Institut für Maschinelle Sprachverarbeitung |
Summer Semester 24

LINGUISTIC ROADMAP OF DARIJA

Helping Darija Linguists Grow

COURSE:
Text Technology

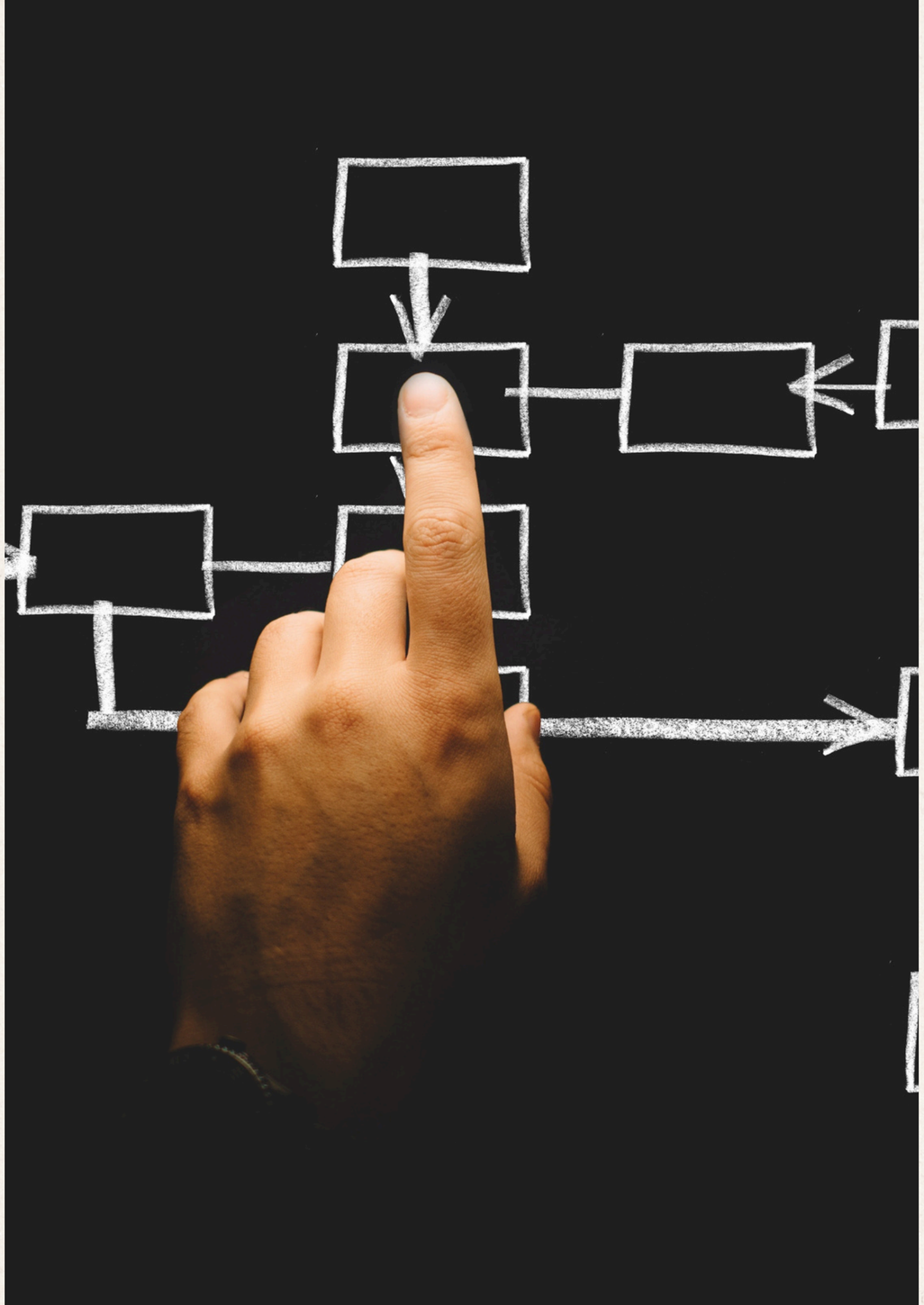
STUDENTS:
Yassir EL ATTAR & Shawn Chua

PROFESSOR:
Kerstin Jung

Agenda

Overview

Where did this idea come from?

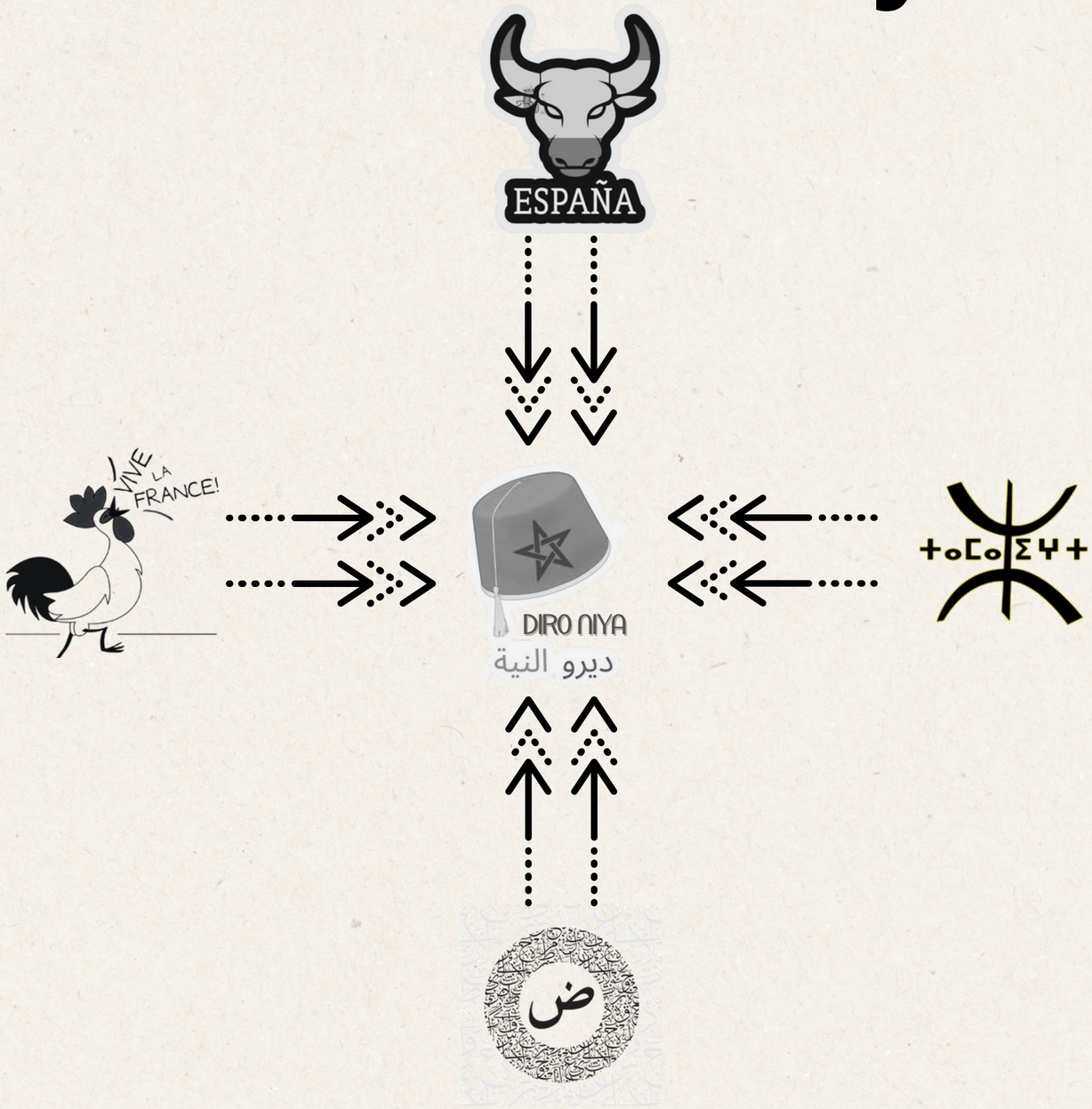


01 Discussion of our native languages and their variations

02 Darija is a new and developping language in Morocco

03 Many linguistics aspects of Darija are still uncovered

Why Darija?



What resulted?

Arabic	*Latin*	*Darija_ Ar*	*Darija*
ق	-	ق	9/ q
ح	-	ح	7
ع	-	ع	3
ص	-	ص	S
ض	-	ض	D
ط	-	ط	T
-	g	ڤ	g
-	p	پ	p
-	v	ف	v

Objectives and Goals

By the end of our project, we will create a platform to:

Goal # 1

Be an interactive database for linguists interested in Darija

Goal # 2

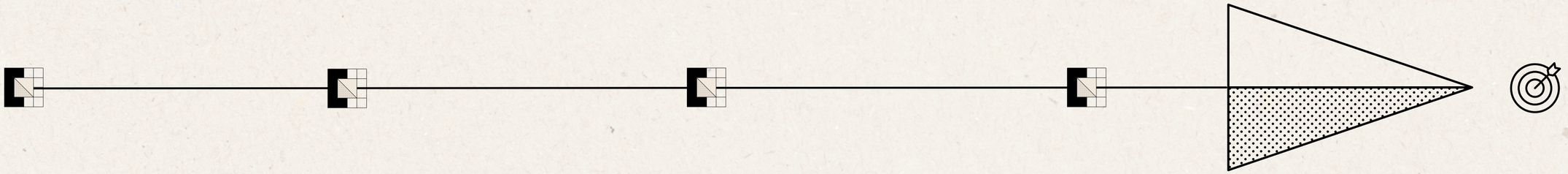
Improve the current Dataset and solve the common problems

Goal # 3

Help linguists contribute and access the database

Timeline

Project timeline



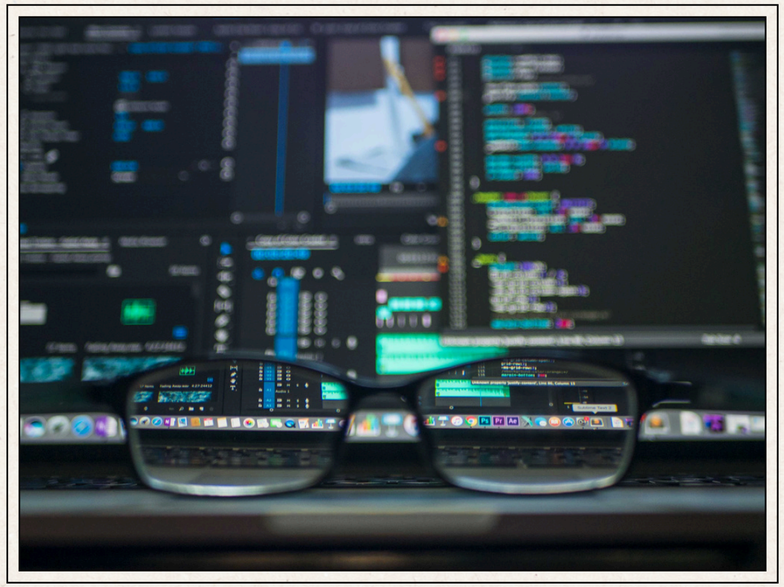
1st period
Collecting data

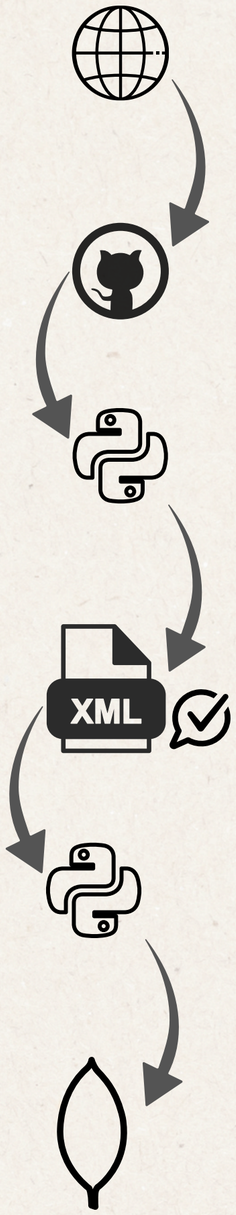
2nd period
Cleaning & storing data

3rd Period
Building technologies

4th Period
Accessing & testing the platform

Final Period
Publishing

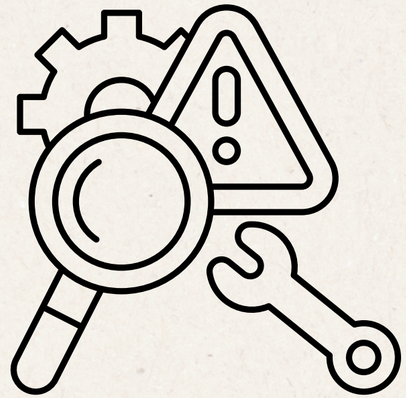




Collect

- From online search to
- Github Darija Open Dataset (DODa)
- Python code to convert to XML
- Validation of XMLs with XSDs
- Python to convert from XML to MongoDB

Problems encountered



The original dataset still had some issues to deal with

Issue #1

Not all entries in the dataset have the same number of spelling forms

Issue #2

Missing values for entries (English/and Darija_Ar)

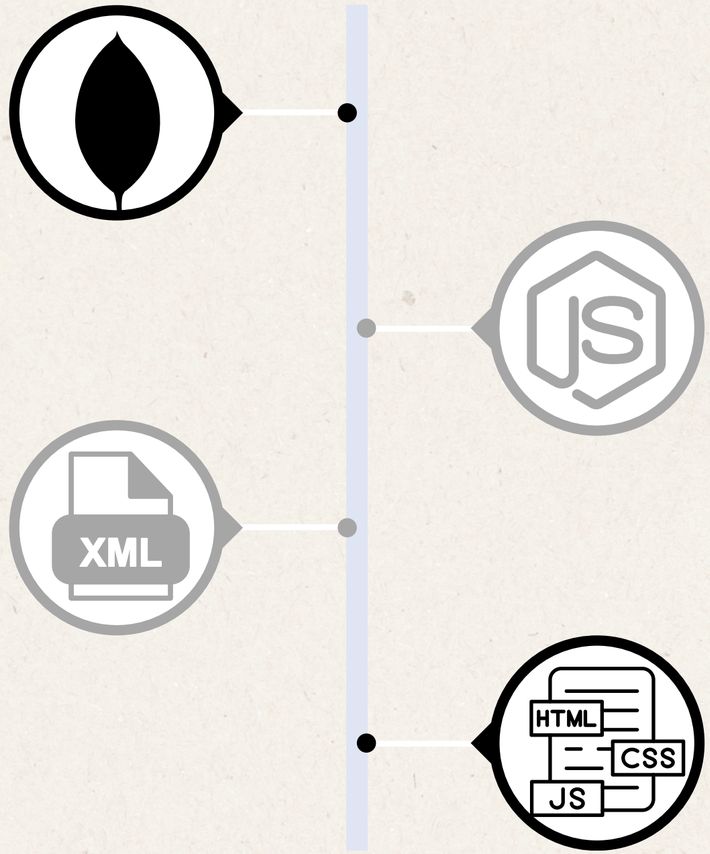
issue #3

Some records/entries appeared in multiple collections/tables

	n1	n2	n3	n4	darija_ar	eng
2	mousi9a	musiqa	mosi9a		موسيقا	music
3	rasm				رسم	painting
4	kamanja				كامانجا	violin
5	3oud				عود	oud
6	chTi7	chTih			شطيح	dance
7	ghna	ghona			غنا	singing
8	chi3r				شعر	poetry
9	fenn	fann	fnn		فنّ	art
10	msr7	masra7	masrah	mesre7	مسرح	theater

Prepare

- MongoDB and Nodejs interaction using Expressjs Framework
- Creating necessary queries of NoSQL
- XSDs to validate XML files for each collection
- Converting XML files to JSON using JS and add them to database
- HTML page to display & access the data



Problems encountered

we encountered several issues when preparing the platform, such as:

Issue #1

Creating XSDs to validate XML files according to the norms of the data structure

Issue #2

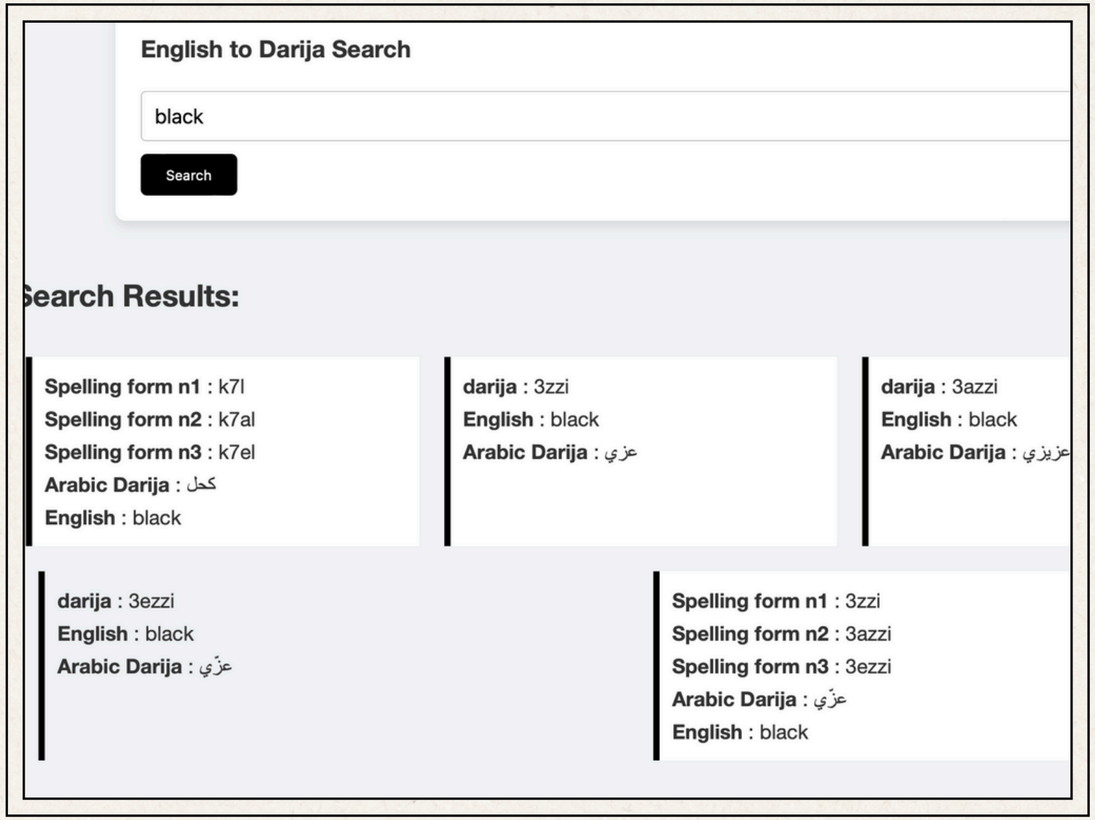
Connecting MongoDB to NodeJS using ExpressJS Server

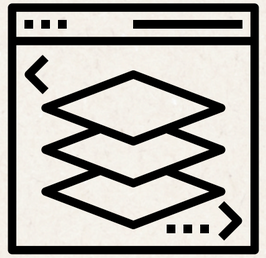
Issue #3

Dublicate entries and records in database cause incorrect results

Issue #4

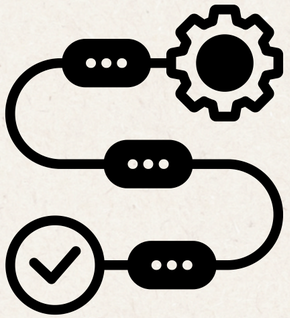
Problems with using JS to validate XML files with XSDs





Access

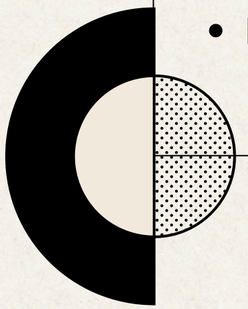
- HTML Page to access data
- Contribution using XML files
- Validation of XML files
- XML to MongoDB
- Statistics of the database



Extensions

ADDITIONAL TECHNOLOGIES AND TECHNIQUES USED

• **XSD additional options**	*Validate XML element to select from options*
• **Javascript**	*Javascript manipulation of JSON*
• **NodeJS & ExpressJS**	*Easy interactions between MongoDB (noSQL) & FrontEnd Page (HTML)*
• **NoSQL**	*More detailed noSQL queries for specific tasks on MongoDB*



Thank you

CONTACT US

E-mail {yassir.el.attar,shawn.chua}@ims.uni-stuttgart.de

Program Computational Linguistics M.Sc.

Department IMS - Faculty 5

References

DATABASE & CONTENT

Darija Open Dataset. [Online] Available at: https://darija-open-dataset.github.io/

GitHub - darija-open-dataset/dataset. [Online] Available at: https://github.com/darija-open-dataset/dataset/ .

MongoDB. [Online] Available at: https://www.mongodb.com .

IMAGES & ICONS

Icons - Available at: https://www.canva.com

Spain Symbol - Available at: https://www.pinterest.com/pin/spain-spanish-flag-souvenir-bull-head-sticker--1089378597347561283/

France Symbol - Available at: https://qph.cf2.quoracdn.net/main-qimg-967efe8c5aae711f6c6d7cddf5bace2c-lq

Tamazight Symbol - Available at: https://tunisie-amazigh.blogspot.com/2013/12/blog-post_26.html

Arabic Symbol - Available at: https://www.nawa3em.com/UserFiles/002-daddd.jpg

Darija ‘Diro Niya’ Symbol - Available at: https://www.spreadshirt.fr/shop/design/diro+niya+morocco+qatar+world+cup+avocay+autocollant-D63adb9d93664df7481c8091c?sellable=qrjDBE5oEMHqQp5wykXn-1459-215