Transnational Statistical Judicial Datasets

Acquisition and Integration

Davide Rua Carneiro

Polythecnic of Porto

Portugal

Grant Agreement n. 101007420
This project is funded by the EU's Justice Programme (2014-2020)

Transnational Statistical Judicial Datasets

- Refer to datasets that contain statistical information and data related to judicial systems and legal processes across different countries or jurisdictions
- Aim to provide comparative and cross-national insights into various aspects of the judicial system, such as case outcomes, court proceedings, legal disputes, and legal trends
- The datasets may include information on
  - Case filings, case dispositions, case durations, legal disputes, court workload, legal personnel, legal aid, legal reforms and legislation…
- Can be valuable for
  - Comparative legal research, policy analysis, and understanding cross-country variations in legal systems
  - Can help identify patterns, trends, and challenges in the administration of justice
  - Facilitate evidence-based policymaking, and support efforts to improve access to justice and legal systems globally
- But can be ethically and legally problematic because:
  - Potential infringement on fundamental rights (privacy)
  - Concerns for intrusiveness, undue digital nudging, and digital surveillance of users
  - Interference with the professional activities of justice officers through digital surveillance

Co-funded by the Erasmus+ Programme of the European Union

Transnational Statistical Judicial Datasets

- In the context of the ODR Scheme project, we elaborated a document that summarizes the proposed approach to acquire and integrate transnational judicial data
- The document contains three main parts
  - Data Extraction – describes the automated process through which data elements and structures can be identified and extracted from existing ODR processes
  - Data Structures – proposes a series of data elements and structures, developed in the context of the project, and their foreseen purposes
  - Data Integration – details an approach for integrating data structures of similar scope but of different origins into unified trans-national judicial datasets

Co-funded by the Erasmus+ Programme of the European Union

EUI ROBERT SCHUMAN CENTRE

1. Data Extraction

How to extract data automatically from ODR processes

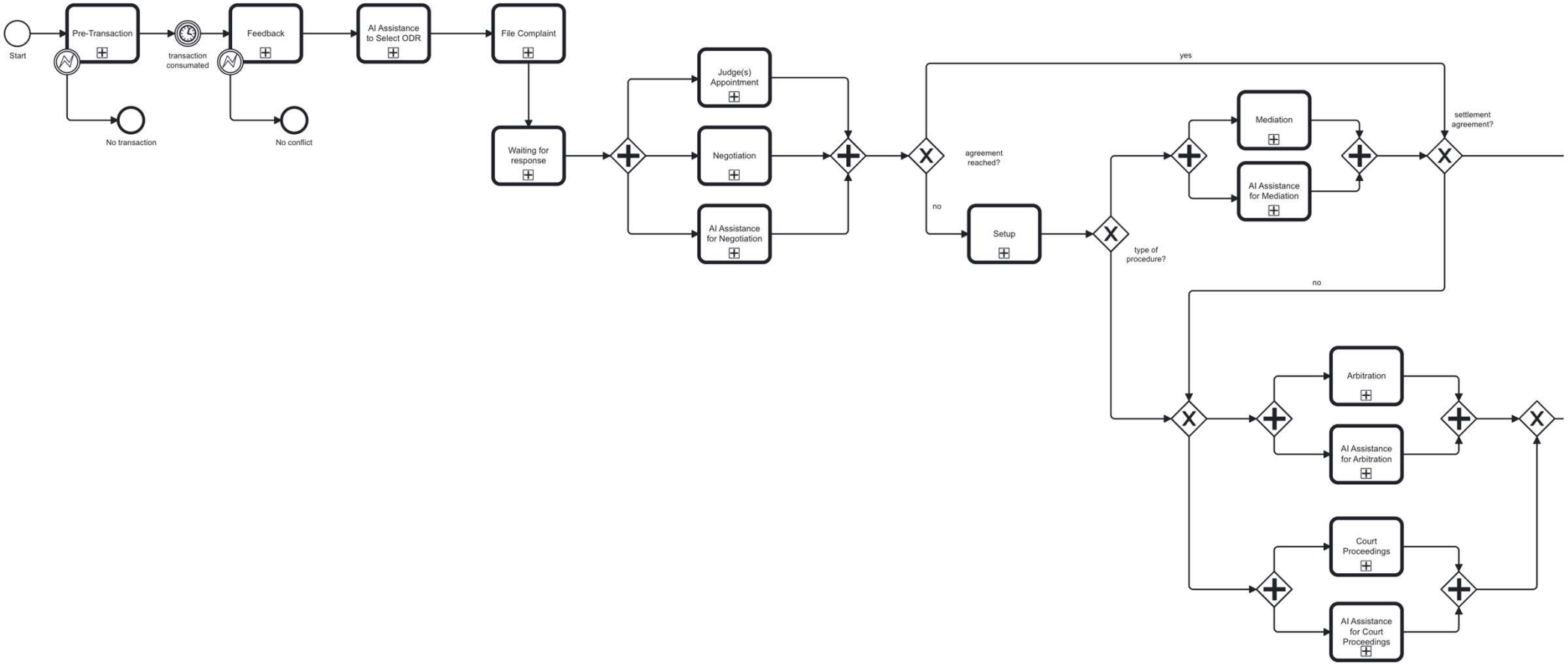
Co-funded by the Erasmus+ Programme of the European Union

ODR Processes

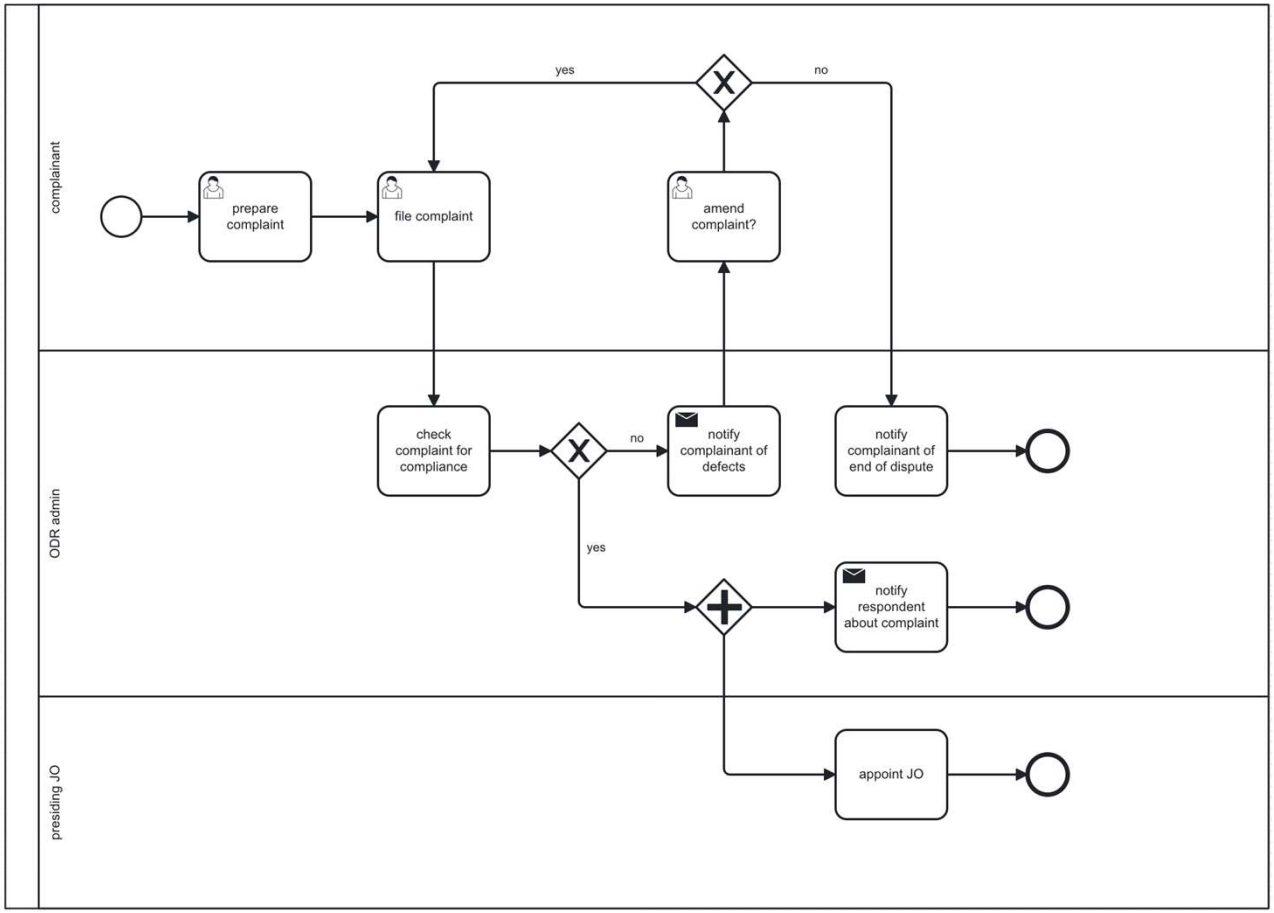
- We assume the existence of digitalized ODR processes, implemented using a standard notation (e.g. BPMN)
- Business Process Model and Notation
  - A graphical modeling language to represent business processes visually
  - Provides a standardized notation that can be easily understood by business users, developers, and stakeholders
  - Designed to provide a common language and framework for describing, analyzing, and improving business processes
  - It consists of a set of symbols and conventions for representing activities, events, gateways, flows, and artifacts

Co-funded by the Erasmus+ Programme of the European Union

BPMN – high level view (excerpt)



BPMN – high level view (File Complaint)



Co-funded by the Erasmus+ Programme of the European Union

BPMN – Forms / variables

Complaint

Are you making the complaint on behalf of somebody else?*

Select

In the name of a consumer or a trader?*

Select

Are you a consumer or a trader?*

Select

Complainant's details

Please enter the complainant's personal contact details in the form. This will make it easier for the respondent to identify them and for the dispute resolution body to contact them (if the complainant and the respondent agree to use the dispute resolution body to resolve their case).

First Name*

Last Name*

Organization Name*

Phone Number

BPMN – Underlying structure


```
{
  "components": [
    {
      "text": "# Complaint",
      "type": "text",
      "id": "Field_1urdpco"
    },
    {
      "values": [
        {
          "label": "Yes",
          "value": "yes"
        },
        {
          "label": "No",
          "value": "no"
        }
      ],
      "label": "Are you making the complaint on behalf of somebody else?",
      "type": "select",
      "id": "Field_1ta5z5u",
      "key": "complaint_on_behalf_of_somebody_else",
      "validate": {
        "required": true
      }
    },
    {
      "values": [
        {
          "label": "Consumer",
          "value": "consumer"
        },
        {
          "label": "Trader",
          "value": "trader"
        }
      ],
      "label": "In the name of a consumer or a trader?",
      "type": "select",
      "id": "Field_0cmah8a",
      "key": "complaint_in_the_name_consumer_or_trader",
      "conditional": {
        "hide": "=complaint_on_behalf_of_somebody_else!=\"yes\""
      },
      "validate": {
        "required": true
      }
    },
```

BPMN – Underlying structure

```
</bpmn:userTask>
<bpmn:exclusiveGateway id="Gateway_0f2k92g">
  <bpmn:incoming>Flow_06dvgwb</bpmn:incoming>
  <bpmn:outgoing>Flow_059iaas</bpmn:outgoing>
  <bpmn:outgoing>Flow_0yepp0l</bpmn:outgoing>
</bpmn:exclusiveGateway>
<bpmn:userTask id="Activity_15s313y" name="Appoint JO">
  <bpmn:incoming>Flow_0lzwjp2</bpmn:incoming>
  <bpmn:outgoing>Flow_0fclfnv</bpmn:outgoing>
</bpmn:userTask>
<bpmn:userTask id="Activity_1og7h5v" name="Set up case file (Presiding Judge)">
  <bpmn:incoming>Flow_0fclfnv</bpmn:incoming>
  <bpmn:outgoing>Flow_1pr87xq</bpmn:outgoing>
</bpmn:userTask>
<bpmn:sequenceFlow id="Flow_10kzgcu" name="yes" sourceRef="Gateway_0h5wrkm" targetRef="Gateway_0bkt7tg">
  <bpmn:conditionExpression xsi:type="bpmn:tFormalExpression">=agreement_JO_yes_no = "yes"</bpmn:conditionExpression>
</bpmn:sequenceFlow>
<bpmn:sequenceFlow id="Flow_0xoxzl3" name="no" sourceRef="Gateway_0h5wrkm" targetRef="Activity_0k8cklt">
  <bpmn:conditionExpression xsi:type="bpmn:tFormalExpression">=agreement_JO_yes_no = "no"</bpmn:conditionExpression>
</bpmn:sequenceFlow>
<bpmn:sequenceFlow id="Flow_0wbd5nw" sourceRef="Activity_0k8cklt" targetRef="Gateway_1q6ziyl" />
<bpmn:sequenceFlow id="Flow_13jem2d" sourceRef="Gateway_1q6ziyl" targetRef="Activity_00b4ajj">
  <bpmn:conditionExpression xsi:type="bpmn:tFormalExpression">=select_proceeding = "mediation"</bpmn:conditionExpression>
</bpmn:sequenceFlow>
<bpmn:sequenceFlow id="Flow_0ro4z3l" sourceRef="Gateway_1q6ziyl" targetRef="Activity_0hjn15w">
  <bpmn:conditionExpression xsi:type="bpmn:tFormalExpression">=select_proceeding = "online_hearing"</bpmn:conditionExpression>
</bpmn:sequenceFlow>
```


Co-funded by the Erasmus+ Programme of the European Union

Extracting Data Elements

- We propose to automatically extract data elements from the XML and JSON specifications

	form_id	zeebe_user_task_form	activity_name	element_id	level_1	level_2	type
0	Form_a581d9b9-a0cc-48d3-89fc-219b37065949	userTaskForm_0dd00cs	AI Negotiation service 1	ri_language	1	1	select
1	Form_a581d9b9-a0cc-48d3-89fc-219b37065949	userTaskForm_0783ces	Negotiation service 1	ri_language	2	1	select
2	form_select_proceeding	userTaskForm_33isn7p		select_proceeding_select	3	1	select
3	Form_agreement_yes_no	userTaskForm_1m2a768		agreement_date_yes_no	4	1	select
4	Form_agreement_yes_no	userTaskForm_3v97i01		settlement_yes_no	5	1	select
5	Form_hearing_yes_no	userTaskForm_3bp8i11		hearing_yes_no	6	1	select
6	Form_d1edad6c-						
7	Form_d1edad6c-	values	labels	required	disabled	condition	
8	Form_d1edad6c-	['en', 'pt']	['English', 'Portuguese']	FALSO	FALSO		
		['en', 'pt']	['English', 'Portuguese']	FALSO	FALSO		
		['online_hearing', 'mediation']	['Online hearing', 'Mediation']	FALSO	FALSO		
		['yes', 'no']	['Yes', 'No']	FALSO	FALSO		
		['yes', 'no']	['Yes', 'No']	FALSO	FALSO		
		['yes', 'no']	['Yes', 'No']	FALSO	FALSO		
		['personal', 'business']	['Personal Account', 'Business Account']	FALSO	FALSO		
		['yes', 'no']	['Yes', 'No']	FALSO	FALSO	"=trader_already_a_member != \"yes\""	
		[]	[]	FALSO	FALSO	"=trader_already_a_member != \"yes\""	

Co-funded by the Erasmus+ Programme of the European Union

Extracting Data Elements

- Form ID – a reference to the unique identifier form where a given data element exists
- Activity name – the name of the BPMN activity in which this data element is defined
- Element ID – the unique identifier of the data element
- Numbering – a hierarchical numbering system that is created and allows to uniquely identify each data element, and its position in the hierarchy
- Type – the type of data element (e.g. select, text field, check box, radio button, date picker, …)
- Values – the possible values of the data element (when applicable, such as in single- or multiple-selection controls)
- Required – whether this element is mandatory in the process or not
- Disabled – whether this element is originally disabled or visible (some elements might only be visible depending on certain conditions being met)
- Condition – when applicable, what are the conditions that define the visibility of a given element
- Default value – the default value of an element, when applicable
- Min / Max – the minimum and maximum values of this element, when applicable, for validation purposes

Co-funded by the Erasmus+ Programme of the European Union

EUI ROBERT SCHUMAN CENTRE

2. Data Structures

What can be built with the extracted Data Elements

Co-funded by the Erasmus+ Programme of the European Union

Data Structures

- The Data Elements extracted from the BPMN process can be combined and/or processed into Data Structures, at different levels
  - Level 0, Raw Data - Data structures at this level use data elements that are directly extracted from the BPMN process
  - Level 1, Processed Data - Includes data elements that are obtained through simple data processing pipelines (e.g. sums, differences, extraction of components…), that use one or more raw data sources. Its interpretation and its relationship with the raw data is straightforward
  - Level 2, Derived/aggregated Data - Includes data elements that can only be achieved through more complex data processing operations, eventually evolving multiple elements from lower levels. These elements should provide support for higher-level decision making, and its relationship with the raw data may not be so direct for someone unfamiliar with the data collection/processing pipeline

Co-funded by the Erasmus+ Programme of the European Union

Access to Justice

A. Preferences concerning language(s) for communication
B. Preferences concerning communication means (e.g. e-mail, mobile phone, platform-specific tools)
C. Preferences concerning communication modalities (e.g. audio, video, text)
D. Preferences concerning locations/regions for participating in the proceedings
E. Preferences concerning conflict resolution modalities (e.g. negotiation, mediation)
F. Information regarding maximum costs that parties are willing to support ;
G. Information regarding cultural background

Co-funded by the Erasmus+ Programme of the European Union

Conflict Description

A. The nature of the parties (e.g. consumer, vendor, …)
B. The location of the parties (e.g. country, region)
C. The nature/type of conflict (e.g. general consumer goods or services, healthcare, financial services, …)
D. The issue(s) at stake (e.g. defect, damage caused, delivery, billing, …)
E. Initial proposal for resolution by the complainant
F. Whether there has already been contact between the parties and/or a tentative to solve the conflict
G. Whether any of the parties are bound to a specific dispute resolution body

Co-funded by the Erasmus+ Programme of the European Union

Conflict Resolution Process

A. A characterization of the process (e.g. procedures used, existence of legal representative, AI services used, language(s) used, …)
B. How the conflict was solved (e.g. mediation, arbitration)
C. Whether a settlement agreement was reached
D. Time to resolution
E. Success rate
F. Satisfaction level of the parties
G. Costs

Co-funded by the Erasmus+ Programme of the European Union

Performance of AI techniques

A. A characterization of the automated mean used in a given process (e.g. generative language model, solution generation, summarization tools, case management)
B. Indicators of accuracy of the automated mean (e.g. accuracy, precision, recall, perplexity, F1 Score, BLEU)
C. A characterization of how the accuracy of the automated mean was evaluated (e.g. methodology, characterization of the data, …)
D. Measures of efficiency (e.g. time it takes for the automated mean to provide responses/results)
E. Measures of user satisfaction towards the use of the automated mean
F. Measures of consistency (i.e. measuring the variability of the responses of the automated mean across similar problems)
G. Measures of legal compliance (i.e. measuring to which extent the responses of the automated mean comply with legal regulations and guidelines)
H. A description of the initiatives in place for monitoring the automated mean by the service provide, including for the detection of bias

EUI ROBERT SCHUMAN CENTRE

3. Data Integration

How to integrate datasets of different origins

Co-funded by the Erasmus+ Programme of the European Union

Data Integration

- This section proposes a process through which datasets of different origins but describing equivalent realities can be integrated, especially in what concerns transnational judicial data
- It relies on
  - Metadata of each source of data (in English) so that the relationship between elements of different data structures can be established
  - A set of integration rules, which specify how the fields of different data sources should be integrated into a single, unified, dataset

Co-funded by the Erasmus+ Programme of the European Union

Integration Rules

- Each integration rule, stored in a machine-readable format, is composed of the following information:
  - Data source origin – the file, connection string (if database), or similar resource to the origin of one data source;
  - Origin field – the name of the input field, from the origin data source;
  - Target database – the target database, in which data are being integrated;
  - Target field – the name of the resulting field in the target database;
  - Transformation – the rules for integrating the source field into the target database;

Co-funded by the Erasmus+ Programme of the European Union

Integration Rules

```
{
 "_id": {
  "$oid": "64428e52bf103590fe492e56"
 },
 "origin_uri": "mongodb+srv://user:password@cluster0.1srjw7j.mongodb.net",
 "dest_uri": "mongodb+srv://user:password@cluster0.1srjj9j3.mongodb.net",
 "origin_field": "custo",
 "dest_field": "cost",
 "origin_collection": "portugal"
 "dest_collection": "final",
 "transformation":
 [
   {
       $project: -> select the relevant field
       {
           _id: 1,
           custo: 1,
       },
   },
   {
       $addFields: -> add new field “cost”, that results from replacing € in the original field “custo”
       {
           cost:
           {
               $replaceAll:
               {
                   input: "$custo",
                   find: "€",
                   replacement: ""
               },
           },
       },
   },
   {
       $project: -> do not include the original field in the final dataset
       {
           custo: 0
       },
   },
   {
       $out: "final" -> write the result of the pipeline in the “final” dataset
   },
 ],
}
```


Co-funded by the Erasmus+ Programme of the European Union

EUI ROBERT SCHUMAN CENTRE

Data Integration

- A visual editor will be developed to allow national authorities (e.g. experts from ministries of justice, experts from statistics bureaus) who want to create transnational datasets to create their own integration from available sources
- Integration rules are created in a structured format and using a Database query language, so that they can run automatically

Co-funded by the Erasmus+ Programme of the European Union

Conclusions

- Bridging transnational datasets is challenging
  - Different laws
  - Different concepts
  - Different technologies
  - …
- It is also a multi-dimensional problem
  - Legal
  - Technical
- This is a proposal created in the scope of the project, which needs to be validated by experts from different fields
  - We count on your feedback!

Co-funded by the Erasmus+ Programme of the European Union