

# Papy-S-Net : A Siamese Network to Match Papyrus Fragments

---

Antoine Pirrone, Marie Beurton-Aimar, Nicholas Journet

June 7, 2019



université  
de BORDEAUX



**LaBRI**

Horizon 2020  
European Union funding  
for Research & Innovation

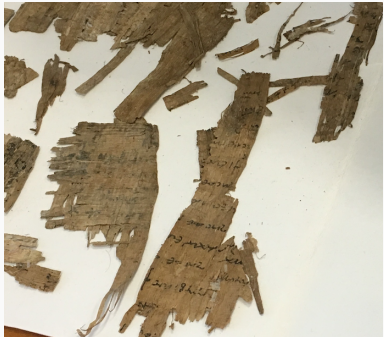
# Context

- *GESHAEM* Project (European Project, ERC)
- Digitalize and study the content of papyri



# Papyrologist work

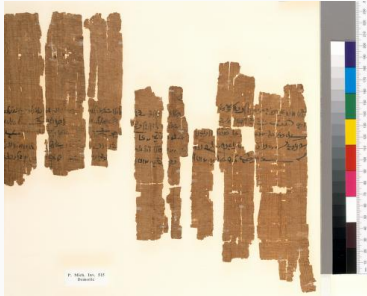
- Select funeral ornaments
- Remove the paint from the ornaments
- Separate the papyrus fragments
- Analyze the text : agricultural production, tax policy ...



# Papyrologist work

## Resolve a complexe puzzle:

- Missing fragments
- No continuity between contours
- Fragments used for funeral ornaments come from different papyri <sup>1</sup>



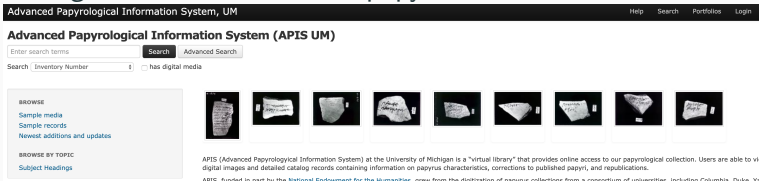
- 1 papyrus
- 12 fragments

<sup>1</sup>image from <https://quod.lib.umich.edu/a/apis> Accessed: June 04, 2019

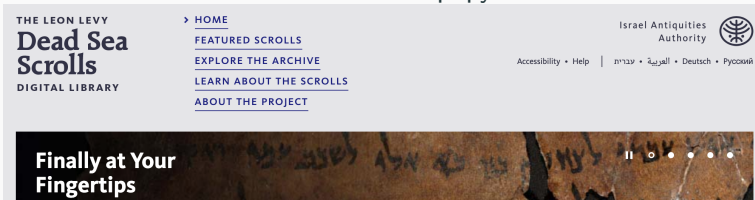


# A common objective for many projects

- *GESHAEM* project (4 years) : 500 fragments to reconstruct
- *Michigan Collection* : 26.000 papyri

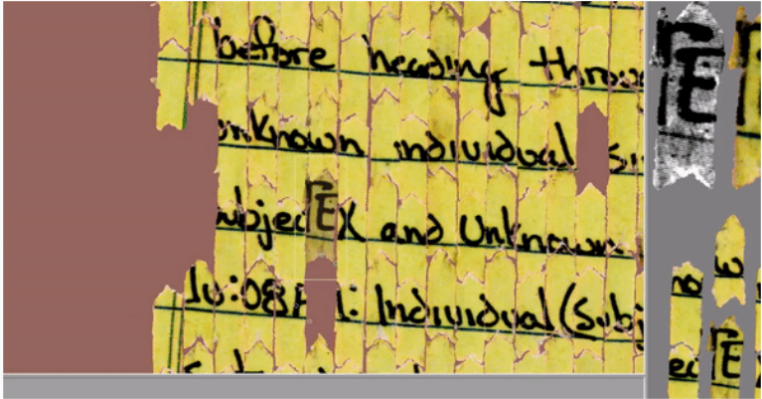


- *Dead Sea Scrolls Collection* : 2000 papyri



## Related Work

- Mainly methods for recovering shredded documents (Butler et al. 2012)

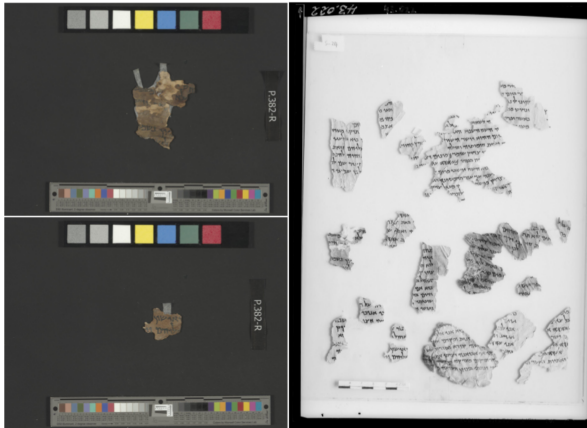


- Optimization problem (text/shape/color continuity)
- Crowd sourcing problem

# Related Work

## For Papyrus

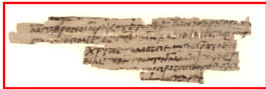
- Improve the digitalization process
- Identify duplicated fragments (Levi et al. 2018)



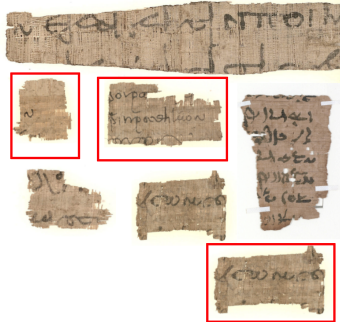
# Our proposal

## Helping the papyrologists to sort the fragments <sup>2</sup>

Request fragment

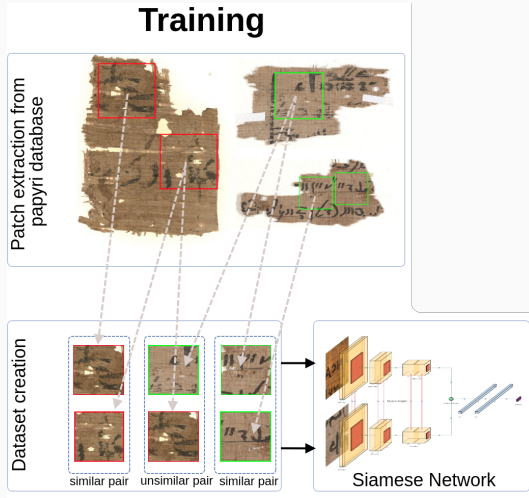


Get matching fragments within a papyrus database



<sup>2</sup>image from <https://quod.lib.umich.edu/a/apis> Accessed: June 04, 2019

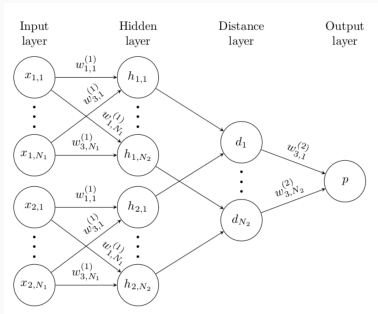
# Our proposal : a supervised learning approach



- Learning with patches
- Building similar and dissimilar pairs

# Siamese networks

- Two identical sub-networks with shared weights
- Defining a similarity metric to know if two inputs are similar or not

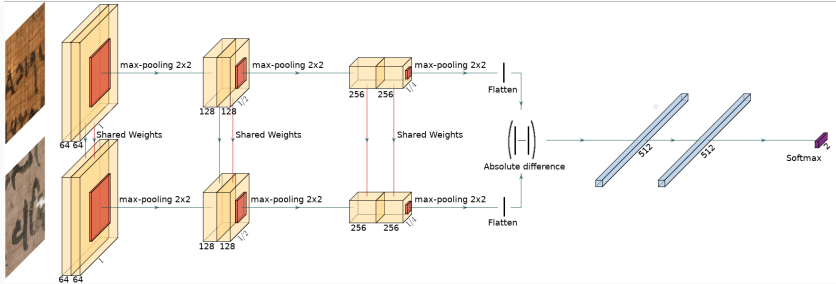


- Used for face recognition,
- Symbol/words matching (Koch et al. 2015, Zhong et al. 2016)



Queries	Top-5 results:					
<i>Letters</i>	<i>Letters</i>	<i>Letters</i>	<i>Letters</i>	<i>Letters</i>	<i>Orders</i>	
<i>and</i>	<i>and</i>	<i>and</i>	<i>and</i>	<i>and</i>	<i>and</i>	
<i>che</i>	<i>che</i>	<i>che</i>	<i>che</i>	<i>che</i>	<i>che</i>	
<i>a</i>	<i>a</i>	<i>a</i>	<i>a</i>	<i>a</i>	<i>as</i>	
<i>Winchester</i>	<i>Winchester</i>	<i>Winchester</i>	<i>Winchester</i>	<i>Winchester</i>	<i>Nicholas</i>	<i>Winchester</i>

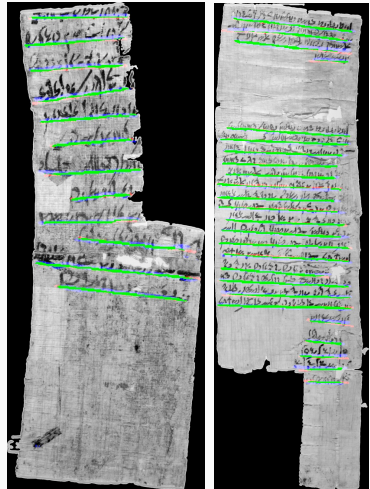
# Papy-S-Net



- Convolution / Pooling → get texture / text shape
- Absolute difference of the outputs of the branches
- Fully connected network for classification
- Softmax output

# Patch extraction

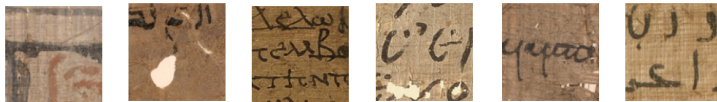
- Heuristic to avoid accepting patches with too much background
- Baseline segmentation method **ARU-Net** (Grüning 2018)
- Re-trained on our dataset
- Quite good segmentation results





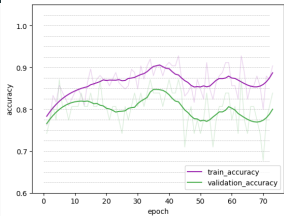
# Learning process

- 500 fragments <sup>3</sup> :
  - -600 to +400 BCE
  - 7 languages
  - From 1000x1000 to 6000x6000 pixels
  - From different digitalization campaigns
- 12.000 extracted patches
- {Train / Validation} 90% / {Test} 10 %
- {Train} 80% {Validation} 20%

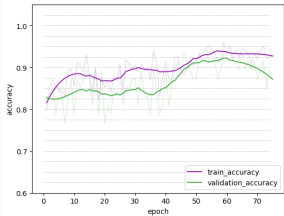


<sup>3</sup> coming from <https://quod.lib.umich.edu/a/apis> Accessed: June 04, 2019

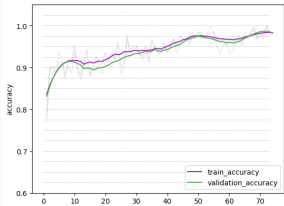
# Learning process - (Training/Validation) on 90% of the dataset



1. Patches containing only texture



2. Random patches



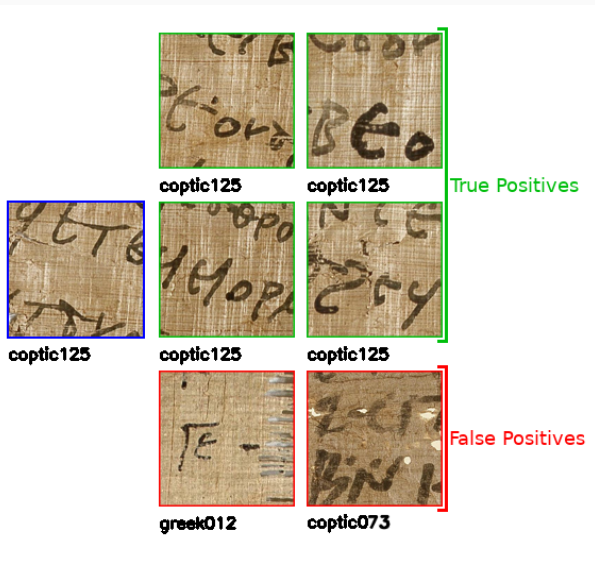
3. Patches all containing text

## Testing Papy-S-Net - on 10% of the dataset

Rates	Random		Without Text		With Text	
	PS-Net	Koch	PS-Net	Koch	PS-Net	Koch
True Pos.	0.80	0.74	0.75	0.76	<b>0.82</b>	0.72
False Pos.	0.09	0.12	0.08	0.13	<b>0.06</b>	0.14
True Neg.	0.92	0.88	0.82	0.87	<b>0.94</b>	0.86
False Neg.	0.20	0.26	0.25	0.24	<b>0.18</b>	0.28

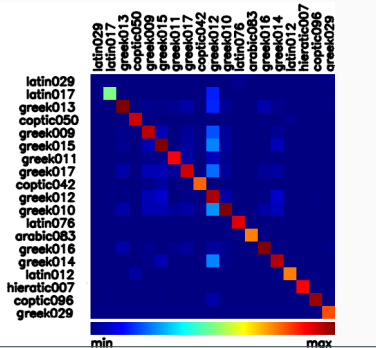
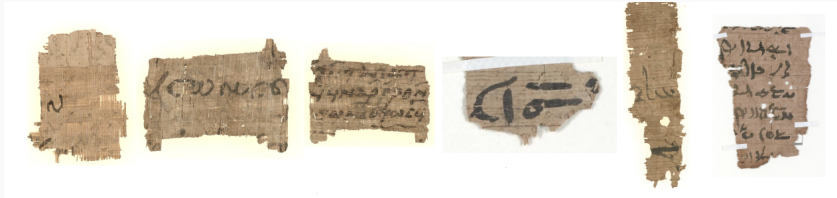
- Best results with **Papy-S-Net** on patches *With text*
  - but imply a preprocessing phase
- *Random* results quite close, but less guarantees : depends on the proportion of text in the fragments

# Examples of matchings



# Testing a real use case

About 60 fragments from 19 papyri to reconstruct<sup>4</sup>



- 79% True Positives
- 17% False Positives

<sup>4</sup>image from <https://quod.lib.umich.edu/a/apis> Accessed: June 04, 2019

## Conclusion and Futur works

- **Papy-S-Net** is a good first step towards more advanced works
- Can already be used to perform a significant filtering

### Future works

- Experiments on *GESHAEM* database, need for fine tuning ?
- Better results with data augmentation at patch level ?
- Language detection
- Papyri content classification (caution note, fiscal registers, administrative letters, sells contracts etc.)
- Actually resolving the puzzle



P. Butler, P. Chakraborty, and N. Ramakrishan.

**The deshredder: A visual analytic approach to reconstructing shredded documents.**

*In 2012 IEEE Conference on Visual Analytics Science and Technology (VAST),* pages 113–122. IEEE, 2012.



T. Grüning, G. Leifert, T. Strauß, and R. Labahn.

**A two-stage method for text line detection in historical documents.**

*arXiv preprint arXiv:1802.03345,* 2018.



G. R. Koch.

**Siamese neural networks for one-shot image recognition.**  
2015.



G. Levi, P. Nisnevich, A. Ben-Shalom, N. Dershowitz, and  
L. Wolf.

**A method for segmentation, matching and alignment of  
dead sea scrolls.**

*In 2018 IEEE Winter Conference on Applications of Computer  
Vision (WACV), pages 208–217. IEEE, 2018.*



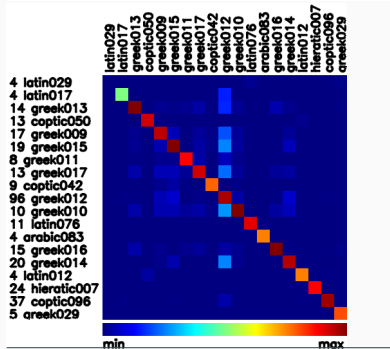
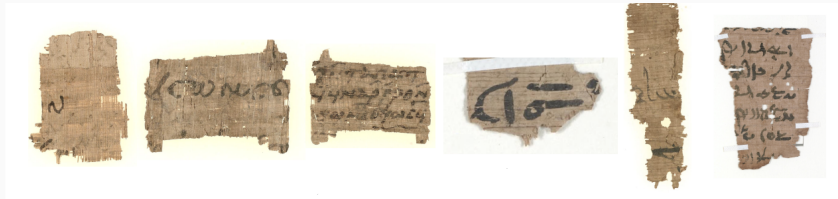


Z. Zhong, W. Pan, L. Jin, H. Mouchre, and C. Viard-Gaudin.  
**Spottingnet: Learning the similarity of word images with convolutional neural network for word spotting in handwritten historical documents.**

*In 2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR), pages 295–300, Oct 2016.*

# Testing a real use case

327 patches, bout 60 fragments from 19 papyri to reconstruct<sup>5</sup>



- Cases of the matrix → number of patches classified as coming from each papyrus
- 79% True Positives
- 17% False Positives

<sup>5</sup> image from <https://quod.lib.umich.edu/a/apis> Accessed: June 04, 2019