



NLP techniques on Public Procurement textual data

Challenges and Opportunities

María Navas-Loro, Óscar Corcho
Ontology Engineering Group
Universidad Politécnica de Madrid, Spain



mnavas@fi.upm.es



<https://marianavas.linkeddata.es>



2022-10-21



NextProcurement



- Public procurement accounts for a 14% of the annual budget of the different governments of the European Union.
- Nevertheless, there is room for improvement regarding public procurement documents processing.



- European Project
2022-2024 (3 years)



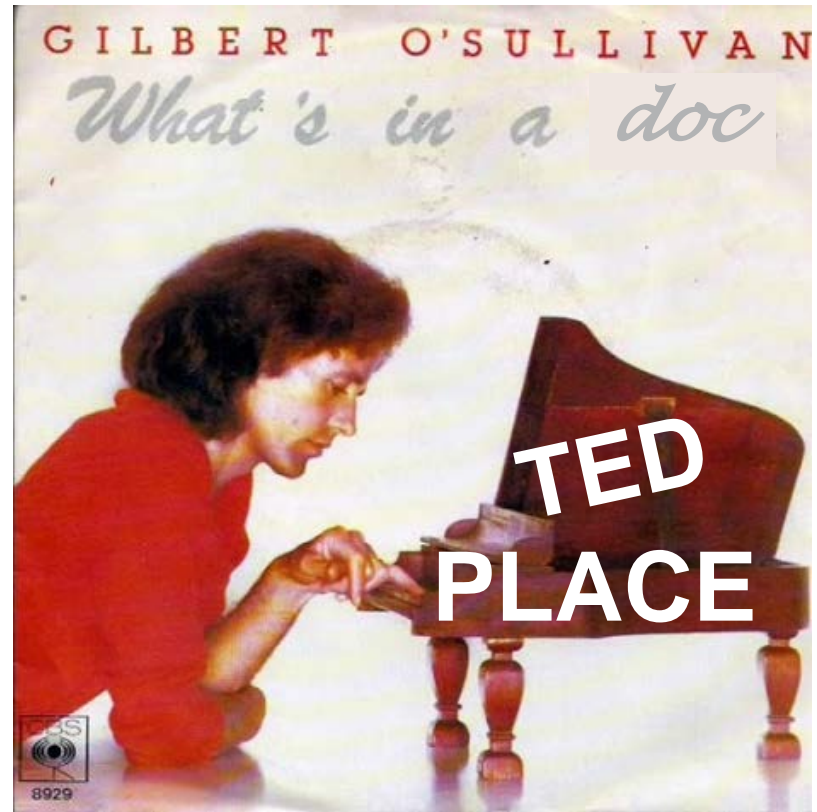
- Currently extracting
requisites (so more fun!):
 - Metadata is a “chorizo” of NEs
 - OCR of awful scanned documents
 - Entity linking
 - CIF
 - Company names
 - ...



- Additionally, we (Oscar, David Chaves) are working with the EU on Public Procurement Data Spaces (an European Platform for Public Procurement)

What's in a doc?

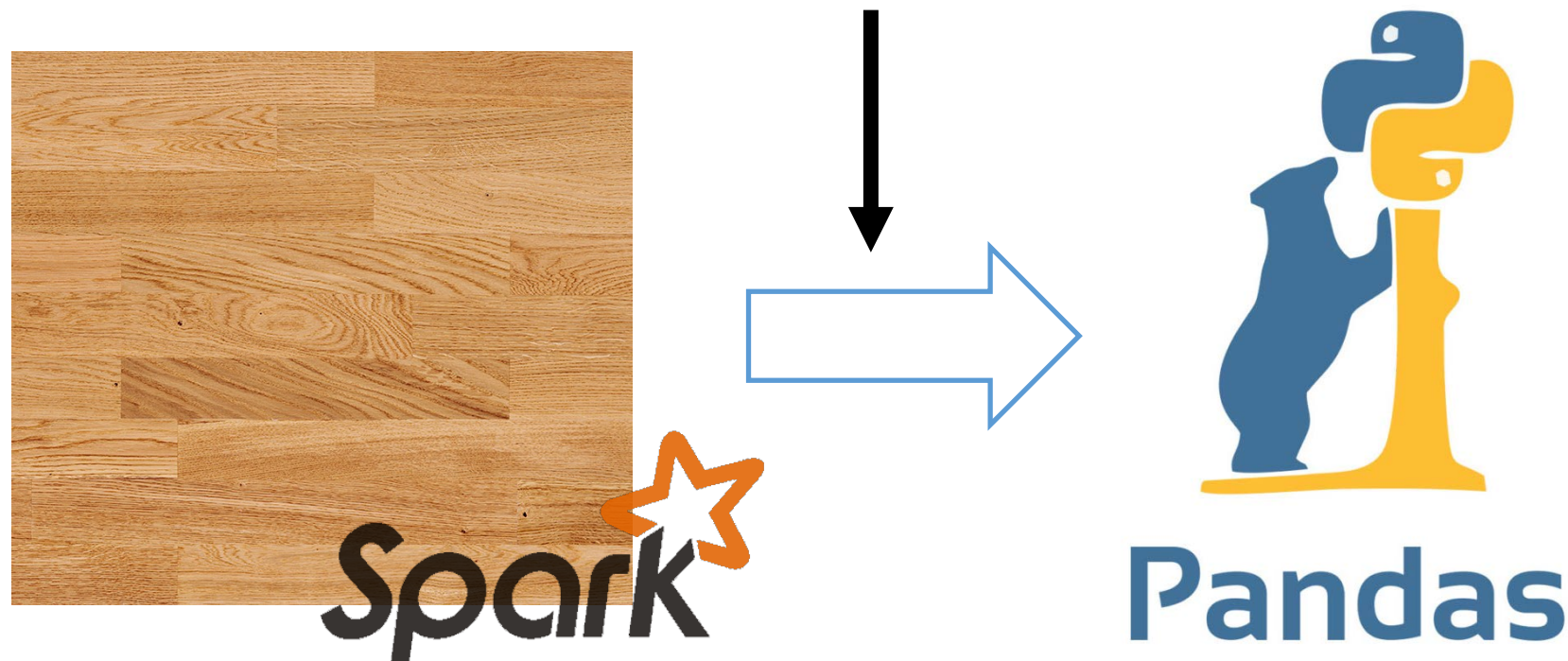
TENDERS



Data in .parquet from UC3M (easy to read in Python*)

```
import pandas as pd

df = pd.read_parquet('file.parquet', engine='fastparquet')
print(df)
```



* Also working on Morph-KGC mappings 😊

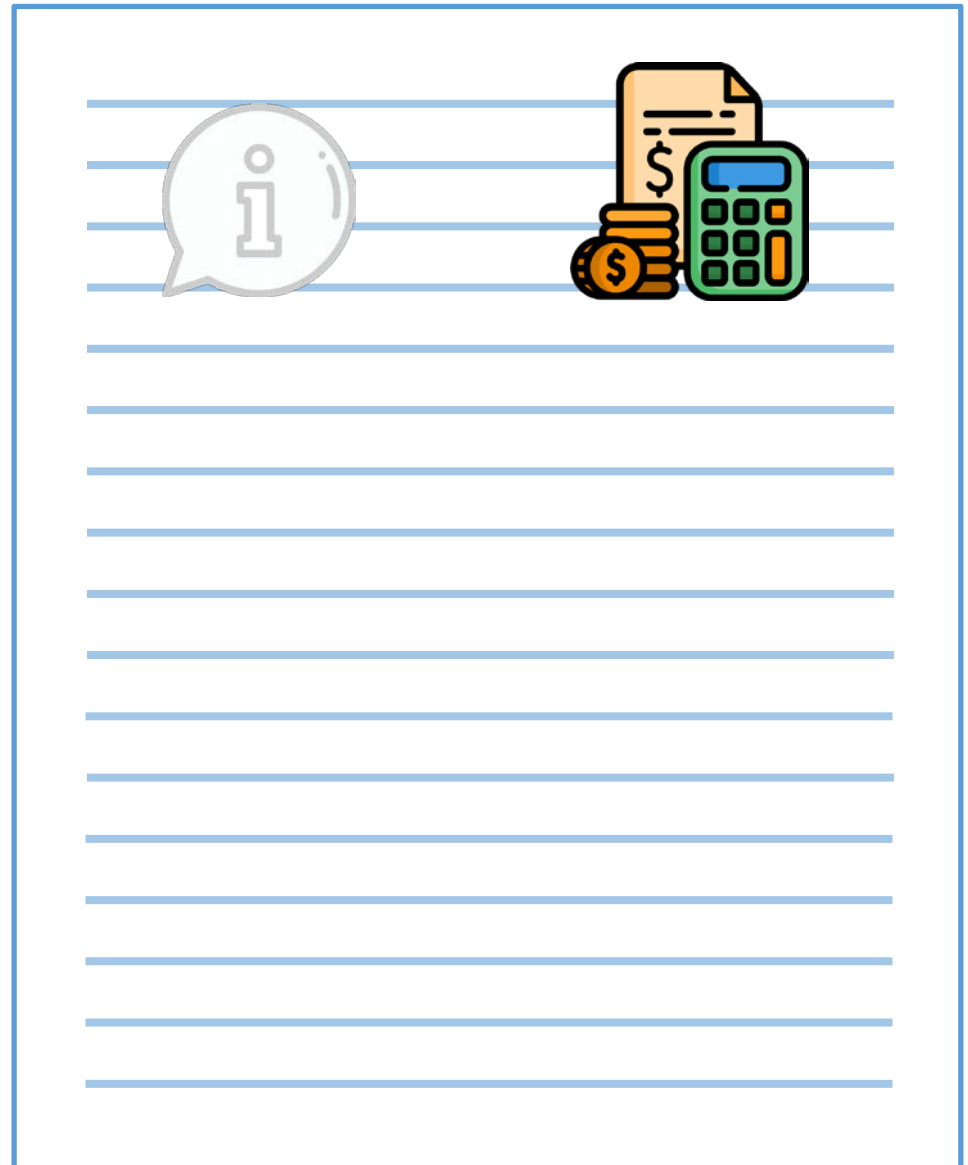
General Information

- Summary
- Title
- File number
- Status
- Name
- Organic location
- **Object of the contract**
- Contract type
- **CPV Classification**
- National subidentity code
- Processing type
- ID
- Lot



Budget

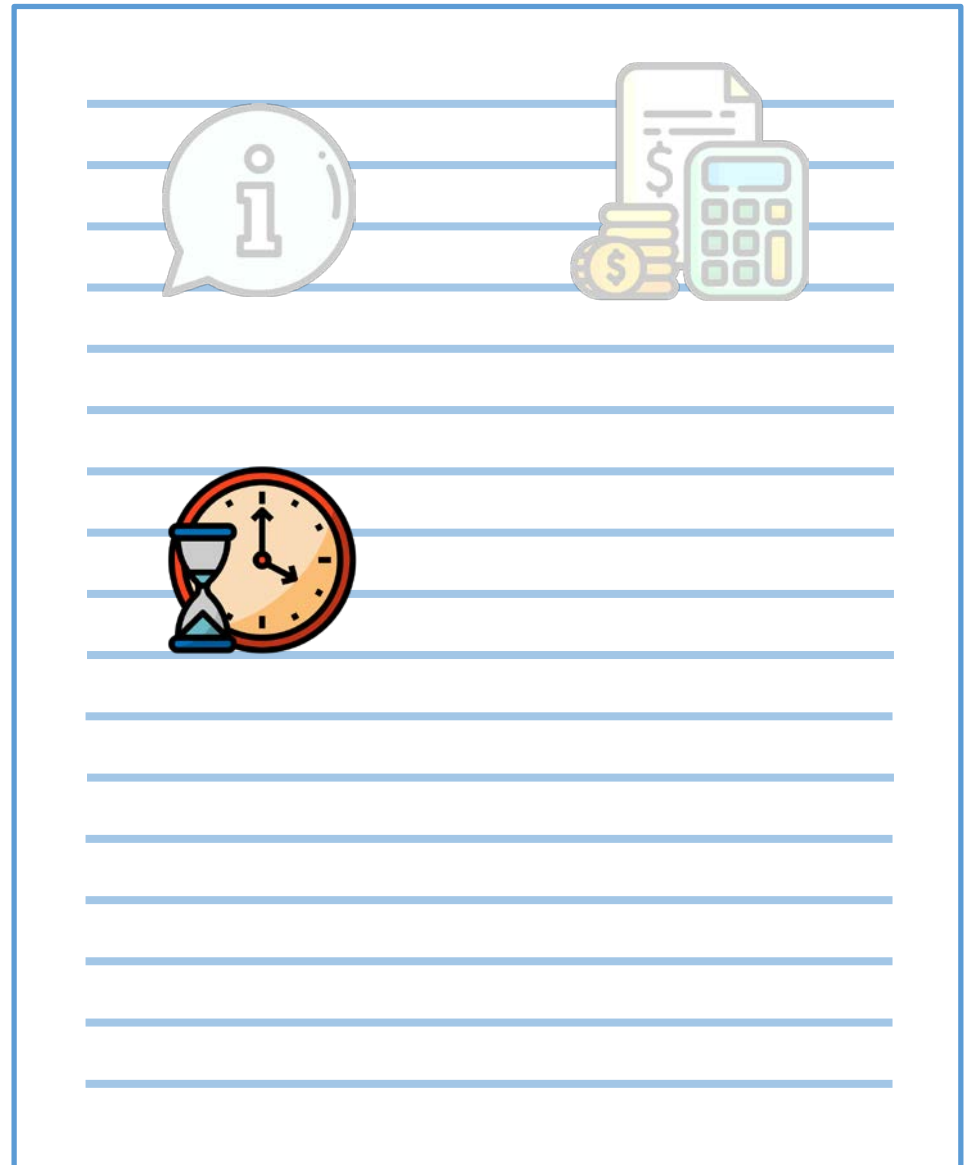
- Contract value estimation
- Base budget without taxes



A large rectangular box with a blue border, containing horizontal blue lines for writing. At the top left is a light blue speech bubble icon with a lowercase 'i'. At the top right are icons for a document with a dollar sign, a stack of gold coins with a dollar sign, and a green calculator.

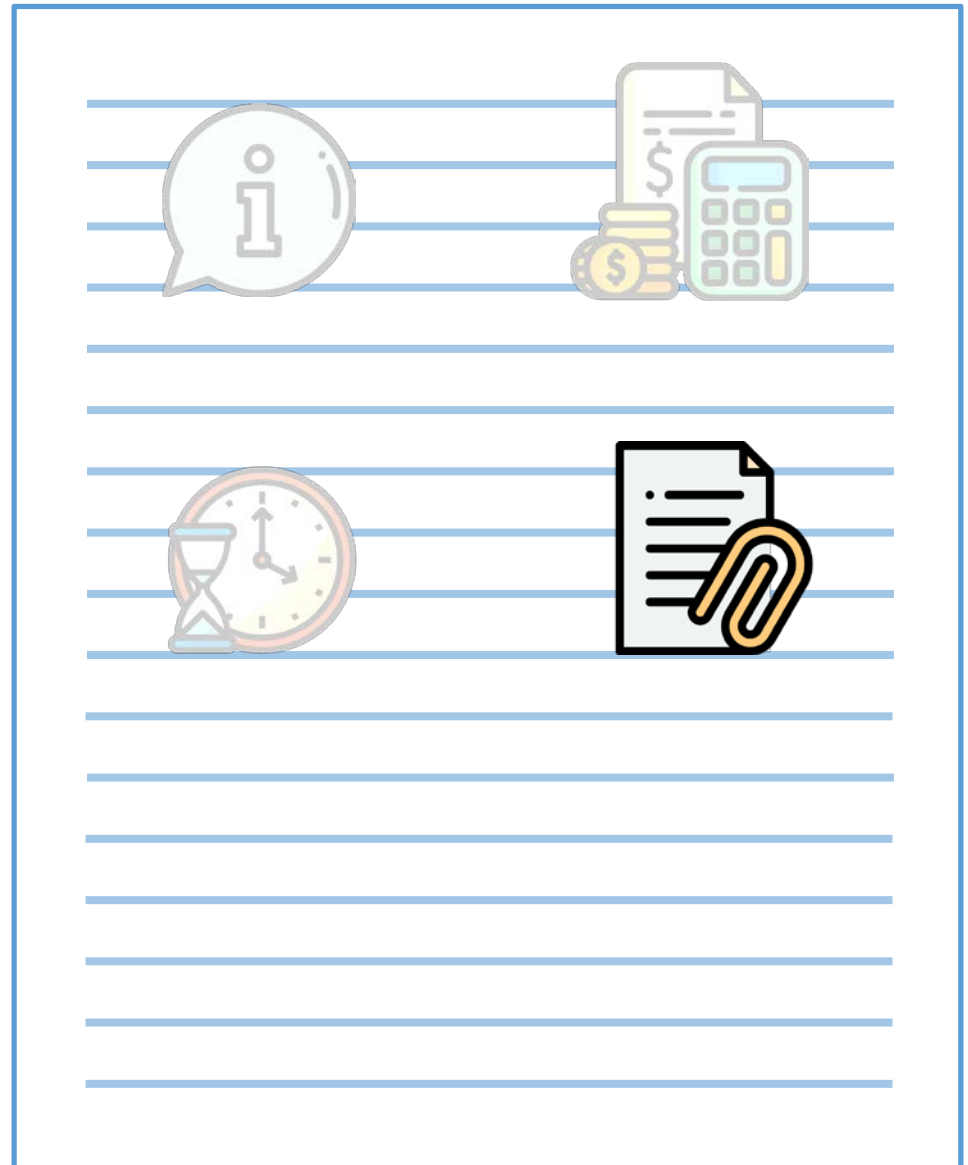
Deadlines

- Execution deadline (Duration)
- Bid Presentation deadline (Date)
- Bid Presentation deadline (Time)
- Type of announcement
- Publication mode
- Publication date
- Execution deadline (Begin)
- Execution deadline (End)



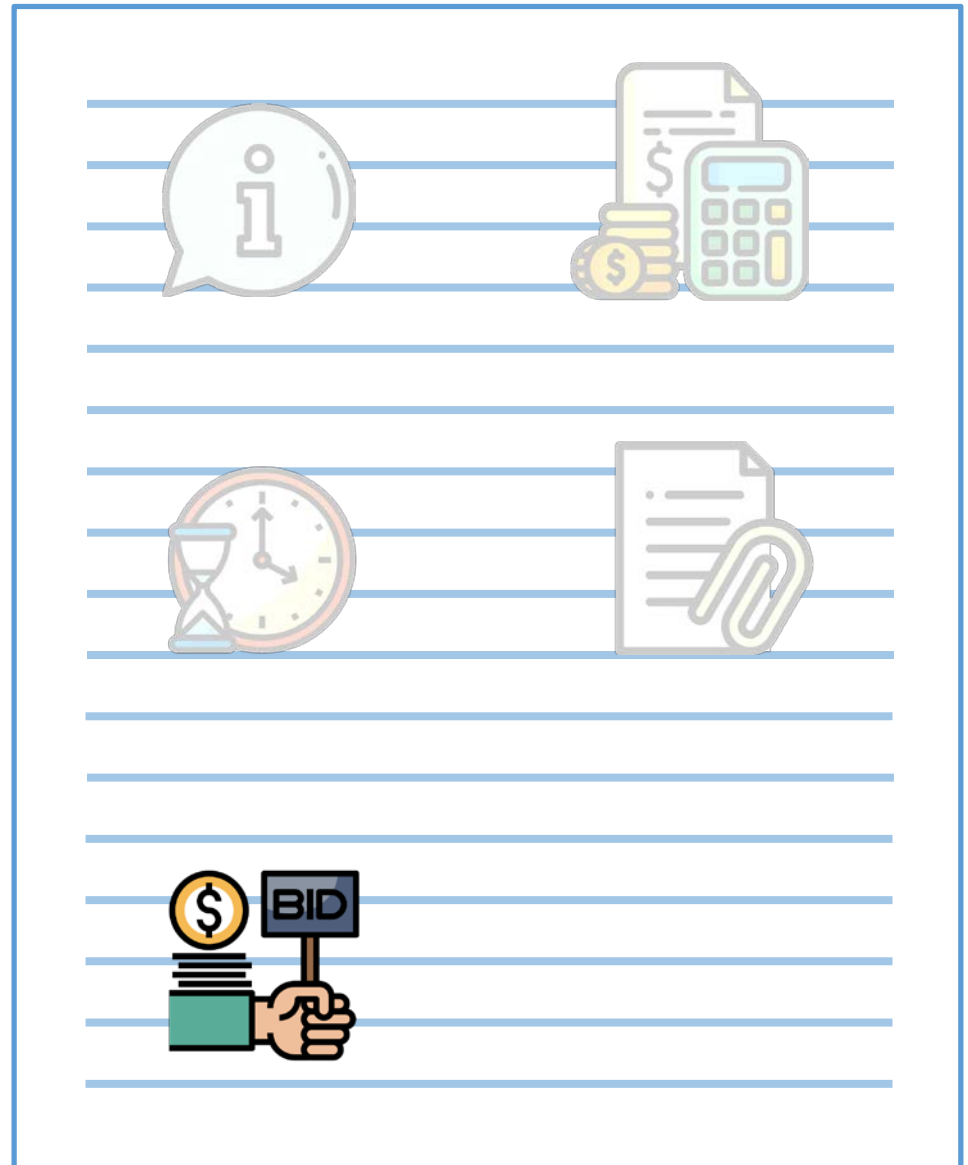
Attachments

- Pliego de cláusulas administrativas
 - Pliego de cláusulas administrativas (URI)
 - Pliego de Prescripciones técnicas
 - Pliego de Prescripciones técnicas (URI)
-
- PDFs
 - OCRs



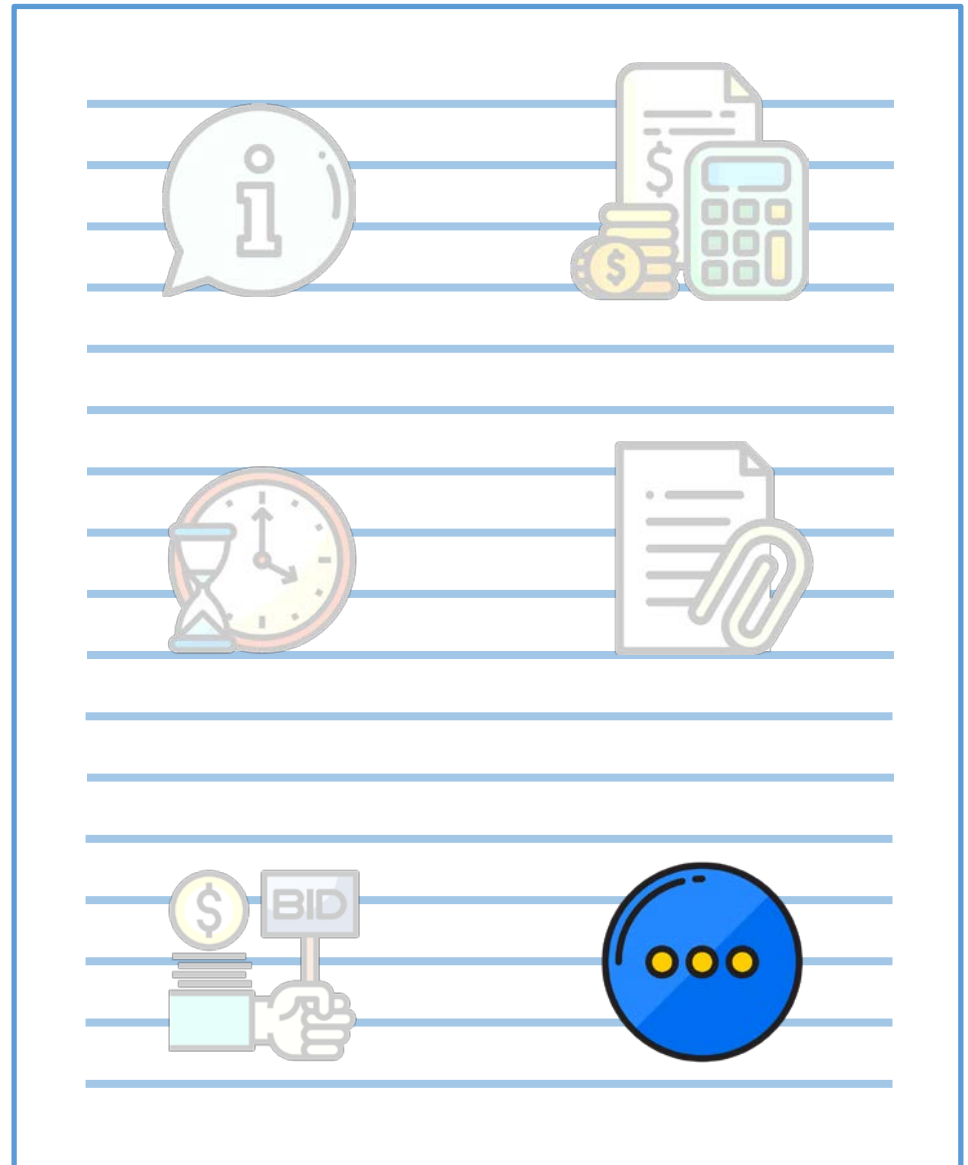
Bidders

- Amount of Bidders
- Identifier
- **Awarded bidder**
- Total amount offered (without taxes)
- Bids (Date)
- Bids (Time)
- URL hiring organization profile
- Bid presentation



Others

- ContractFolderStatus -
- LocatedContractingParty -
- ParentLocatedParty -
- ParentLocatedParty -
- ParentLocatedParty -
- PartyName - Name
- Domain
- file name
- Entry
- deleted_on
- Updated





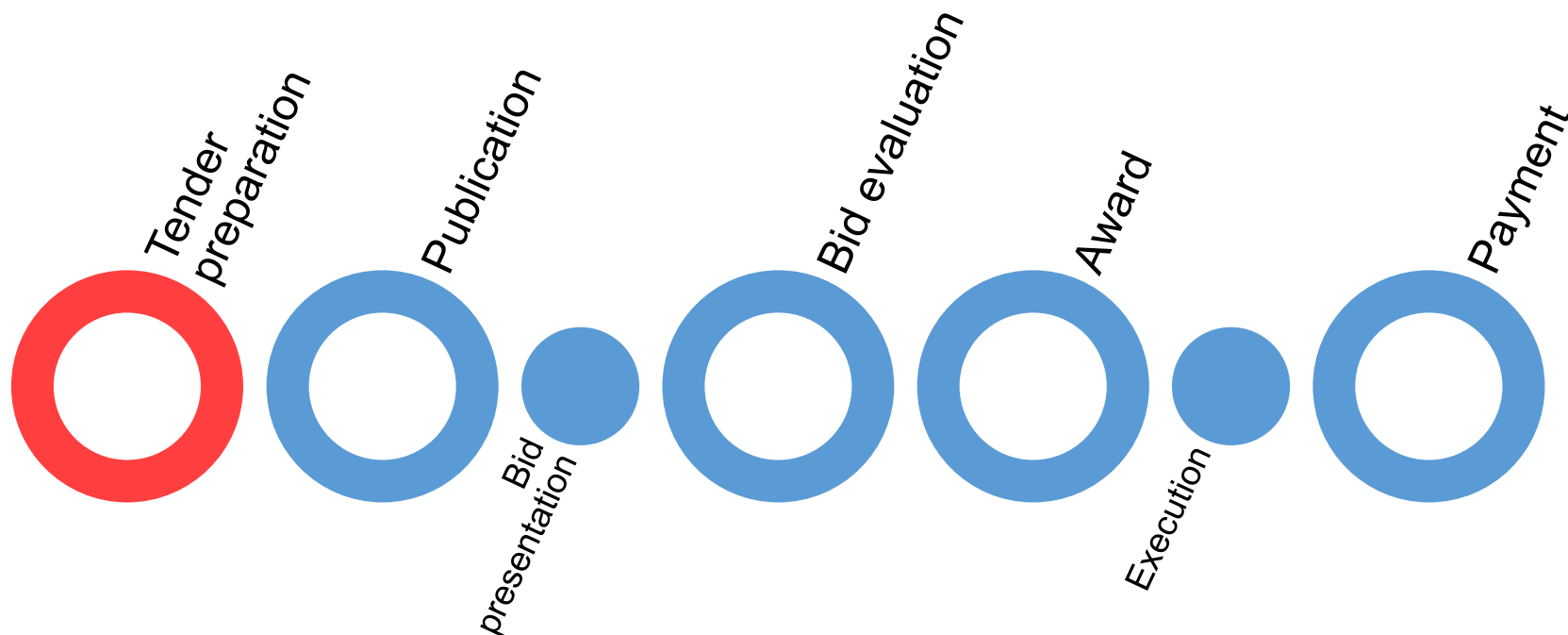
How it works?

TIMELINE



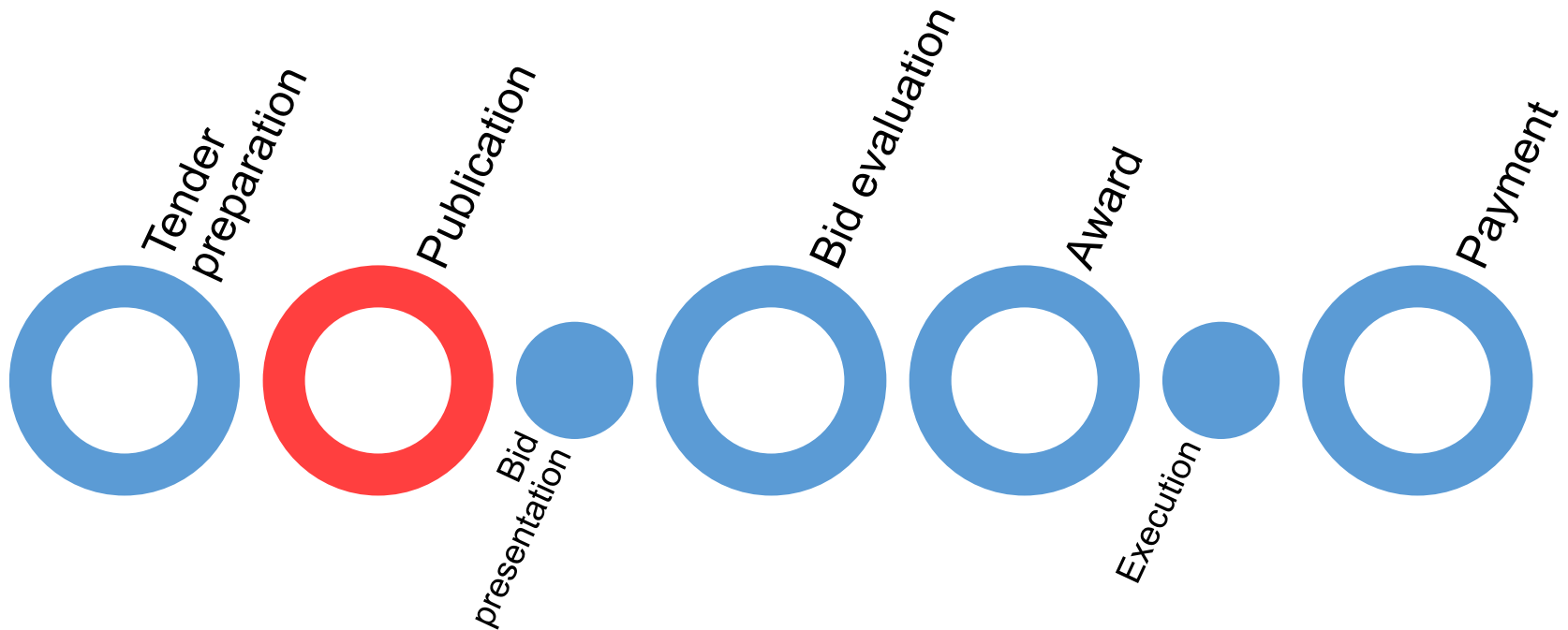
The administration drafts
the tender

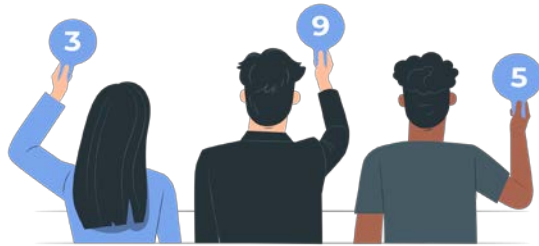
1



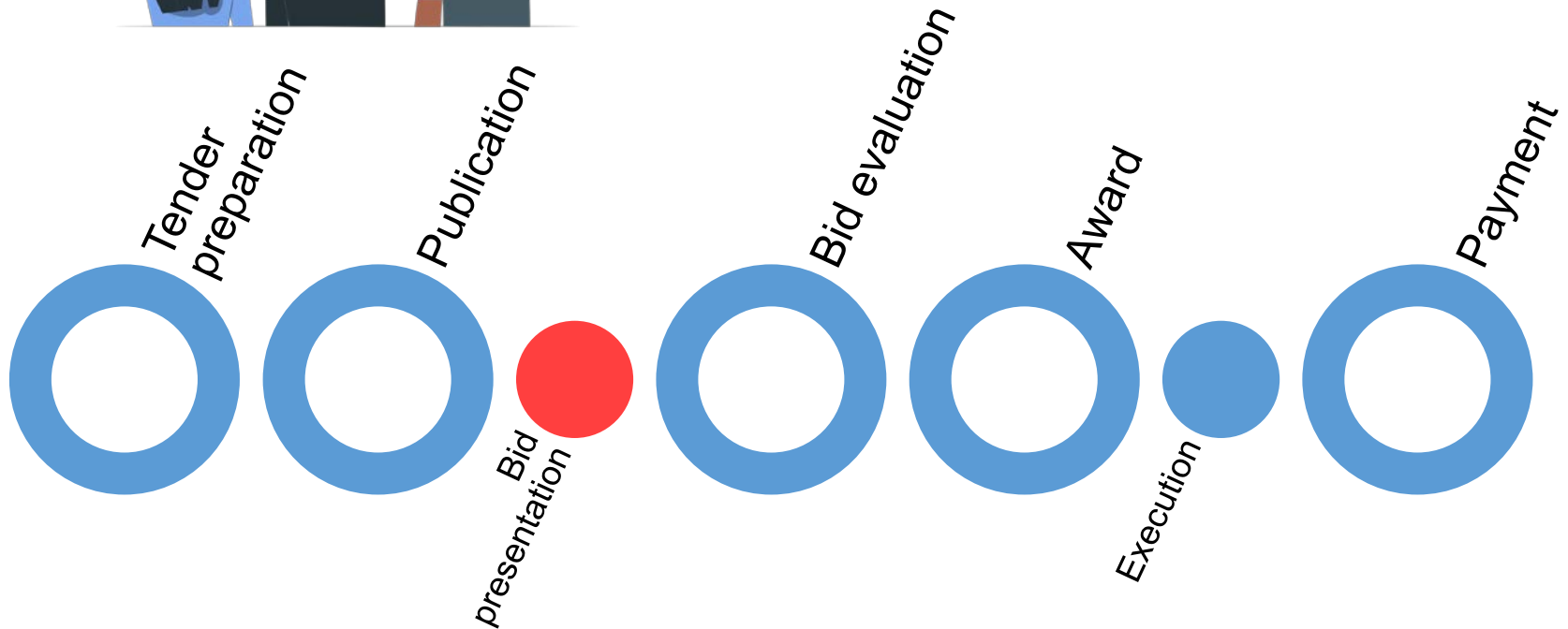


The administration publishes the tender



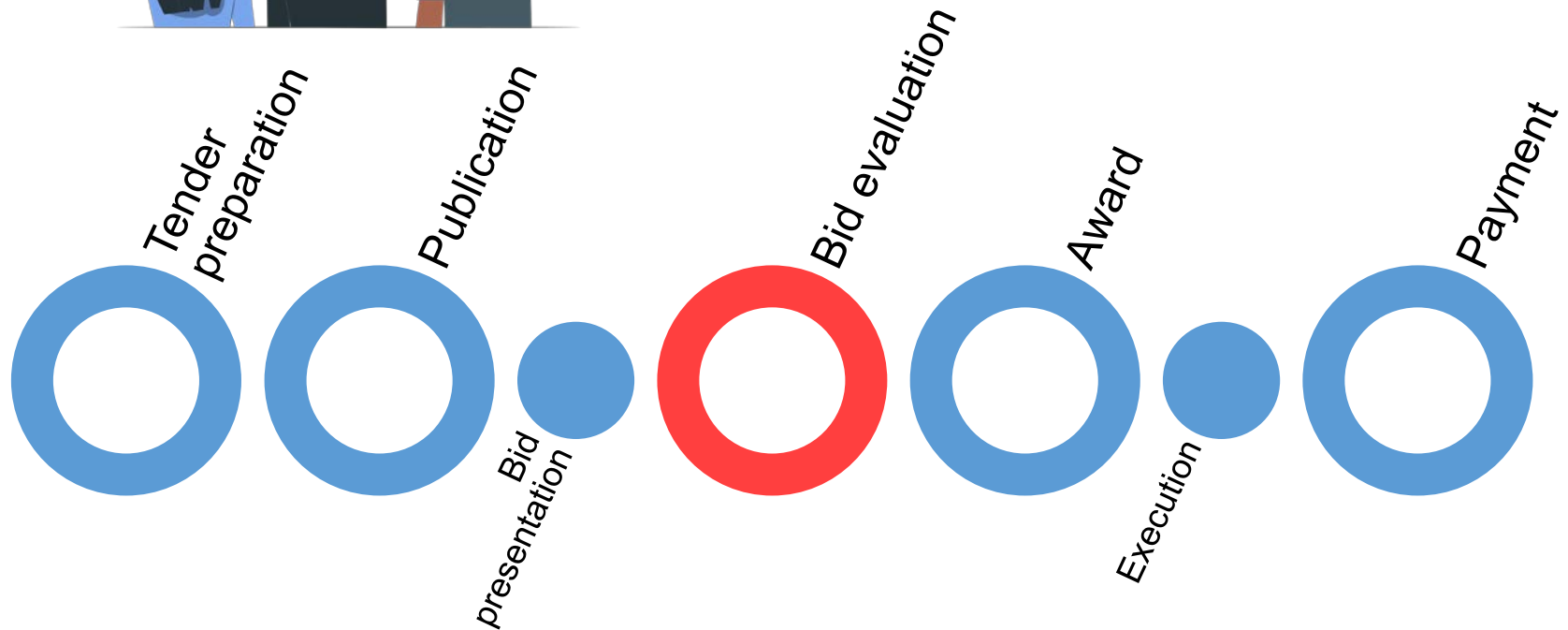


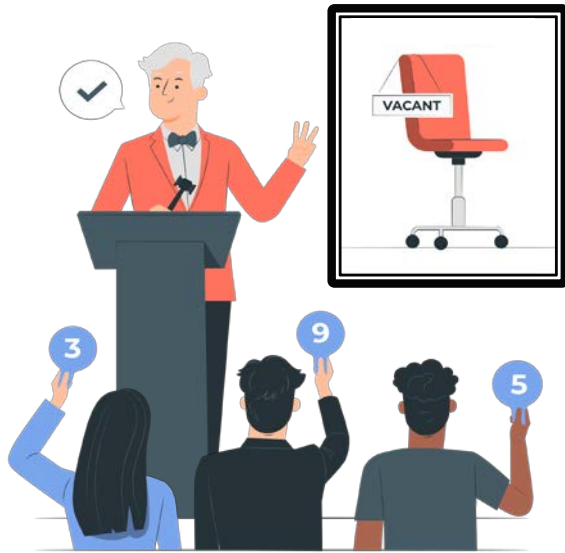
Companies send in
their bids



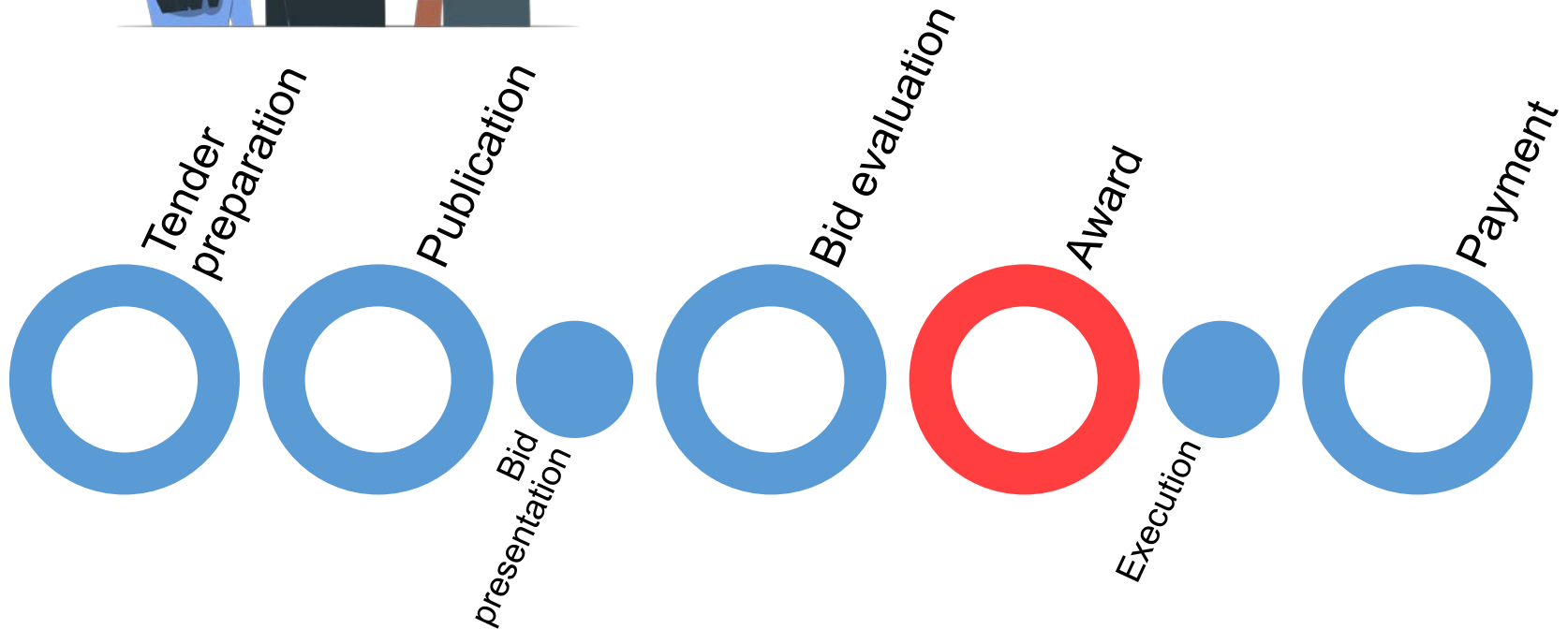


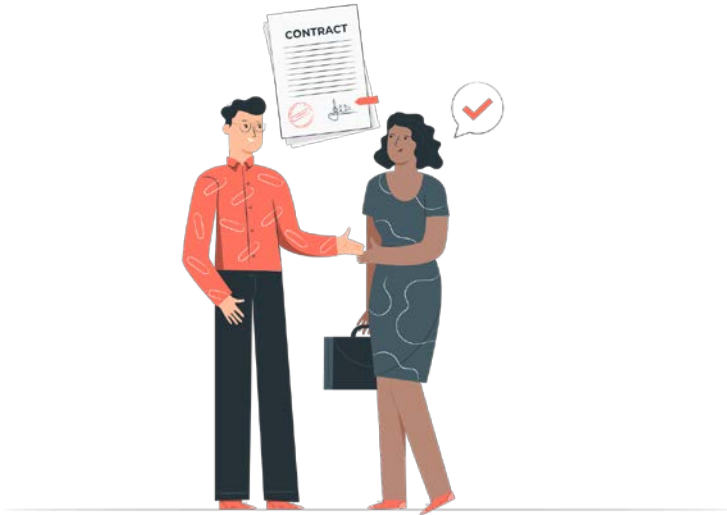
The administration
evaluates the bids



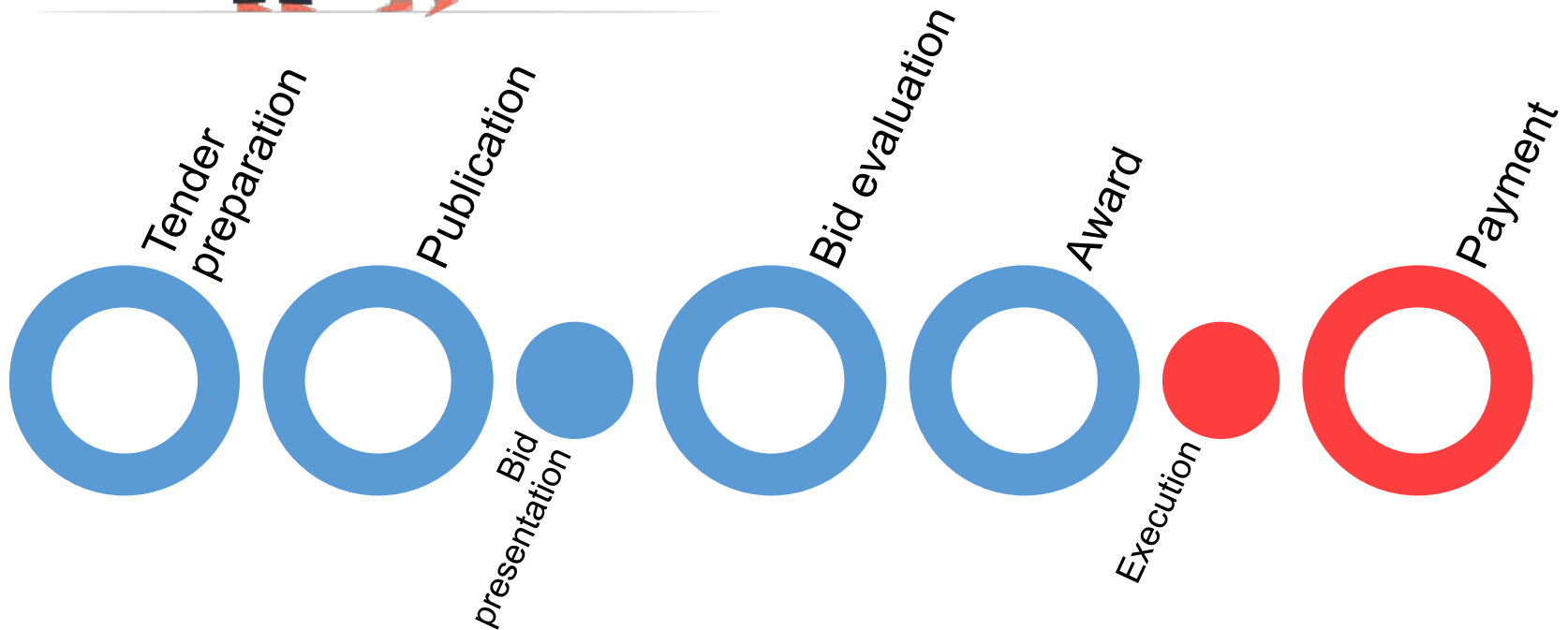


The administration
awards the tender



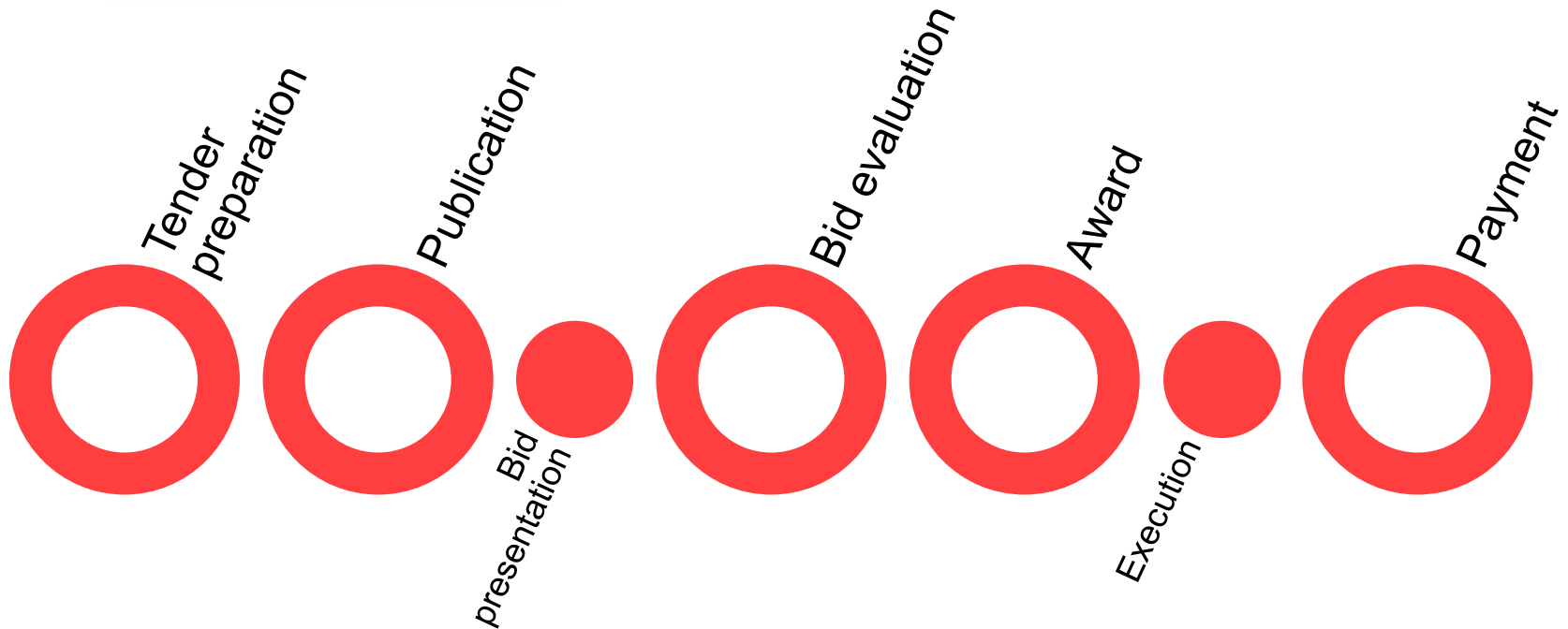


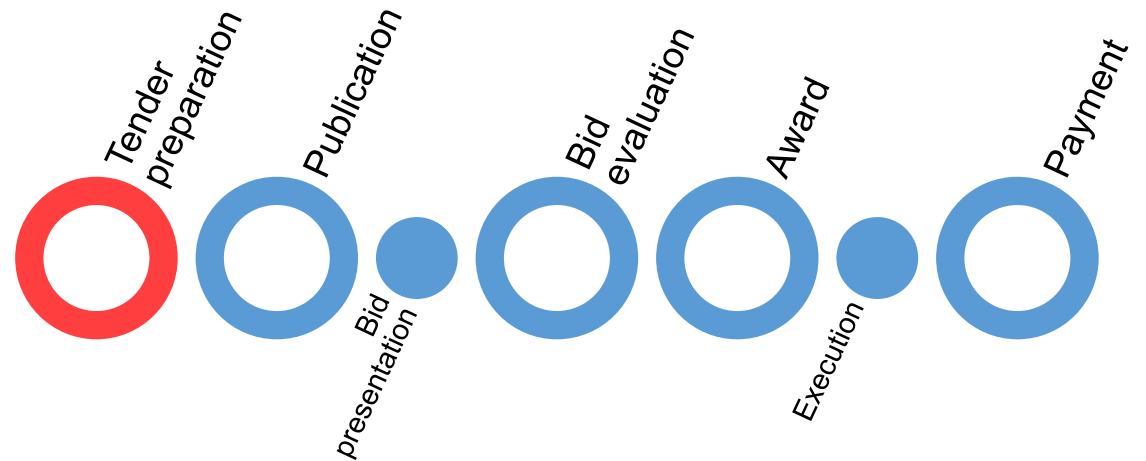
Execution and payment
of the contract





EXTRA
Analyzing the data





Tender Preparation

RESEARCH LINE 1

Problem 1: Imprecision

The wording of the tenders may present imprecisions that are difficult to manage.

- Imprecise words, such as "solvency", are frequently used without being clear what they imply.



Possible Solutions

- Detect the worst written parts of existing tenders and create a taxonomy of the worst errors. Words that tend to be imprecise, such as "solvency", could be detected.
- A style recommender could also be generated on the basis of correct tenders.
- A metric could be designed to measure the clarity and readability.



Main problem here is the need of annotations and expert knowledge to detect and classify main errors

Problem 2: Variability and Inefficiency

Tenders are expressed in many different ways depending on the writer.

- Even if they ask for the same products or services.
- The same work of drafting is done multiple times, incurring great inefficiency.



Possible Solutions

- Locate similar texts in order to use them as a reference
 - based on text similarity
 - other similar aspects (such as budget)



Can be combined with next problem

Problem 3: Criteria

Taxonomy to classify tenders in different areas.

- More than 9k possible codes!
- Different classification guidelines.
- Different levels of granularity used, less specific CPV being commonly used.
- Human errors.



Possible Solutions

- A system that receives the description of a tender and recommends the best fitting CPV codes could help with these issues (ONGOING).



Data available (TED, PLACE, Hacienda...), no need of human annotation!

Problem 4: Budget estimation

- No tools to help building a budget.
 - What is the average price of X?
- Many Awards just because the bid is the cheapest.
- What is the relation between the tender budget and the awarded bid budget?

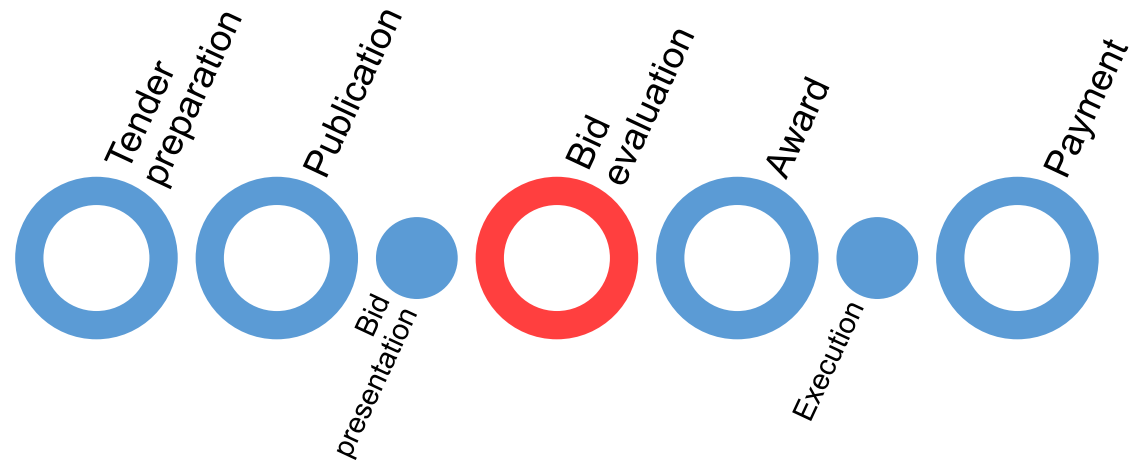


Possible Solutions

- Use text similarity to find average prices for similar objects.



Related to previous problems.



Bid Evaluation

RESEARCH LINE 4

Problem 1: Low amount of bidders

The more companies that apply for a tender, the higher the quality of the final Award. But:

- Many tenders without bidders
- Many tenders with just one bidder.
- Low amount of SMEs being awarded.



Possible Solutions

- Clusterings taking into account the amount and type of clauses (e.g. social) to understand low bidding.



Related to next problem.

Problem 2: Who bids?

- Difficult to avoid problems such as:
 - Collusion
 - Awarding of contracts to the same companies without justified reasons.
- The analysis of competition would gradually improve the publication of tenders.



Possible Solutions

- Profiling: define typologies among bidders, levels of participation...
- Analysis of the relation among amount of bidders and tenders (object, clauses,...)
- Create a network of bidders/tenders.



Data is there, basic Machine Learning looks promising... TFM?

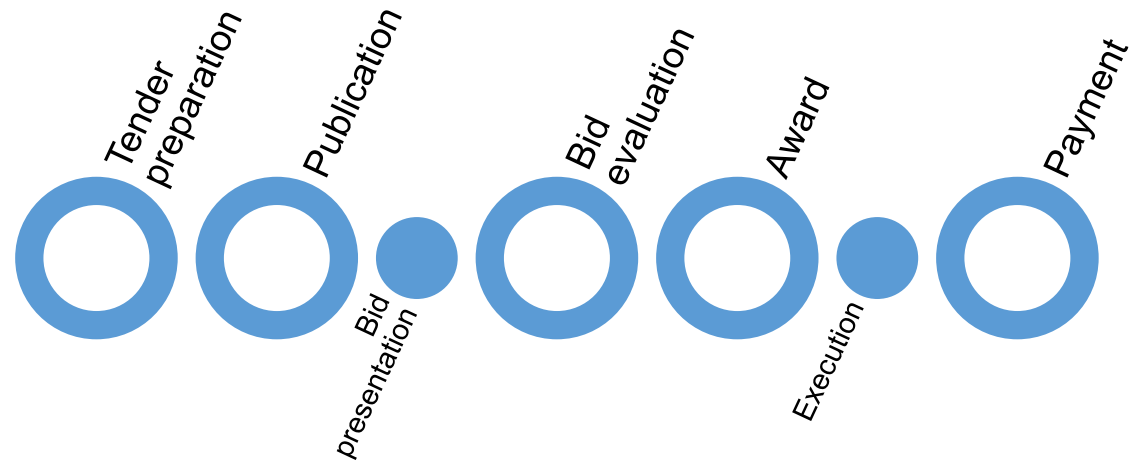
Problem 3: How do clauses affect?

- Different clauses with different evaluation criteria.
 - Could we develop a typology? (e.g., social, environmental, innovation...)
 - Could we develop evaluation models?
- Sometimes, there are “special execution conditions”, how do they affect to bidding and execution?



Possible Solutions

- Identify and generate a repository of clauses related to objects/topics/CPVs...
- Identify and generate a repository of validation criteria related to clauses.
- A KB or KG?



Reusing data

EXTRA RESEARCH LINE

Problem: How to retrieve info

- They want to be able to make queries to the system.
 - How many pills X were bought last year?
- Could this be done with a KG?

Possible Solutions

- QA system
- Preprocessing NLP in tenders to derive triples.
- Probably SPARQL queries should be expressed as natural language.



- Many open problems!
- Many different research lines → not all of them feasible/of interest.
- If you are interested, they committed to annotate/validate.
- Great opportunity to assemble and adapt different results from other domains.



For more information:

- **Multi-label Text Classification for Public Procurement in Spanish** (SEPLN)
María Navas-Loro, Daniel Garijo, Oscar Corcho
- **NextProcurement: Challenges in Public Procurement in Spain** (AI4LEGAL)
María Navas-Loro

Follow [@JaimeObregon](https://twitter.com/JaimeObregon) in Twitter

<http://nextprocurement-project.com/#> (project)

<https://procurement.linkeddata.es/> (our contributions)

Planning a workshop proposal about Public Procurement to [ICAIL](#) (one of the best conferences in the legal domain, suggestions of other venues are welcomed!)



19th International Conference on Artificial
Intelligence and Law - ICAIL 2023

19th-23rd June 2023, Braga, Portugal

ICAIL 2023



Credits:

- FlatIcon: www.flaticon.es
- StorySet: <https://storyset.com/>

If you think that your tool can be applied to the domain and you want me to present your work to the partners, let me know!



THANK YOU FOR YOUR ATTENTION



COMMENTS WELCOME 😊



NLP techniques on Public Procurement textual data

Challenges and Opportunities

María Navas-Loro, Óscar Corcho
Ontology Engineering Group
Universidad Politécnica de Madrid, Spain



mnavas@fi.upm.es



<https://marianavas.linkedata.es>



2022-10-21



NextProcurement

