl:output
        method="xml"
        encoding="UTF-8"
        indent="yes" ></xsl:output>
    <!-- reste du document XSLT -->
<xsl:template match="/">
    <xsl:copy>
        <!-- Copier tous les attributs -->
        <xsl:copy-of select="@*"/>
        <!-- Appliquer les règles à son contenu -->
        <xsl:apply-templates/>
    </xsl:copy>
</xsl:template>
        ...........................
        ...........................
        ...........................
        ...........................

</xsl:stylesheet>
```

Figure 4. Transformation for the WSDL file of the establishment functionality
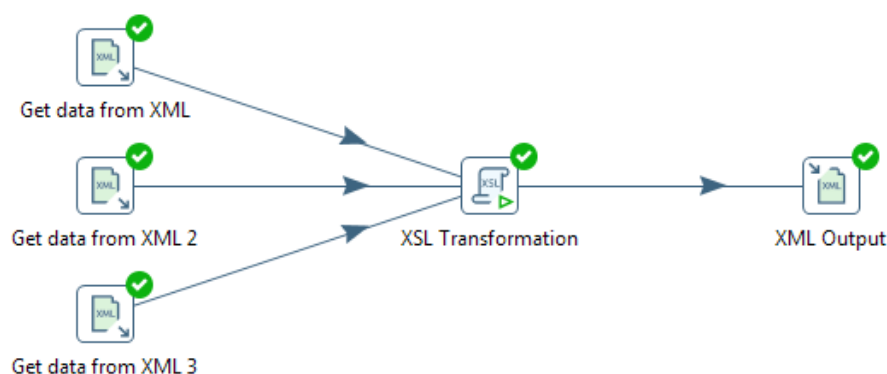
Figure 5. ETL transformation for the same functionality

4. CONCLUSION AND FUTUR WORK

In this article we have described and compared the existing E-orientation platforms in the Moroccan and French Community, Then we have generated for each feature a descriptive file WSDL using the platform axis (integrated in Eclipse), During this work, we presented both ETL solutions such as Apatar OS Data Integration, CloverETL Engine, KETL, Pentaho Data Integration, Scriptella, Talend Open Studio, then their features were compared between them, and to process our WSDL files we chose to work with the tool pentaho data integration, this processing is done in two steps: the first step is to make the WSDL files of each feature "standard", the second step is to process these files via XSLT. At the end of this work, we have been able to display generic files that will serve as a basis for the expected meta-model. In the next work we will model these WSDL files in UML and we will propose a general and enriched meta-model that brings together several functionalities of the E-orientation platforms using the model driven architecture (MDA).

REFERENCE

[1] F. Guerss, M. Aitdaoud, K. Douzi, M.Talbi, A. Namir "Implementation of a computerized system for the orientation of the Moroccan student in the university," *Procedia - Social and Behavioral Sciences,* vol. 182, pp. 381-387, May 2015.

[2] J. Adib, A. Marzak, H. Moutachaouik, M. Hain "The generation of the components meta-models of an E-orientation system using the WSDL-UML transformation," *The 2nd International Conference on. Smart Applications and Data Analysis for Smart Cities*, pp. 1-7, 2018.

[3] "MDA," *Object Management Group*. [Online]. Available: http://www.omg.org/mda/. [Accessed: 10-Feb- 2017].

[4] "Onisep," [Online]. Available: http://www.orientation.com/ [Accessed November 11, 2016].

[5] "Orientation," *digiSchool orientation*, [Online]. Available: http://www.orientation.com/ [Accessed: November 11, 2016].

[6] "9rayti," *Education Media Company*, [Online]. Available: http://www.9rayti.com/ [Accessed November 11, 2016].

[7] "Orientation-chabab," [Online]. Available: http://www.orientation-chabab.com/ [Accessed November 11, 2016].

[8] J. Adib, A. Marzak, H. Moutachaouik, M. Hain "Towards a Meta-model of E-orientation platforms," *IOSR Journal of Computer Engineering,* vol. 19, no. 4, Ver. IV, pp. 74-79, Jul-Aug 2017.

[9] "Apache Axis Architecture Guide," The Apache Software Foundation, 2002. [Online] Available: http://xml.apache.org/axis/

[10] "Open eclipse community," *Eclipse Foundation*, [Online]. Available: http://www.eclipse.org/. [Accessed November 11, 2016].

[11] "Web Services Description Language (WSDL)," *W3C Note*, 2012. [Online]. Available: http://www.w3.org/TR/wsdl.

[12] S. Rekkal, A. Bakhta, L. Lakhdar, "A new approach to gather similar operations extracted from web services," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, no. 2, pp. 1437-1444, Apr. 2019.

[13] S. Dupor and V. Jovanović, "An approach to conceptual modelling of ETL processes," *2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, Opatija, pp. 1485-1490, 2014.

[14] "SQL," SQL Cours et tutoriels sur le langage SQL, Accessed 10 Feb 2017. [Online]. Available: https://sql.sh

[15] M.Q. Memon, et al*.,* "Implementation of Business Intelligence for Sales Management," *International Journal of Electrical Engineering and Computer Science*, vol. 9, no. 3, pp. 784-788, Mar. 2018.

[16] "ETL Tools," *DB Software Laboratory*, [Online]. Available: https://www.etl-tools.com/. [Accessed: 10 Feb 2017].

[17] Tim A. Majchrzak, Tobias Jansen, Herbert Kuchen "Efficiency Evaluation of Open Source ETL Tools," *Proceedings of the 2011 ACM Symposium on Applied Computing (SAC)*, TaiChung, Taiwan, pp. 287-294, Mar. 2011.

[18] "Pentaho data integration," *Hitachi Vantara*, [Online]. Available: https://www.hitachivantara.com/en-us/products/big-data-integration-analytics/pentaho-data-integration.html/ [Accessed: 10-Feb- 2017].

[19] "Talend," Talend Data Fabric™, [Online]. Available: https://www.talend.com/. [Accessed 10 Feb 2017].

[20] N. Sirisha, K. V. D. Kiran, and R. Karthik, "Hadoop Security Challenges and Its Solution Using KNOX," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 12, no. 1, pp. 107-116, Oct. 2018.

[21] T. Kapila, T. "Pentaho BI & integration with a custom Java web application," [Online]. Available: http://www.neevtech.com/blog/2014/08/13/pentaho-bi-integration-with-a-custom-java-web-application-2/ [Accessed November 11, 2016].

[22] D. Tarnaveanu, "Pentaho Business Analytics: a Business Intelligence Open Source Alternative," *Database System Journal*, vol. 3, no. 3, pp. 23-34, 2012.

[23] M. Parra, A. Syed, M. Azeem, N. Halgamuge "Pentaho and Jaspersoft: A Comparative Study of Business Intelligence Open Source Tools Processing Big Data to Evaluate Performances," *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 10, pp. 20-29, 2016.

[24] "XSLT," *W3Schools*, [Online]. Available: https://www.w3schools.com/xml/xsl_intro.asp [Accessed: 10 Feb 2017].

[25] "Extensible Markup Language (XML)," *W3C Recommendation*, 2012. [Online]. Available: http://www.w3.org/XML/.