# French-German AI Summer School

## Transformer models and Text data availability

**ILIADE** Team
(**I**nte**LI**gence **A**rtificielle, **D**onnées et **E**xpérimentations)

**Speakers** : Baptiste Doyen (Data Scientist)

22/06/2021

BANQUE DE FRANCE
EUROSYSTÈME

Industrial Data Analytics
& Machine Learning

MACHINE LEARNING AND MASSIVE DATA ANALYSIS

# Introduction : *at the cross-roads of two situations*

## Inside context

Diversity of *activities* where **text data** plays a major role inside our organization :

- ☐ *Regulatory* (loan contracts, investment programs brochures)
- ☐ *Web Monitoring* (Twitter, websites scrapping)
- ☐ *Economic Analysis* (written reports from analysts, FIBEN database)
- ☐ *Financial Education* (search engine in natural language for SME owners)
- ☐ *Cybersecurity* (command lines, shell programs)

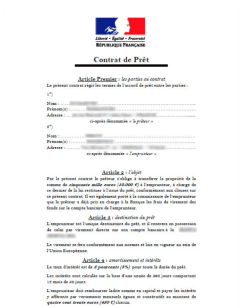



FIGURE – Loan Contract

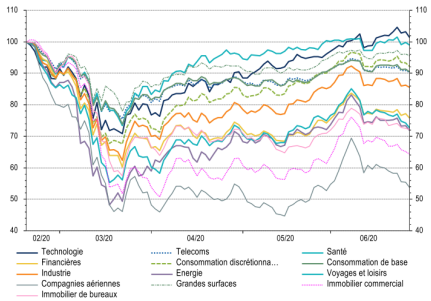



FIGURE – Data-2-Text in econometry


FIGURE – SME owners Google-like

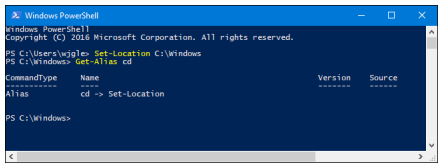



FIGURE – Some PowerShell code...

BANQUE DE FRANCE
EUROSYSTÈME

# Introduction : *at the cross-roads of two situations*

## Outside context

Powerful and easily accessible **NLP technologies** since 2017 : *Attention mechanism*[1], *Transformer models* and *fine-tuning paradigm* powered by *GPUs* hardware

*And our favorite toolkit :*

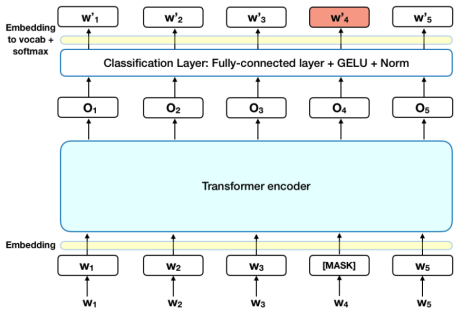



FIGURE – Open Source Weights & Architectures : BERT[2], CamemBERT[3], ...

PyTorch

FIGURE – Deep Learning made 'easy'

Transformers

FIGURE – Fine-tuning made 'easy'

FIGURE – HPC server with GPUs

BANQUE DE FRANCE
EUROSYSTÈME

# Introduction : *at the cross-roads of two situations*

👎 But there is a **bad news** : resources are scarce, incomplete or too abundant...

👍 And a **good news** also : Transformer models offer a very interesting variety of approaches able to tackle multiple settings with *zero-shot learning*, *few-shot learning*, *fine-tuning* and *language modelling*

| DL **task** \ GB of **Text** | 0 | 0.001-0.01 *collected by 'hand'* | 0.1 *Benchmarks datasets* | 10 *Wikipedia dump* |
|---|---|---|---|---|
| *Classification* | ✓ | ✓ | x | x |
| *NER(+RE)* | ? | ✓ | ? | x |
| *MLM(+NSP)* | | | | ✓ |
| *Summarization* | ✓ | ✓ | ✓ | x |
| *Generation* | ✓ | ? | ? | x |

In-house task/data availability table : in-reach, possibly in-reach and out-of-reach

BANQUE DE FRANCE
EUROSYSTÈME

# Summary

1. Part I : Zero-Shot Learning for validating real-estate loan object
2. Part II : Fine-tuning a NER for extracting personnal entities on loan contracts
3. Part III : PowerShell-CyBERT : a Masked Language Model on PowerShell command lines for detecting LOLbins attacks

BANQUE DE FRANCE
EUROSYSTÈME

1. Part I : Zero-Shot Learning for validating real-estate loan object
2. Part II : Fine-tuning a NER for extracting personnal entities on loan contracts
3. Part III : PowerShell-CyBERT : a Masked Language Model on PowerShell command lines for detecting LOLbins attacks

BANQUE DE FRANCE
EUROSYSTÈME

## Actual problem

Validate or not the **eligibility of a loan object**.
This object is valid iff it refers to a personnal real-estate loan (e.g. : "*purchase of a house with a living area of 100m2 including 5 rooms*").
Loans for professional purpose are not accepted (e.g. : "*Financing purchase of a brand-new MAN 35480 BL truck year 2014*")

## Data at our disposal

**None !** (at least for training...) | Some test data will be collected finally

## Transformers called to the rescue...

**Zero-Shot Learning** Models : fine-tuned NLI models turned as universal classifiers

- **NLI** : $p_\Theta(seq_1, seq_2) \in \{entailment, neutral, contradiction\}$
- **Classifiers** : $p(seq|hyp) = \tilde{p}_\Theta(seq, hyp) \in \{entailment, contradiction\}$
- **Models** : CamemBERT-base-xnli (110M) ; CamemBERT-large-xnli (335M) and XLM-RoBERTa-large-xnli [4] (15 languages, 550M)

BANQUE DE FRANCE
EUROSYSTÈME

## In practice : there is no magic...

Usual real-life ML hurdles are still well-grounded here

- **Pre-processing** : models are sensible to negation ("*transformation of non-habitable premises into habitation*")
- **Model benchmark** : how to estimate performance on real data ? XNLI benchmark dataset[5] was not very useful... (81.9% ; 85.2% and $\sim 90\%$ accuracy)
- **Model size** : the larger the better in terms of performance but it poses rapidly a size problem (XLM model $> 2Gb$).

  To address this problem : we used **dynamic quantization** with ONNX[6]

  - quantization : mapping *float*32 parameters to *int*8 values using

$$W_{float32} = \text{scale}_W * (W_{int8} - \text{zero}_W) \quad (1)$$

    where $W$ is a weight matrix and $\text{scale}_W, \text{zero}_W$ are constant terms
  - dynamic : $\text{scale}_W, \text{zero}_W$ are computed during runtime according to $W$ input
  - shortcomings : more unstable compared to static (CPUs) but not possible (yet) for large models $> 2Gb$

BANQUE DE FRANCE
EUROSYSTÈME

- **Hyperparameter tuning** : here it's not a numerical one ! "*hypothesis*" in $p(.|hypothesis)$ can be tuned
- **Uncertainty** : how to handle $50\% - 50\%$ predictions ? (Empirical) By averaging $n$-grams predictions ($n = len(seq) - 1$)

## Numerical results

Let's see the results on an in-house test set composed of 44 sequences (29 positives and 15 negatives)

| | | Predicted Pos | Predicted Neg |
|---|---|---|---|
| Actual | Pos | 25 | **15** |
| | Neg | **4** | 0 |

Non-Quantized CamemBERT-large-xnli with uncertainty catching-up ($F_1 = 0.72$)

| | | Predicted Pos | Predicted Neg |
|---|---|---|---|
| Actual | Pos | 27 | **2** |
| | Neg | **2** | 13 |

Quantized and Non-Quantized XLM-RoBERTa ($F_1 = 0.93$)

| | | Predicted Pos | Predicted Neg |
|---|---|---|---|
| Actual | Pos | 29 | 0 |
| | Neg | **1** | 14 |

Quantized and Non-Quantized XLM-RoBERTa with uncertainty catching-up ($F_1 = \mathbf{0.98}$)

BANQUE DE FRANCE EUROSYSTÈME

1. Part I : Zero-Shot Learning for validating real-estate loan object
2. Part II : Fine-tuning a NER for extracting personnal entities on loan contracts
3. Part III : PowerShell-CyBERT : a Masked Language Model on PowerShell command lines for detecting LOLbins attacks

BANQUE DE FRANCE
EUROSYSTÈME

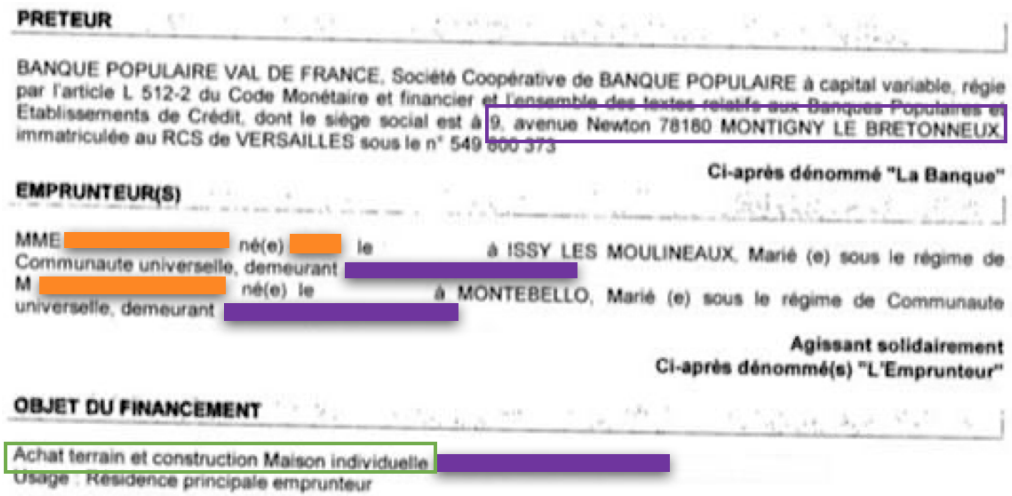
## Actual problem

Extract personnal informations (e.g. : names and adresses) from loan contracts

## Data at our disposal

Hundreds of PDFs documents from various banks. A pre-processing OCRisation phase is hence necessary to collect raw text from these documents.
**After annotations** : 677 annotated documents, only 1/4 of tokens carry an entity

Region of interest on loan contract



PRETEUR

BANQUE POPULAIRE VAL DE FRANCE, Société Coopérative de BANQUE POPULAIRE à capital variable, régie par l'article L 512-2 du Code Monétaire et financier et l'ensemble des textes relatifs aux Banques Populaires et Etablissements de Crédit, dont le siège social est à 9, avenue Newton 78180 MONTIGNY LE BRETONNEUX, immatriculée au RCS de VERSAILLES sous le n° 549 800 373

Ci-après dénommé "La Banque"

EMPRUNTEUR(S)

MME né(e) le à ISSY LES MOULINEAUX, Marié (e) sous le régime de Communauté universelle, demeurant

M né(e) le à MONTEBELLO, Marié (e) sous le régime de Communauté universelle, demeurant

Agissant solidairement
Ci-après dénommé(s) "L'Emprunteur"

OBJET DU FINANCEMENT

Achat terrain et construction Maison individuelle

Usage : Résidence principale emprunteur

## Transformers called to the rescue...

Fine-tuning of a **NER** (token classification) model using CamemBERT-base

- In practice, there is **some "magic"**...

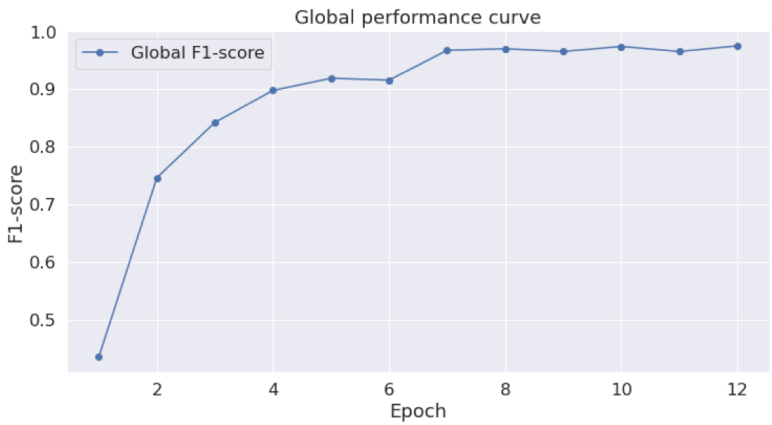



FIGURE – *F*1-score on all entites

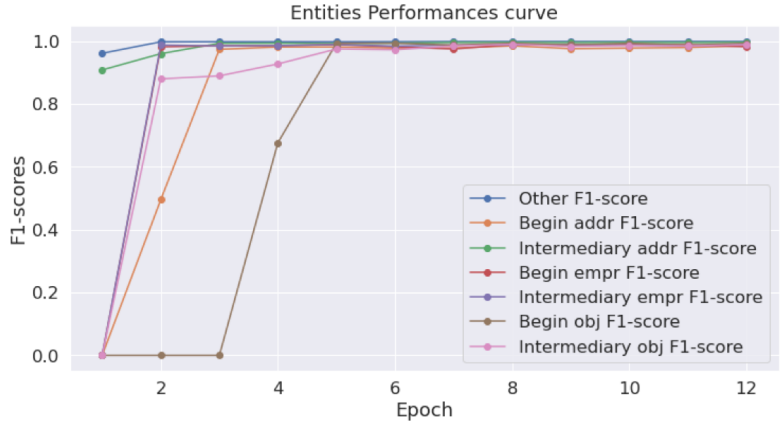



FIGURE – *F*1-score for each entity

- And it's **not a surprise** !

| **Benchmark** : | Zhang et al. (20)[7] | Barriere et al. (19)[8] | Ours (20) |
|---|---|---|---|
| $F1$ | 0.83 | 0.96 | 0.97 |
| $N_{entities}$ | 7 | 10 | 3 |
| $N_{docs}$ | 90-120 | 585 | 677 |
| Model | BERT Base-Chinese | CamemBERT | CamemBERT |

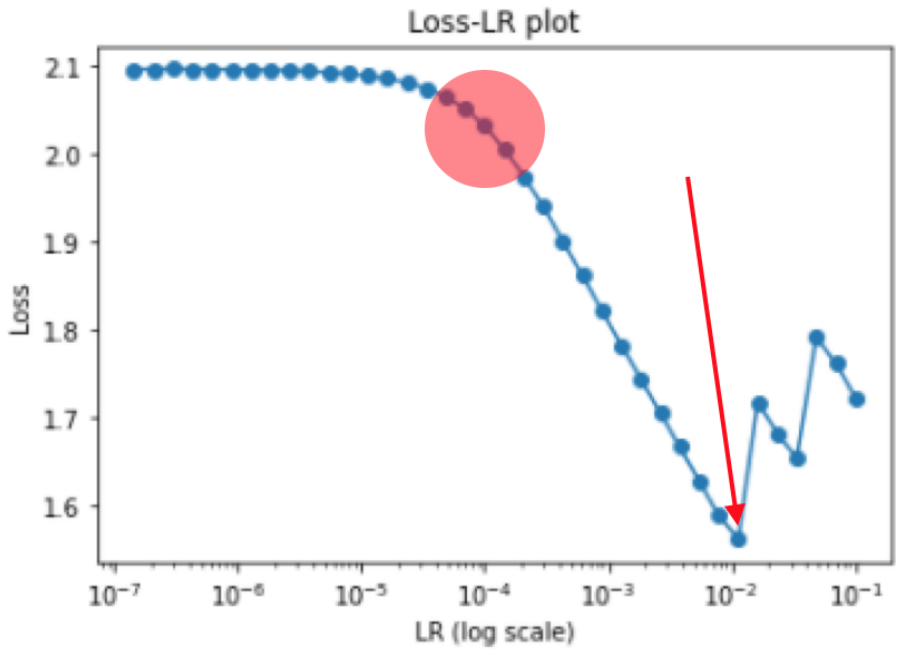
BANQUE DE FRANCE
EUROSYSTÈME

## **Going further** : AutoML for NER

Is it possible to find a way to make the training procedure adaptative for this type of NER task ? *Suggested procedure* :
$\{AutoLRFinder\} \bigoplus \{EarlyStopping \text{ (on valid } loss)\} \bigoplus \{\text{grid search on } batch_{size}\}$

Description of *AutoLRFinder* procedure :

- Compute $Loss(lr)$ with $lr(x) = \alpha * exp(x * \beta)$, s.t. $lr(0) = lr_{start}$ and $lr(N_{steps}) = lr_{end}$
- Find $lr_{max} = \underset{lr}{\operatorname{argmin}}\{Loss(lr)\}$
- Find $lr_{steepest} = \underset{]-\infty;lr_{max}]}{\operatorname{argmax}}\{\frac{\partial Loss(lr)}{\partial lr}\}$




BANQUE DE FRANCE
EUROSYSTÈME

## **Going further** : MetaNER **without** training data

**In an usual setting** : raw data are collected, entities are annotated from it and serves training.
**But, inversely**, is it possible to go from entities directly and build a context around for generating documents and train a model from these generated data ?

Example of *Entities2Context* data generation procedure, inspired by [9] :

- **Entity** : 9 avenue Montigny le Bretonneux, 75009 Paris
- **Input** : <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> 9 avenue Montigny le Bretonneux, 75009 Paris <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask>
- **Final Output** : _La _partie _sud _de _la _maison _est _de _plain _pied _9 _avenue _Mont igny _le _Breton neux _, _750 _09 _Paris _située _dans _le _prolongement _de _la _maison _à _portes _intérieures

At each step, a random <mask> token is selected and replaced by another token $\in$ *Vocab* sampled using CamemBERT as a bidirectional MLM

BANQUE DE FRANCE
EUROSYSTÈME

1. Part I : Zero-Shot Learning for validating real-estate loan object
2. Part II : Fine-tuning a NER for extracting personnal entities on loan contracts
3. Part III : PowerShell-CyBERT : a Masked Language Model on PowerShell command lines for detecting LOLbins attacks

BANQUE DE FRANCE
EUROSYSTÈME

## Actual problem

Detect **LOLbins attacks** from all logs events collected by the cybersecurity departement. LOLbin ("Living Off the Land binary") : an usual Windows binary application (.exe) executed in a diverted way.

## Data at our disposal

Millions of raw logs events, each event containing various data like IP addresses, time of the execution, machine of execution, status of user, ...
**Main data** : **PowerShell command line** that triggered the application

Example of command lines (*regsvr32.exe*, used by Windows to register dlls)

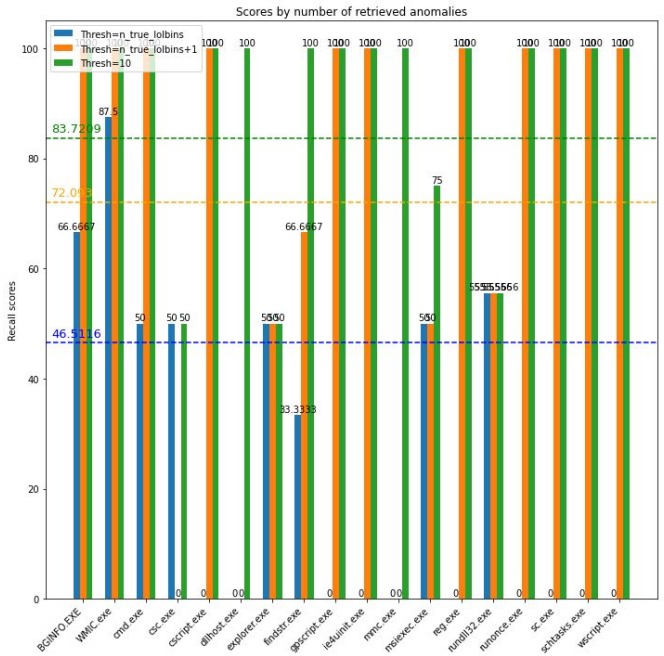
- Usual way : *regsvr32.exe /s /u C :\\Program Files\\McAfee\\ScriptSn.dll*
- LOLbin way[10] : *regsvr32.exe /s /u /i :file.sct scrobj.dll* (*Squiblydoo* attack)

## Transformers called to the rescue...

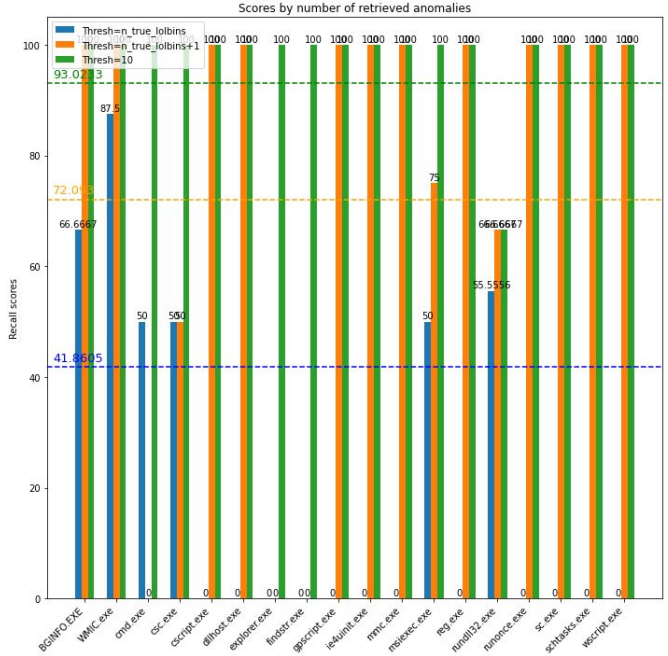
How to compute an **embedding** for the command line by leveraging all the available data ?
Training a **RoBERTa-MLM**[11] on these millions of command lines ($\sim$ 12M unique) proved to be a tractable and performing option

- **Pre-processing** : how to prepare as well as possible sequences for MLM ?
  - Avoid **redundancy within the sequence** by "naturalizing" (e.g. : strip punctuations)
  - Avoid **redundancy between sequences** (e.g. : hide numbers)
  - Limit **sequence lengths** between min-max bounds
- Choice of the **Tokenizer** : custom or pre-trained ?
  - Custom : training of a **BPE (*Byte-Pair-Encoding*) tokenizer** (of which vocabulary size ?)
  - Pre-trained : tokenizer from **RoBERTa-base** model pre-trained on plain english
- **Results** : a comparison of recall metrics with another language model

Word2Vec




CyBERT




BANQUE DE FRANCE
EUROSYSTÈME

Thank you for your attention ! :)
Any questions ?

BANQUE DE FRANCE
EUROSYSTÈME

# Bibliography


[1] Vaswani et al. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[2] Jacob Devlin et al. BERT : pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805.

[3] Louis Martin et al. Camembert : a tasty french language model. *CoRR*, abs/1911.03894, 2019. URL http://arxiv.org/abs/1911.03894.

[4] Alexis Conneau et al. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116, 2019. URL http://arxiv.org/abs/1911.02116.

[5] Alexis Conneau et al. XNLI : evaluating cross-lingual sentence representations. *CoRR*, abs/1809.05053, 2018. URL http://arxiv.org/abs/1809.05053.

[6] Bai et al. Onnx : Open neural network exchange. https://github.com/onnx/onnx, 2019.

[7] Ruixue Zhang et al. Rapid adaptation of BERT for information extraction on domain-specific business documents. *CoRR*, abs/2002.01861, 2020. URL https://arxiv.org/abs/2002.01861.

[8] Valentin Barrière and Amaury Fouret. May I check again ? - A simple but efficient way to generate and use contextual dictionaries for named entity recognition. application to french legal texts. *CoRR*, abs/1909.03453, 2019. URL http://arxiv.org/abs/1909.03453.

[9] Alex Wang and Kyunghyun Cho. BERT has a mouth, and it must speak : BERT as a markov random field language model. *CoRR*, abs/1902.04094, 2019. URL http://arxiv.org/abs/1902.04094.

[10] Oddvar Moe et al. Living off the land binaries and scripts (and also libraries). https://github.com/LOLBAS-Project/LOLBAS, 2020.

[11] Yinhan Liu et al. Roberta : A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. URL http://arxiv.org/abs/1907.11692.


BANQUE DE FRANCE
EUROSYSTÈME