Data Pipeline Course – A22

Professor: Catherine Faron

Project Contributors:
Nken Babeth: nken.babeth@edu.dsti.institute
Nopchanok Duangdeeden: nopchanok.duangdeeden@edu.dsti.institute
Teresa Graffi: teresa.graffi@edu.dsti.institute

LANGUAGE HOLIDAY AGENCY PROJECT

Report on the modelling of an XML database which manages the data of a language holiday agency.

Modelling choices behind XML schema

INTRODUCTION

The XML database proposed is called "LANGUAGE-HOLIDAY-AGENCY", below is a description of the elements of the document. It is structured based on the region group to which each country belongs. We proposed 3 different countries and 2 cities for each country, for this reason we got a total of 6 offices, 12 different employees, 9 different offers, 18 different activities, and more than 18 different participants. The agency is divided into two groups, a European group with attribute `id="E"` and an Asian one with `group id="A".` Each country has two different offices based in 2 different cities, for France, the offices are in Paris and Nice, for Italy there are Rome and Bologna, while for Thailand the offices are in Bangkok and Phuket. Since Paris, Rome, and Bangkok are the capital cities of these countries they are also the main offices of the agency, for this reason, they propose two offers for the language holiday, while the other cities propose one offer. Below follows a representation of how our database is structured, the symbol `"(+)"` represents an element which has child-elements.

MAIN STRUCTURE OF THE DOCUMENT

```
<LANGUAGE-HOLIDAY-AGENCY>
        <group id='E'>
        <name> European </name>
                <country name='France'>
                        <office city='Paris'>
                        (+)     <location>
                                <n-employees> </n-employees>
                                <manager> </manager>
                        (+)     <employee> </employee>
                        (+)     <employee> </employee>
                        (+)     <offer> </offer>
                        (+)     <offer> </offer>
                        </office>
                        <office city='Nice'>
                                …
                        (+)     <offer> </offer>
                        </office>
                </country>
                <country name='Italy'>
                        <office city='Bologna'>
                                …
                        (+)     <offer> </offer>
                        </office>
                        <office city='Rome'>
                                …
                        (+)     <offer> </offer>
                        (+)     <offer> </offer>
                        </office>
                </country>
        </group>
        <group id='A'>
                <name> Asian </name>
                <country name='Thailand'>
                        <office city='Bangkok'>
                        …
                        (+)     <offer> </offer>
                        (+)     <offer> </offer>
                        </office>
                        <office city='Phuket'>
                                …
                        (+)     <offer> </offer>
                        </office>
                </country>
        </group>
</LANGUAGE-HOLIDAY-AGENCY>
```

For each office, the database contains information regarding the location of the workplace, employees' information, and more importantly, the different offers proposed by the office. The location is stored using three separate child elements, `postal-code`, `road-name`, and `building-number`.

For what concerns the information on the employees we stored some personal information like name and gender, and some data related to their work like email, address, office location, and language offer specialization.

The address of each one of the employees corresponds to the address of the office where they work, the office location corresponds to the city of the office, and the language of the offer specialization is the main language spoken in the country they work in. The email of the employees serves also as the identifier for each person that is stored in the database, since email addresses are unique.

OFFER ELEMENT

```
<offer>
        <id> </id>
(+)     <tour-leader>
(+)     <participant>
(+)     <participant>
(+)     <period season='Summer / Autumn / Winter / Spring'>
(+)     <language-class>
</offer>
```

The child element `offer` of office can be distinguished using an ID, which is made of 3 digits, the first digit stand for the country, 1 for France, 2 for Italy, 3 for Thailand, the second digit is the city, 1 for the first city described which is the capital city for each country and 2 for the second city, the third one indicates the number of the offer that can be either 1 or 2. There are other 5 child elements for each offer, `tour leader`, `participant`, `period`, and `language class`, each one of them contains other elements.

The tour leader must correspond to one of the two employees described in `office` element and is the same for each offer proposed by the same office. Information regarding the tour leader is stored and must correspond to the information stored in the `employee` element.

PARTICIPANT ELEMENT

```
<participant>
        <personal-info>
                <name> </name>
                <age> </age>
                <gender> </gender>
                <email> </email>
                <address> </address>
                <nationality> </nationality>
                <native-language> </native-language>
                <starting-proficiency-level> </starting-proficiency-level>
                <end-proficiency-level> </end-proficiency-level>
        </personal-info>
        <preferences>
                <group> alone, mixed ages, peers </group>
                <activity-main-type> sport / culture / culinary / relax </activity-main-type>
                <activity-main-type> sport / culture / culinary / relax </activity-main-type>
                <activity-secondary-type> individual / small group / big group </activity-secondary-type>
                <activity-secondary-type> individual / small group / big group </activity-secondary-type>
                <budget currency='$'> </budget>
                <duration> </duration>
                <season> Summer / Autumn / Winter / Spring </season>
                <start-date> </start-date>
                <end-date> </end-date>
                <accommodation> guest house / studio-apartment / college / not included </accommodation>
                <accommodation-board> half board, room and board, breakfast, not included </accommodation-board>
                <type-of-holiday> families holiday / school break / school group / preformed group / couples / friends / cultural exchange </type-of-holiday>
                <final-test> external certificate of proficiency / attendance certificate / proficiency certificate </final-test>
        </preferences>
</participant>
```

For each offer, there are at least 2 participants. The child element `participant` is more articulated than other elements. It contains two sub elements: `personal-info` and `preferences`. As suggested by the names, the first one stores personal information of the participant. Since the holiday is mainly focused on learning a language, `starting-proficiency-level` and `end-proficiency-level` are stored to assign participants to the right language class, and to register the proficiency level achieved at the end of the holiday.

The `preferences` element is the key to understanding the different offers proposed by the different offices of the agencies. It contains all the preferences expressed by the participant, and this information will be considered when describing the different aspects of the language holiday, indeed the whole offer depends to the season that the participant chooses as preference.

Among the preferences, every participant has the possibility to choose the kind of group to join during the holiday thanks to the `group` element; the choice is among `mixed ages`, `peers`, or `alone`, respectively join a group composed of people of mixed ages or joining a group with people with the same age of the participant or not joining a group at all but enjoying the holiday unaccompanied.

The agency organises also activities during the holiday, for this reason the participant can express up to two preferences for the type of activity he would like to attend among `sport`, `culture`, `culinary`, and `relax`, moreover, there is also the possibility to specify two different kinds of group to join during the activities between `small group`, `big group`, and `individually`.

Key elements that must be taken into consideration when organizing language holidays are how much the participant would like to pay for the holiday and how long his holiday will last. The `budget` is expressed in dollars using the attribute `currency='$'`, however, it must be in line with the duration of the trip, since a very long journey cannot have a small budget. For the same reason, the budget must agree as well with the type of accommodation and the type of board.

The `duration` of the holiday must be at least 1 week to allow the learning of the language and the carrying out of the activities, up to a maximum of 6 months. The participant will choose the `season` of the year; the choice is very important because the organization of the holiday is based on it since the activities will change accordingly to the season chosen.

The date of the trip can vary as well but the difference between `end-date` and `start-date` must coincide with the duration of the holiday and the month must belong to the season chosen.

The elements `accommodation` and `accommodation-board` are strictly related to each other and to the budget. There are four types of accommodation, some are more expensive than others, from the least expensive they are: `not included`, `guesthouse`, `college`, and `studio/apartment`. The same holds for the kinds of board, which are `not included`, `breakfast`, `half board`, and `room and board`.

Furthermore, the participant will specify the `type-of-holiday` preferred among several choices.

At the end of the language holiday the level reached will be tested and stored in the personal information of the participant in '`end-proficiency-level`'. The proficiency can be assessed in different ways and the participant can express his preference in the `final-test` element among `attendance certificate`, `proficiency certificate`, and `external certificate of proficiency`.

All the information described above is stored for each one of the participants of a certain offer proposed by the office of the language agency.

PERIOD ELEMENT

```
<period season='Summer/Autumn/Winter/Spring'>
            <start-date> </start-date>
            <end-date> </end-date>
            <activity>
                        <name> </name>
                        <occurrence> </occurrence>
                        <main-type> </main-type>
                        <secondary-type> </secondary-type>
                        <n-participants> </n-participants>
                        <participant>
                                    <name> </name>
                                    <age> </age>
                                    <gender> </gender>
                                    <email> </email>
                        </participant>
                        <participant>
                                    …
                        </participant>
            </activity>
            <activity>
                        …
            </activity>
            <staff>
                        <name> </name>
                        <gender> </gender>
                        <email> </email>
                        <address> </address>
                        <language> </language>
                        <type> </type>
            </staff>
</period>
```

Another child element of "`offer`" is the period, which has the attribute "`season`". The attribute must correspond to the season specified by the participants in the preferences and according to it, the activities will vary. The first child elements of period element are `start-date` and `end-date`, which will correspond to those expressed in the preferences of the participants.

One of the key child elements is `activity`, there can be more than one activity for each period, and for each activity, some information will be stored such as the name of the activity, how many times it will take place, the type of activity among relax, culture, sport and culinary, the kind of group of participants that will join, the total number of participants and of course who are the participants. The `participant` must correspond to the participant of the offer element but for each participant, just the strictly necessary information is saved, i.e., `name`, `age`, `gender`, and `email`. All this information is stored for each activity but of course, the participants of the activities can vary, but still must respect the preferences of the participants.

The last key child element of period is `staff;` here one of the employees stored in office will appear again with a different child element, which is `type`. The type changes according to the activity that has been organized for a certain period.

LANGUAGE-CLASS ELEMENT

```
<language-class>
            <id> F1 </id>
            <n-students> </n-students>
            <language> </language>
            <proficiency> </proficiency>
            <final-test> </final-test>
            <final-test> </final-test>
            <professor>
                        <name> </name>
                        <gender> </gender>
                        <email> </email>
                        <address> </address>
                        <language> </language>
                        <proficiency> </proficiency>
            </professor>
            <participant>
                        <name> </name>
                        <age> </age>
                        <gender> </gender>
                        <email> </email>
                        <nationality> </nationality>
                        <native-language> </native-language>
            </participant>
            <participant>
                        …
            </participant>
</language-class>
```

The offer of course includes a language-class as well, which is the last child element of the offer element. The `language-class` element has some sub elements that aim at describing and differentiating each one of the language courses offered by the agency, always respecting the preferences of the participants. Each class has an `id` which is built using the first letter of the language taught in the class and the number of the starting level of proficiency. Then there is the number of students, the language taught, the starting and ending proficiency, and two possible types of the final test to assess the proficiency of the students at the end of the course.

Of course, the professor is one of the child elements of language class as well; the `professor` element includes some sub elements used to describe information about the professor. The professor must correspond to one of the two employees of the office, the one that wasn't chosen to be part of the staff of the activities.

The last child element of the language class is the `participant`, there can be more than one participant for each language class.

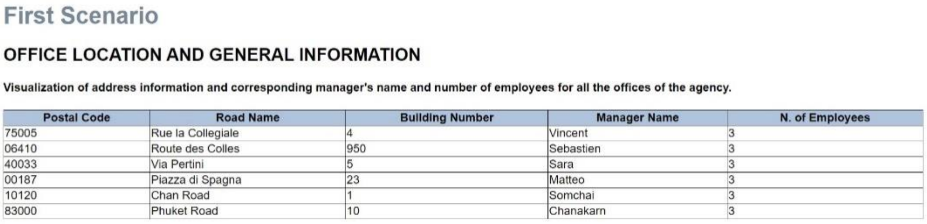
Advantages and Disadvantages of the modelling for data processing

Advantages:
The structure based on the location of the agency makes it very easy to identify a specific country. Moreover, in case the agency would open a new agency, either in one of the countries proposed or in another country, the inclusion of this new element won't cause any problem, since the structure will not need to be modified.
To add a new participant to the database the process is very simple, a new node must be created under the 'offer' element, afterwards the participant can be added in all the preferred activities and language classes.
Additional activities can be added in any moments adding a new node under the season element; it's not even mandatory to add a participant to the activity, for this reason new activities can be created and used for future holidays as well.

Disadvantages:
A disadvantage of the database is that there isn't an element 'clients' that stores data of all the participants who have joined a holiday in the past or that are going to participate in the future. However this could be implemented easily as a sub element of 'office', in this way the structure won't change but additional data can be stored.
The cost of the holiday is uncertain because it depends on many factors, among these, the season, length of stay, accommodation type and board. All these elements depend on the preferences of each participant and on the budget availability.

Scenarios to answer use cases

1. In the first case we imagined a situation in which is required the data concerning the offices' information, including location and general information like the name of the manager of each office and how many people are employed.



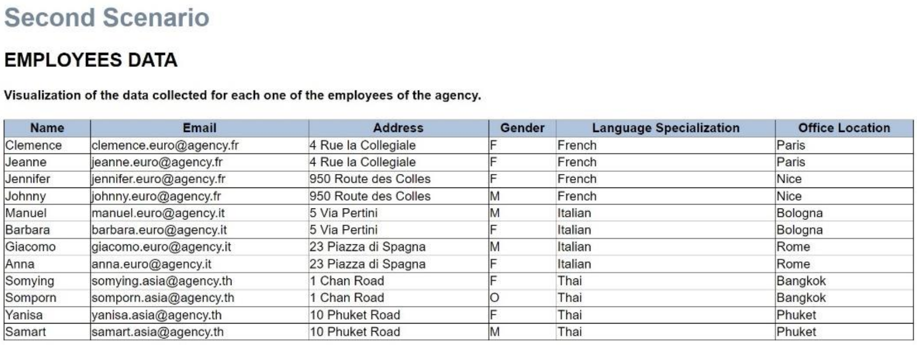
First Scenario

OFFICE LOCATION AND GENERAL INFORMATION

Visualization of address information and corresponding manager's name and number of employees for all the offices of the agency.

Postal Code	Road Name	Building Number	Manager Name	N. of Employees
75005	Rue la Collegiale	4	Vincent	3
06410	Route des Colles	950	Sebastien	3
40033	Via Pertini	5	Sara	3
00187	Piazza di Spagna	23	Matteo	3
10120	Chan Road	1	Somchai	3
83000	Phuket Road	10	Chanakarn	3

2. The second scenario satisfies the need of visualizing a sort of database that contains data on the employees. This kind of database is needed for all workplaces and makes it easy to access information on the employees like email or in which office they are currently working.



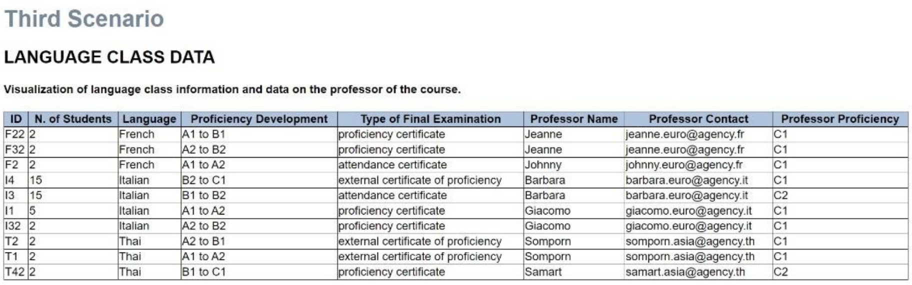
Second Scenario

EMPLOYEES DATA

Visualization of the data collected for each one of the employees of the agency.

Name	Email	Address	Gender	Language Specialization	Office Location
Clemence	clemence.euro@agency.fr	4 Rue la Collegiale	F	French	Paris
Jeanne	jeanne.euro@agency.fr	4 Rue la Collegiale	F	French	Paris
Jennifer	jennifer.euro@agency.fr	950 Route des Colles	F	French	Nice
Johnny	johnny.euro@agency.fr	950 Route des Colles	M	French	Nice
Manuel	manuel.euro@agency.it	5 Via Pertini	M	Italian	Bologna
Barbara	barbara.euro@agency.it	5 Via Pertini	F	Italian	Bologna
Giacomo	giacomo.euro@agency.it	23 Piazza di Spagna	M	Italian	Rome
Anna	anna.euro@agency.it	23 Piazza di Spagna	F	Italian	Rome
Somying	somying.asia@agency.th	1 Chan Road	F	Thai	Bangkok
Somporn	somporn.asia@agency.th	1 Chan Road	O	Thai	Bangkok
Yanisa	yanisa.asia@agency.th	10 Phuket Road	F	Thai	Phuket
Samart	samart.asia@agency.th	10 Phuket Road	M	Thai	Phuket

3. The third scenario contains data regarding the language classes that have been organised by the agency. This visualization is very useful in the organisation phase of the classes; indeed, each professor can easily identify which course they will be covering and the corresponding number of students.



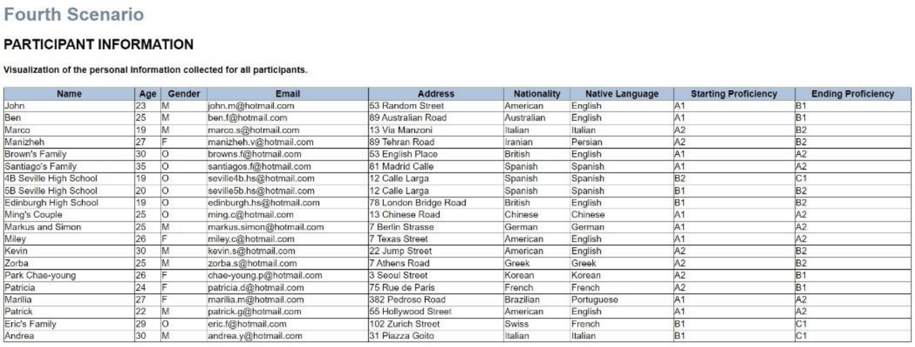
Third Scenario

LANGUAGE CLASS DATA

Visualization of language class information and data on the professor of the course.

ID	N. of Students	Language	Proficiency Development	Type of Final Examination	Professor Name	Professor Contact	Professor Proficiency
F22	2	French	A1 to B1	proficiency certificate	Jeanne	jeanne.euro@agency.fr	C1
F32	2	French	A2 to B2	proficiency certificate	Jeanne	jeanne.euro@agency.fr	C1
F2	2	French	A1 to A2	attendance certificate	Johnny	johnny.euro@agency.fr	C1
I4	15	Italian	B2 to C1	external certificate of proficiency	Barbara	barbara.euro@agency.it	C1
I3	15	Italian	B1 to B2	attendance certificate	Barbara	barbara.euro@agency.it	C2
I1	5	Italian	A1 to A2	proficiency certificate	Giacomo	giacomo.euro@agency.it	C1
I32	2	Italian	A2 to B2	proficiency certificate	Giacomo	giacomo.euro@agency.it	C1
T2	2	Thai	A2 to B1	external certificate of proficiency	Somporn	somporn.asia@agency.th	C1
T1	2	Thai	A1 to A2	external certificate of proficiency	Somporn	somporn.asia@agency.th	C1
T42	2	Thai	B1 to C1	proficiency certificate	Samart	samart.asia@agency.th	C2

4. The fourth visualization shows a kind of database containing personal information of the participants, like a client database. Thanks to it the agency can observe the contact details of the participants and the progress that they did thanks to the language holiday organised.



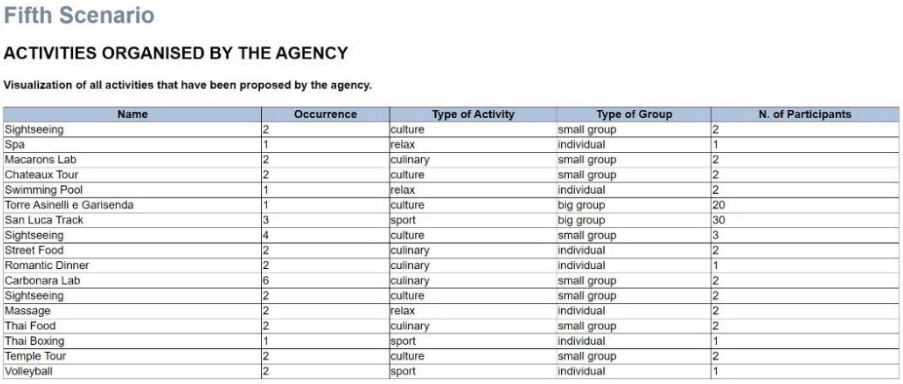
Fourth Scenario

PARTICIPANT INFORMATION

Visualization of the personal information collected for all participants.

Name	Age	Gender	Email	Address	Nationality	Native Language	Starting Proficiency	Ending Proficiency
John	23	M	john.m@hotmail.com	53 Random Street	American	English	A1	B1
Ben	25	M	ben.f@hotmail.com	89 Australian Road	Australian	English	A1	B1
Marco	19	M	marco.s@hotmail.com	13 Via Manzoni	Italian	Italian	A2	B2
Manizheh	27	F	manizheh.v@hotmail.com	89 Tehran Road	Iranian	Persian	A2	B2
Brown's Family	30	O	browns.f@hotmail.com	53 English Place	British	English	A1	A2
Santiago's Family	35	O	santiagos.f@hotmail.com	81 Madrid Calle	Spanish	Spanish	A1	A2
4B Seville High School	19	O	seville4b.hs@hotmail.com	12 Calle Larga	Spanish	Spanish	B2	C1
5B Seville High School	20	O	seville5b.hs@hotmail.com	12 Calle Larga	Spanish	Spanish	B1	B2
Edinburgh High School	19	O	edinburgh.hs@hotmail.com	78 London Bridge Road	British	English	B1	B2
Ming's Couple	25	O	ming.c@hotmail.com	13 Chinese Road	Chinese	Chinese	A1	A2
Markus and Simon	25	M	markus.simon@hotmail.com	7 Berlin Strasse	German	German	A1	A2
Miley	26	F	miley.c@hotmail.com	7 Texas Street	American	English	A1	A2
Kevin	30	M	kevin.s@hotmail.com	22 Jump Street	American	English	A2	B2
Zorba	25	M	zorba.s@hotmail.com	7 Athens Road	Greek	Greek	A2	B2
Park Chae-young	26	F	chae-young.p@hotmail.com	3 Seoul Street	Korean	Korean	A2	B1
Patricia	24	F	patricia.d@hotmail.com	75 Rue de Paris	French	French	A2	B1
Marilia	27	F	marilia.m@hotmail.com	382 Pedroso Road	Brazilian	Portuguese	A1	A2
Patrick	22	M	patrick.g@hotmail.com	55 Hollywood Street	American	English	A1	A2
Eric's Family	29	O	eric.f@hotmail.com	102 Zurich Street	Swiss	French	B1	C1
Andrea	30	M	andrea.y@hotmail.com	31 Piazza Goito	Italian	Italian	B1	C1

5. The fifth scenario shows all the activities that the agency offers. This kind of visualisation can be to show to new clients which kind of activities the agency can organise during a language holiday.



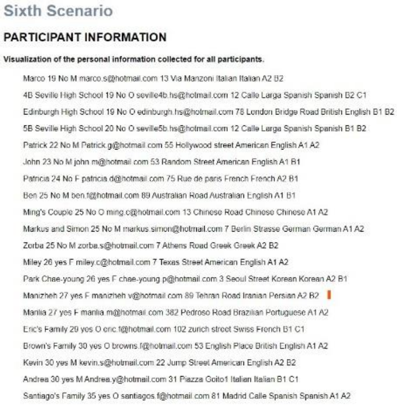
Fifth Scenario

ACTIVITIES ORGANISED BY THE AGENCY

Visualization of all activities that have been proposed by the agency.

Name	Occurrence	Type of Activity	Type of Group	N. of Participants
Sightseeing	2	culture	small group	2
Spa	1	relax	individual	1
Macarons Lab	2	culinary	small group	2
Chateaux Tour	2	culture	small group	2
Swimming Pool	1	relax	individual	2
Torre Asinelli e Garisenda	1	culture	big group	20
San Luca Track	3	sport	big group	30
Sightseeing	4	culture	small group	3
Street Food	2	culinary	individual	2
Romantic Dinner	2	culinary	individual	1
Carbonara Lab	6	culinary	small group	2
Sightseeing	2	culture	small group	2
Massage	2	relax	individual	2
Thai Food	2	culinary	small group	2
Thai Boxing	1	sport	individual	1
Temple Tour	2	culture	small group	2
Volleyball	2	sport	individual	1

6. The sixth scenario shows the exploitation of part of the personal data of the participants in namespaces format. Data is showed in ascending order of age, testing when the age of the participant exceeds 25 ("yes" for age > 25, "No" otherwise). This can be used by the agency in those circumstances where employees need to check the age of participants to offer them suitable activities and to create the groups for the language holiday.



Sixth Scenario

PARTICIPANT INFORMATION

Visualization of the personal information collected for all participants.

Marco 19 No M marco.s@hotmail.com 13 Via Manzoni Italian Italian A2 B2

4B Seville High School 19 No O seville4b.hs@hotmail.com 12 Calle Larga Spanish Spanish B2 C1

Edinburgh High School 19 No O edinburgh.hs@hotmail.com 78 London Bridge Road British English B1 B2

5B Seville High School 20 No O seville5b.hs@hotmail.com 12 Calle Larga Spanish Spanish B1 B2

Patrick 22 No M Patrick.g@hotmail.com 55 Hollywood street American English A1 A2

John 23 No M john.m@hotmail.com 53 Random Street American English A1 B1

Patricia 24 No F patricia.d@hotmail.com 75 Rue de paris French French A2 B1

Ben 25 No M ben.f@hotmail.com 89 Australian Road Australian English A1 B1

Ming's Couple 25 No O ming.c@hotmail.com 13 Chinese Road Chinese Chinese A1 A2

Markus and Simon 25 No M markus.simon@hotmail.com 7 Berlin Strasse German German A1 A2

Zorba 25 No M zorba.s@hotmail.com 7 Athens Road Greek Greek A2 B2

Miley 26 yes F miley.c@hotmail.com 7 Texas Street American English A1 A2

Park Chae-young 26 yes F chae-young.p@hotmail.com 3 Seoul Street Korean Korean A2 B1

Manizheh 27 yes F manizheh.v@hotmail.com 89 Tehran Road Iranian Persian A2 B2

Marilia 27 yes F marilia.m@hotmail.com 382 Pedroso Road Brazilian Portuguese A1 A2

Eric's Family 29 yes O eric.f@hotmail.com 102 zurich street Swiss French B1 C1

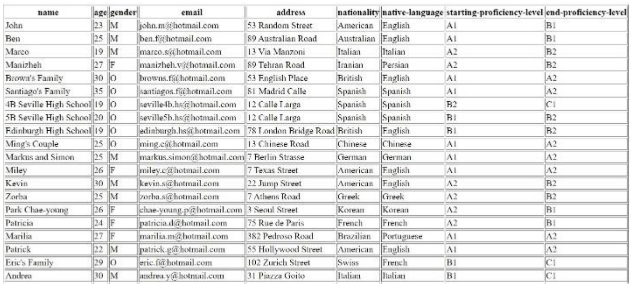
Brown's Family 30 yes O browns.f@hotmail.com 53 English Place British English A1 A2

Kevin 30 yes M kevin.s@hotmail.com 22 Jump Street American English A2 B2

Andrea 30 yes M Andrea.y@hotmail.com 31 Piazza Goito1 Italian Italian B1 C1

Santiago's Family 35 yes O santiagos.f@hotmail.com 81 Madrid Calle Spanish Spanish A1 A2

7. The seventh scenario shows the exploitation of part of the personal information collected on the participants in JSON.



name	age	gender	email	address	nationality	native-language	starting-proficiency-level	end-proficiency-level
John	23	M	john.m@hotmail.com	53 Random Street	American	English	A1	B1
Ben	25	M	ben.f@hotmail.com	89 Australian Road	Australian	English	A1	B1
Marco	19	M	marco.s@hotmail.com	13 Via Manzoni	Italian	Italian	A2	B2
Manizheh	27	F	manizheh.v@hotmail.com	89 Tehran Road	Iranian	Persian	A2	B2
Brown's Family	30	O	browns.f@hotmail.com	53 English Place	British	English	A1	A2
Santiago's Family	35	O	santiagos.f@hotmail.com	81 Madrid Calle	Spanish	Spanish	A1	A2
4B Seville High School	19	O	seville4b.hs@hotmail.com	12 Calle Larga	Spanish	Spanish	B2	C1
5B Seville High School	20	O	seville5b.hs@hotmail.com	12 Calle Larga	Spanish	Spanish	B1	B2
Edinburgh High School	19	O	edinburgh.hs@hotmail.com	78 London Bridge Road	British	English	B1	B2
Ming's Couple	25	O	ming.c@hotmail.com	13 Chinese Road	Chinese	Chinese	A1	A2
Markus and Simon	25	M	markus.simon@hotmail.com	7 Berlin Strasse	German	German	A1	A2
Miley	26	F	miley.c@hotmail.com	7 Texas Street	American	English	A1	A2
Kevin	30	M	kevin.s@hotmail.com	22 Jump Street	American	English	A2	B2
Zorba	25	M	zorba.s@hotmail.com	7 Athens Road	Greek	Greek	A2	B2
Park Chae-young	26	F	chae-young.p@hotmail.com	3 Seoul Street	Korean	Korean	A2	B1
Patricia	24	F	patricia.d@hotmail.com	75 Rue de Paris	French	French	A2	B1
Marilia	27	F	marilia.m@hotmail.com	382 Pedroso Road	Brazilian	Portuguese	A1	A2
Patrick	22	M	patrick.g@hotmail.com	55 Hollywood Street	American	English	A1	A2
Eric's Family	29	O	eric.f@hotmail.com	102 Zurich Street	Swiss	French	B1	C1
Andrea	30	M	andrea.y@hotmail.com	31 Piazza Goito	Italian	Italian	B1	C1

Appendix: tools used during the project

Working environment:

- Notepad++
- Visual Studio Code

Check XML syntax:

- Notepad++ with XML Tools plugins
- Visual Studio Code with XML extension

Validate XML Schema against the XML document:

- Notepad++ with XML Tools plugins
- Visual Studio Code with XML extension

Convert XSL stylesheets into HTML:

- https://www.freeformatter.com/xsl-transformer.html#before-output

Convert JSON to HTML:

- https://codebeautify.org/json-to-html-converter#google_vignette