Modélisation des processus de navigation dans une bibliothèque numérique

R. Champagnat - M. Trabelsi - C. Suire - J. Morcos

Maître de Conférences HDR

Université de La Rochelle - L3i

L3i - équipe Dynamique des Systèmes et Adaptativité (eAdapt)

- 6 permanents + 12 non-permanents + 1 membre associé
- Thème : auto-adaptation des systèmes et services numériques
  - Pilotage auto-adaptatif des systèmes en réseaux avec calculs intégrés : vers des infrastructures numériques intelligentes, éco-efficientes et sécurisées
  - Pilotage auto-adaptatif des services numériques du futur : vers des approches centrées sur l'humain
  - Blockchains éco-efficiente basée sur la confiance : approches décentralisées de partage des connaissances et de protection des données
- Fouille de processus : 3 maîtres de conférences, 3 thèses en cours (1 soutenue), 1 postdoctorant

Objectif

Objectif

Utiliser la fouille de processus pour analyser des services numériques

Question de recherche

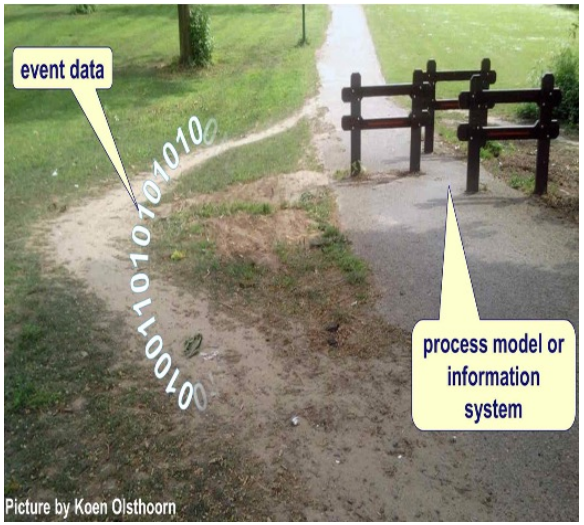
Peut-on extraire des information intéressantes sur les utilisateurs de services numériques avec des processus faiblement structurés en utilisant la fouille de processus ?

Fouille de processus

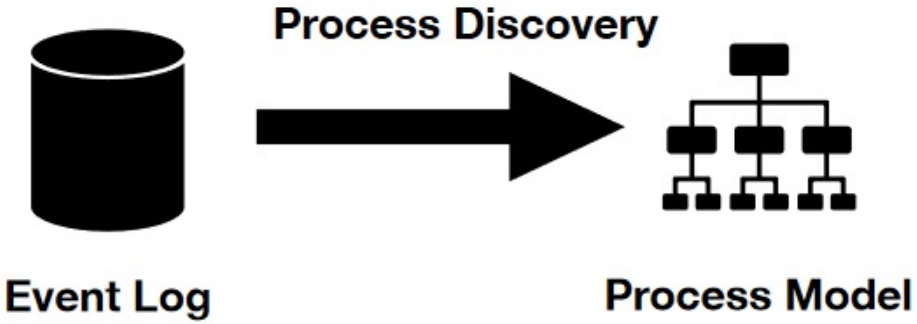
- Fouille de processus : extraire de la connaissance sur les processus métiers d'une organisation à partir des traces d'exécution
- Utilisation :
  - Définir un modèle des processus d'usage
  - Rejouer une exécution pour déterminer sa conformité
- Travaux
  - Identification des parcours utilisateurs dans une bibliothèque numérique
  - Recommandation de ressources pédagogiques pour une formation à distance
  - Détecter des comportements anormaux dans le cadre de la protection de la vie privée

Principes de la fouille de processus I

- Extraire les processus métiers à partir des logs
- Combiner la fouille de données et la modélisation des processus métiers



Principes de la fouille de processus II

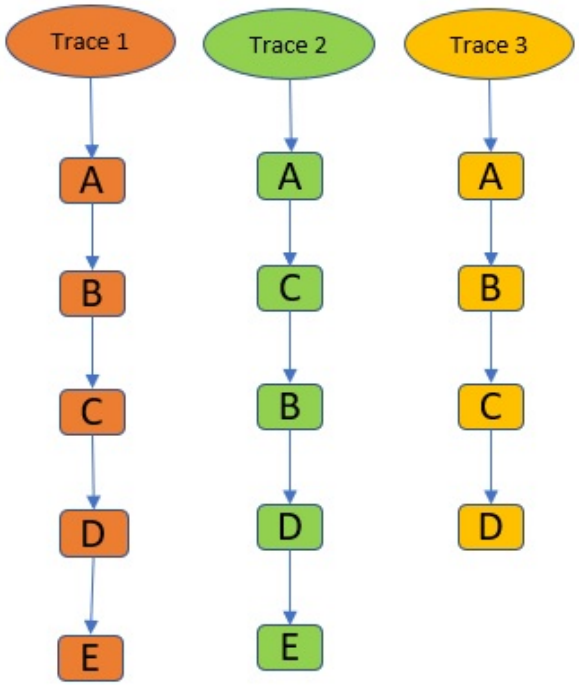


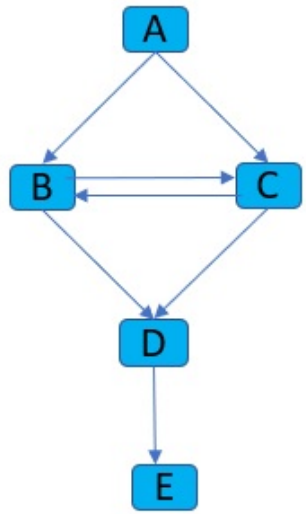
Terminology

- **Event logs** : an event logs $L = \{t_1, t_2, ..., t_k\}$ is a set of k **traces**
- **Trace** : each trace t_i $(1 \leq i \leq k)$ is a set of n_i consecutive events $t_i =< e_{i1}, e_{i2}, ...e_{in_i} >$ made by the same CaseID.
- **Event** : each event e is characterised by its frequency f_e which is the number of times e occurs in all the traces.

Principes de la fouille de processus III

Case id	Event
1	A
2	A
1	B
1	C
3	A
2	C
3	B
2	B
1	D
2	D
2	E
3	C
3	D
1	E





Principes de la fouille de processus IV

- ***Direct succession*** : **x > y**
- ***Causality*** : **x -> y**, if ***x>y*** *and not* ***y>x***
- ***Parallel*** : **x || y**, if both ***x > y and y > x***
- **Choice** : ***x # y*** , if never ***x>y*** *nor* ***y>x***

Case 1	Task A
Case 2	Task A
Case 3	Task A
Case 3	Task B
Case 1	Task B
Case 1	Task C
Case 2	Task C
Case 4	Task A
Case 2	Task B
Case 2	Task D
Case 5	Task A
Case 4	Task C
Case 1	Task D
Case 3	Task C
Case 3	Task C
Case 4	Task B
Case 4	Task D
Case 5	Task E
Case 5	Task D

Causality :
A -> B
A -> C
A -> E
C -> D
B -> D
E -> D

Direct succession :
A>B
B>C
C>B
E>D
A>C
C>D
B>D
A>E

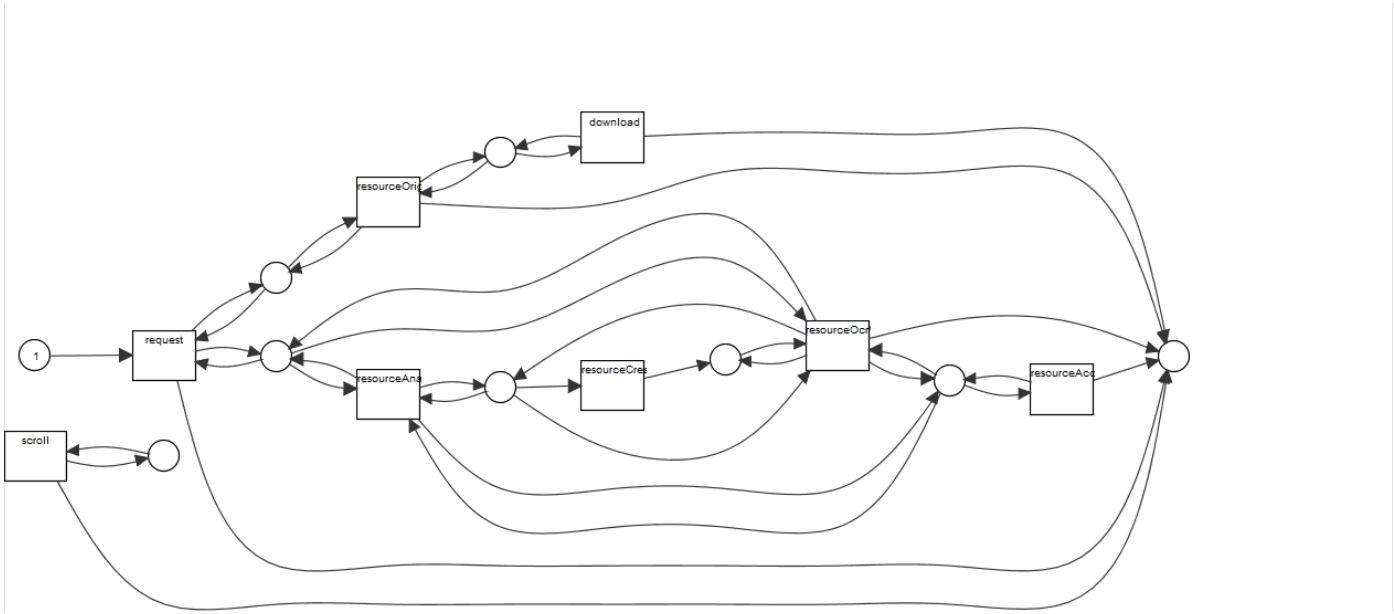
Parallel :
B||C
C||B

Choice :
A # D
B # E
C # E
E # B
D # A
E # C

w = [<A,B,C,D>, <A,C,B,D>, <A,E,D>]

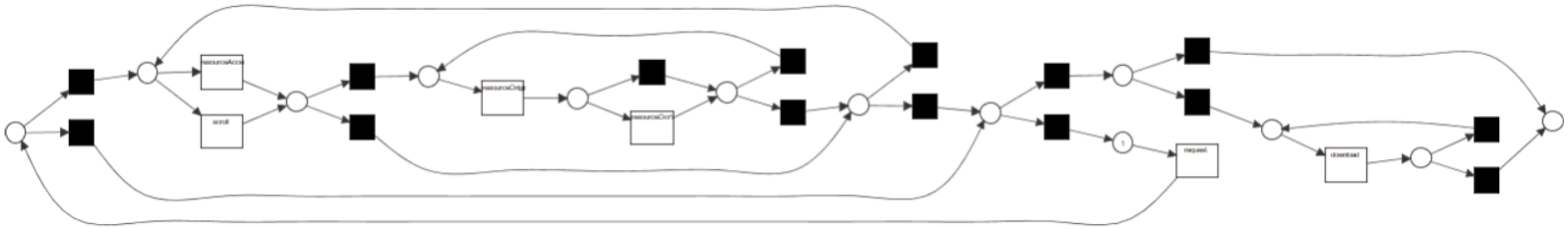
Utilisation pour analyser les usages de Gallica I

- Problématique : extraire des parcours utilisateurs d'un SI possédant peu de processus métiers prédéfinis
- Pourquoi les bibliothèques numériques ?
- Performance des algorithmes existants
  - Alpha Miner

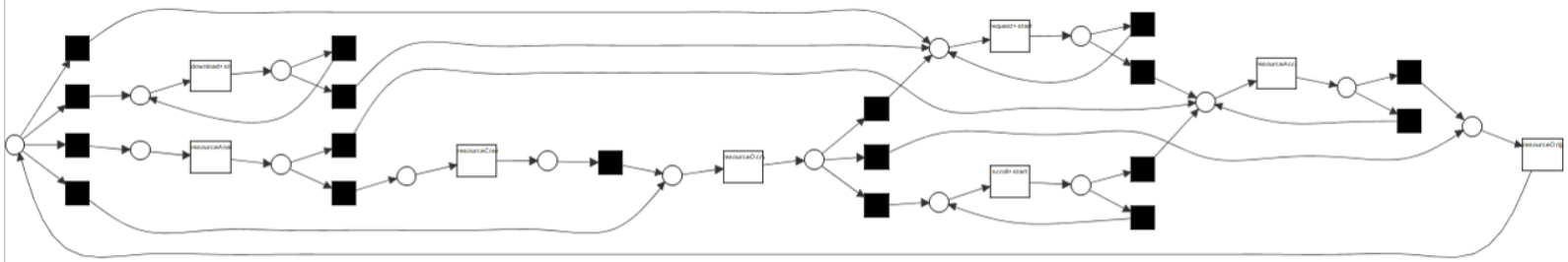


Utilisation pour analyser les usages de Gallica II

- Inductive Miner

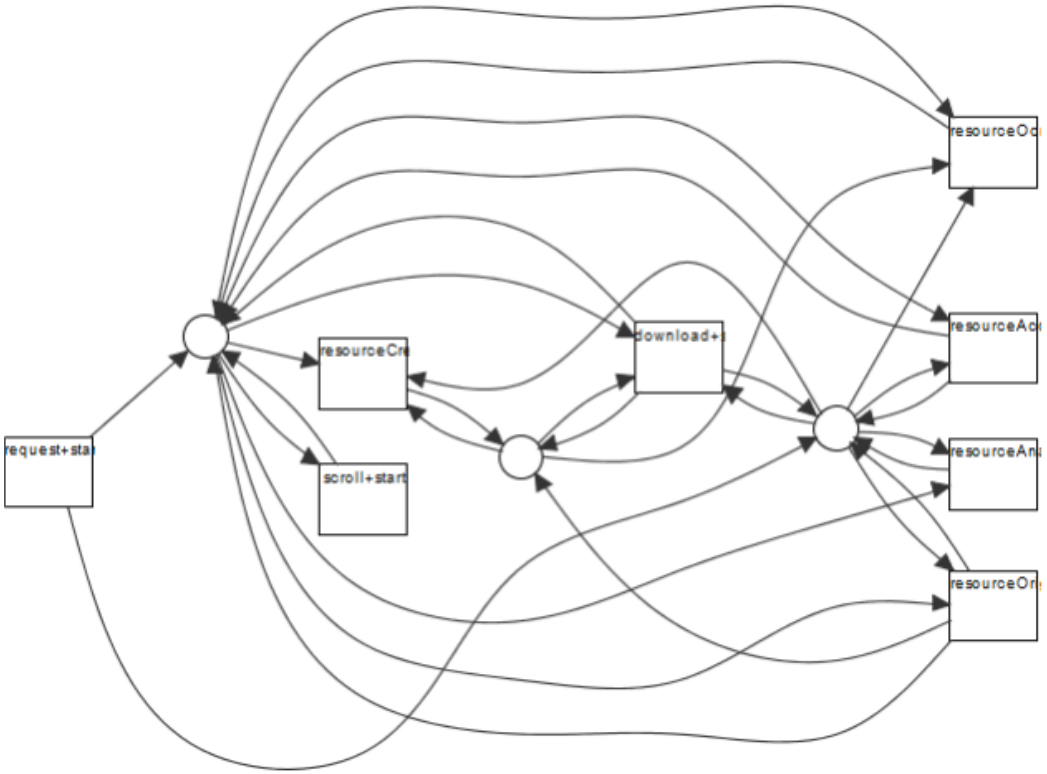


- Heuristic Miner



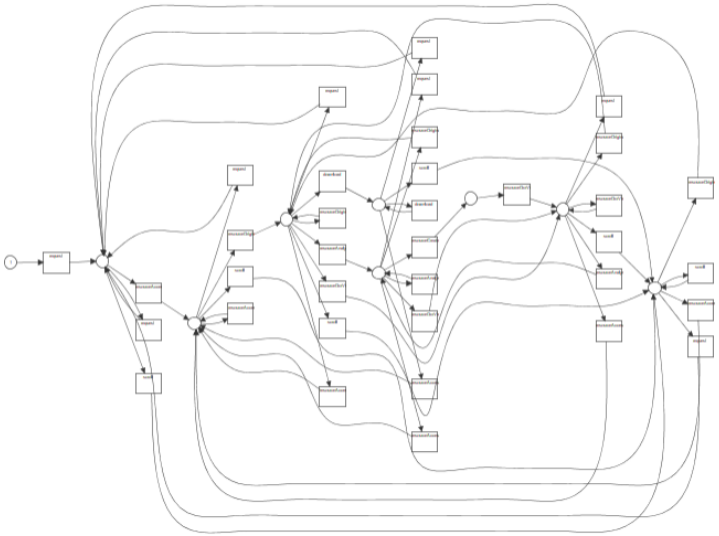
Utilisation pour analyser les usages de Gallica III

- Language Based Regions



Utilisation pour analyser les usages de Gallica IV

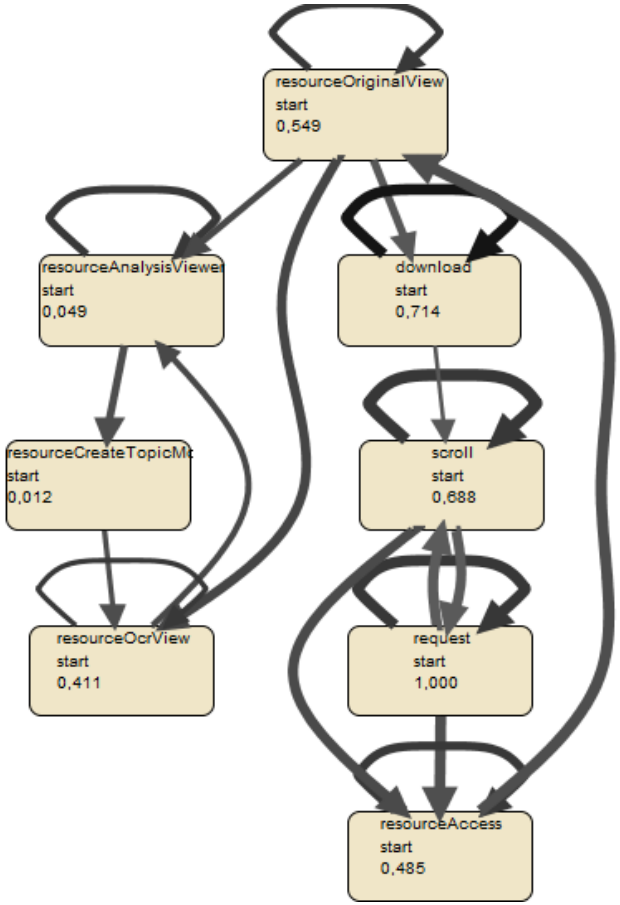
- State Based Regions



- Genetic Miner

Utilisation pour analyser les usages de Gallica V

- Fuzzy Miner



Utilisation pour analyser les usages de Gallica VI

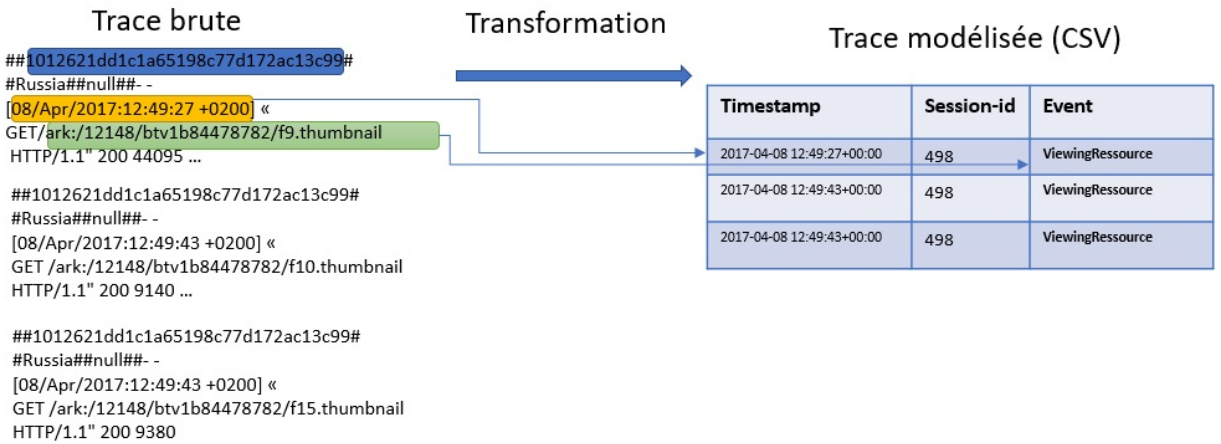
- Result

	Fitness	Precision	Generalisation	Fitness	Precision	Generalisation
Alpha ++	0	0	0	0	0	0
Inductive Miner	**0,9886**	0,2391	0,9994	**0,9315**	0,1437	0,9992
Heuristic Miner	0	0	0	0	0	0
Language Based Regions	0,6163	0,3825	0,9793	0,7835	0,919	0,9622
State Based Regions	0,8995	0,4233	0,9957	0,9560	0,2942	0,9918
Genetic Miner	0,9232	0,1800	0,9939	0,6232	0,8053	0,9963
Fuzzy Miner	--	--	--	--	--	--

Utilisation pour analyser les usages de Gallica VII

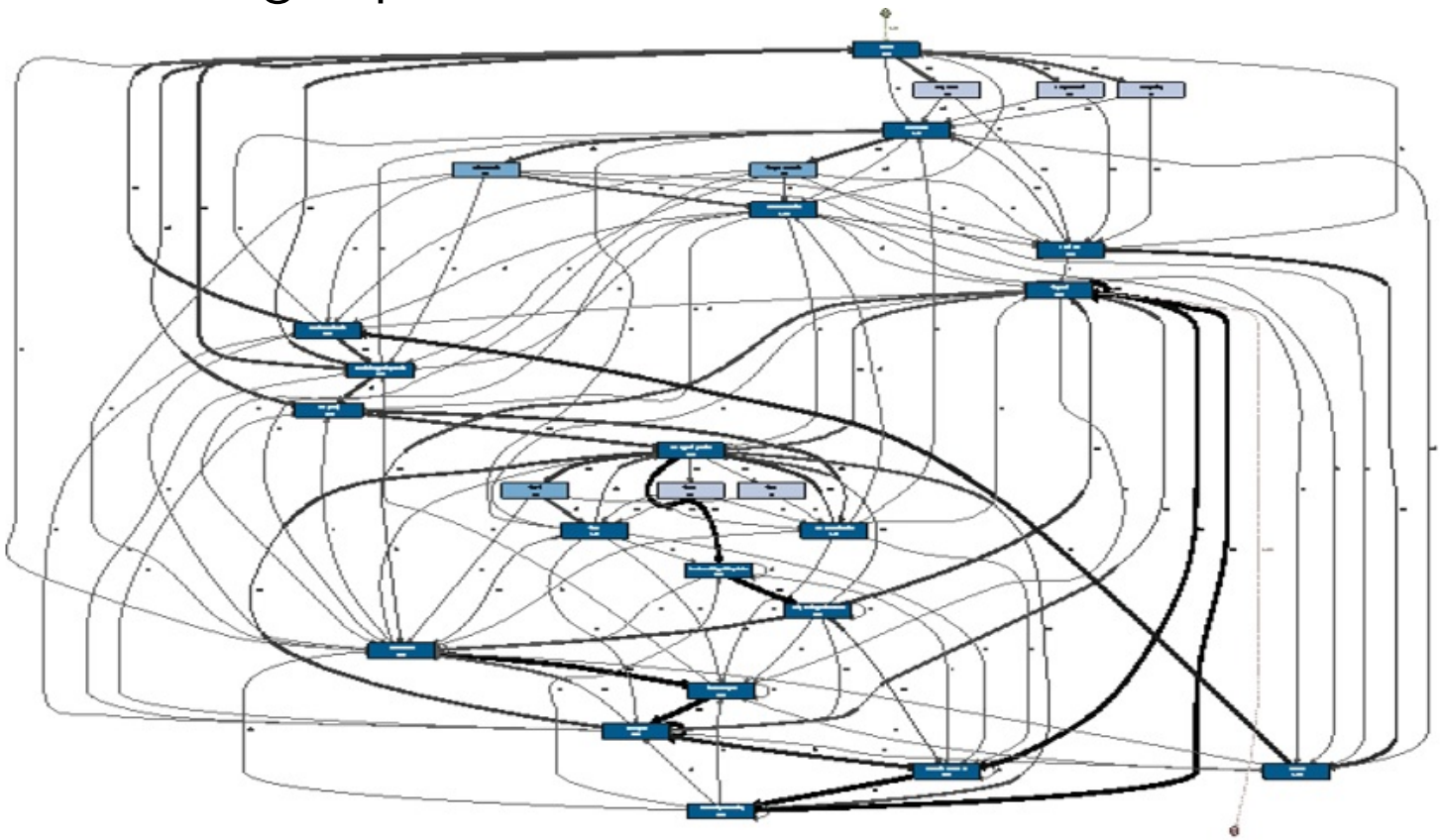
- Les logs réels ne sont pas directement exploitables

```
##1a47161ad98134bf072fe5ea3573fca6##Japan##Tokyo##- -
[10/Apr/2017:07:52:14 +0200] "GET /accueil/?mode=desktop
HTTP/1.1" 200 11311 "-" "Mozilla/5.0 (Macintosh; Intel Mac
OS X 10_12_4) AppleWebKit/603.1.30 (KHTML, like Gecko)
Version/10.1 Safari/603.1.30" "JSESSIONID=AD97; xtidc=1605;
xtan18798=-; xtant18798=1; rxVisitor=1479; xtvrn=$18798$"
```



Utilisation pour analyser les usages de Gallica VIII

- Besoin de regrouper



Spaghetti models $\Longrightarrow$ Clustering

Utilisation pour analyser les usages de Gallica IX

Feature-based Similarity

Converting each trace into a vector of features based on defined characteristics.

Trace-based Similarity

Similarity between two traces can be measured using the syntax similarity.

Model-based Similarity

Process models are considered as input for the clustering in order to structure traces.

- Handling the exponential number of events in DL logs.

Utilisation pour analyser les usages de Gallica X

- Convert traces into vectors (trace clustering => vector clustering)
- Frequencies and relations are not sufficient to encode well a DL user trace
- Frequent Sub-Sequences (FSS) in the traces can contribute to distinguish users and tasks.
- Traces containing FSSs are the most significant traces.
- Converting traces using a particular FSS encoding.
- Each identified FSS in a trace is replaced by its encoding.

Utilisation pour analyser les usages de Gallica XI

- **The length of the FSS** : more the lengh n is higher, more the FSS is the most significant.
- **The frequency of the FSS** : the FSS with the highest frequency f is the is the most important.
- **The frequency of events in the FSS** : the difference between two FSSs with same frequency and lengh is underlined by this factor.
- **The direct succession relation between events in the FSS** : redundant relations between two events in a FSS may indicate important criteria for clustering such as backtracks...

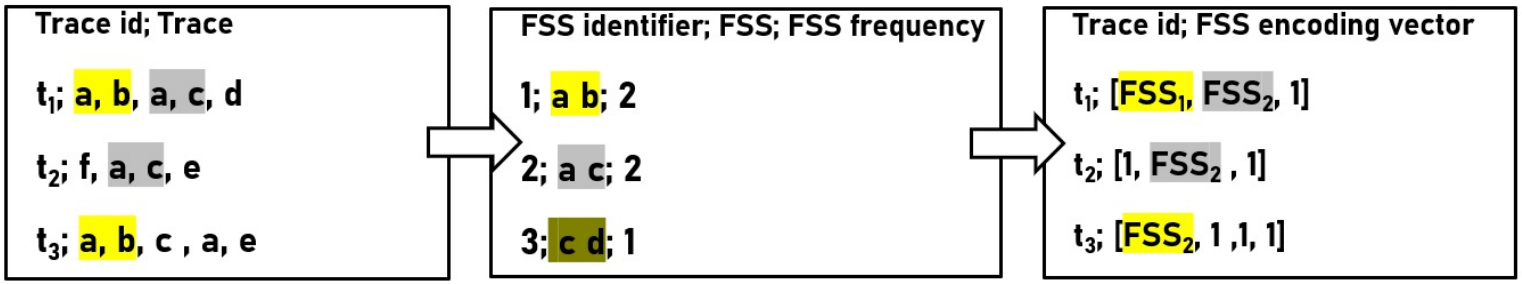
Utilisation pour analyser les usages de Gallica XII

$$Encoding(FSS) = \frac{1}{f_{FSS} \sum_{i=1}^{n-1} f_{e_i} f_{e_{i+1}} f_{r_{i,i+1}}}$$

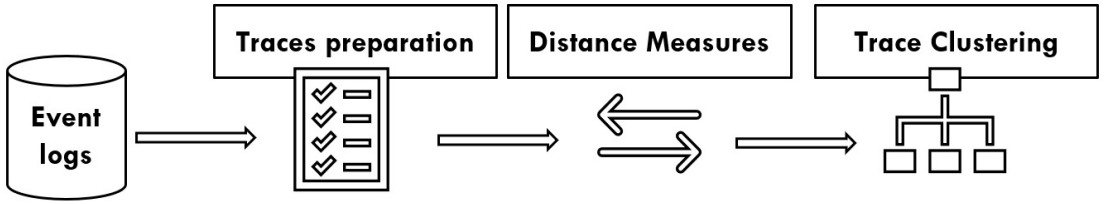
- f_{FSS} is the frequency of the FSS
- n is its length
- f_{e_i} is the frequency of the event
- $f_{r_{i,i+1}}$ is the frequency of the direct relation between events
- FSS Encoding value = [0 - 1]

Utilisation pour analyser les usages de Gallica XIII

- Other events which do not belong to an FSS are considered irrelevant and only their positions account in the clustering.
- Replace such events in the traces by 1.
- Distinguish traces with the same FSS not in the same position



Utilisation pour analyser les usages de Gallica XIV



Utilisation pour analyser les usages de Gallica XV

Clustering evaluation metrics

- **Silhouette (S)** : the separation distance between the generated clusters
- **Davies-Bouldin (D)** : evaluates the intra-cluster similarity and the inter-cluster differences

Utilisation pour analyser les usages de Gallica XVI

Process modeling evaluation metrics

- **Fitness (F)** : determines how well the process model covers the event logs.
- **Precision (P)** : measures whether the process model allows only the behaviour observed in the event logs.
- **Generalization (G)** : measures the ability of the model to generalize the behavior seen in the logs.
- **F-measure** : the trade-off between the Fitness and Precision.

- Clustering level : a labeled dataset assigning a class to each user's trace is required.
- No such data is openly available.
- DLs users' search behaviors [1].

Utilisation pour analyser les usages de Gallica XVII

Lookup traces

- access precisely identified documents with few manipulations
- 40 traces, 140 events, 10 types

Borderline traces

- access documents within a well-defined subject area
- 30 traces, 170 events, 6 types

Exploratory traces

- access to a wide range of documents in different fields and of different types
- 30 traces, 310 events, 5 types

Utilisation pour analyser les usages de Gallica XVIII

Experimental results

- Retrieve the three desired DL users groups described in the simulated data.
- Two baselines inspired from existing clustering methods were implemented.
- **freq-based** method : each event e in the traces will be replaced by its frequency f_e in all the event logs.
- **relation-based** method : consists in the counting of the direct succession relations between events r in all the event logs.

Utilisation pour analyser les usages de Gallica XIX

	freq-based	relation-based	FSS-based
Silhouette	0,731	0,473	**0,817**
Davies-Bouldin	0,282	1,178	**0,238**
Clusters	5 clusters	3 clusters	**3 clusters**
Silhouette	0,730	0,423	0
Davies-Bouldin	0,412	1,178	0
Clusters	2	2 (+50 noises)	1

- The Meanshift algorithm outperforms DBSCAN
- Relation-based & FSS-based methods allow us to obtain the three clusters
- The FSS-based method achieves better results on both D and S

Utilisation pour analyser les usages de Gallica XX

	Real clusters	relation-based	FSS-based
Fitness	1	1	1
Precision	0,5360	0,5965	**0,6345**
F-measure	0.7000	0.7500	**0.7800**
Generalization	0.6719	0.2697	0.4533

- Results on the models obtained from the Meanshift clusters.
- FSS-based approach improves the Precision and F-measure values

Utilisation pour analyser les usages de Gallica XXI

Data visualization

- Global view of users' queries (500 M monthly).
- ELK services to visualise queries.
- Time stamp to filter our dataset (April 2017).

Outliers Detection

- Deleting irrelevant queries mainly related to the design of Gallica
- Deleting queries related to the bots-crawlers
- Deleting $\sim$ 60% of the queries

Users' Queries Tagging

- Standard convention to tag queries.
- Each query is normalised using standard action's name.

Sessionization

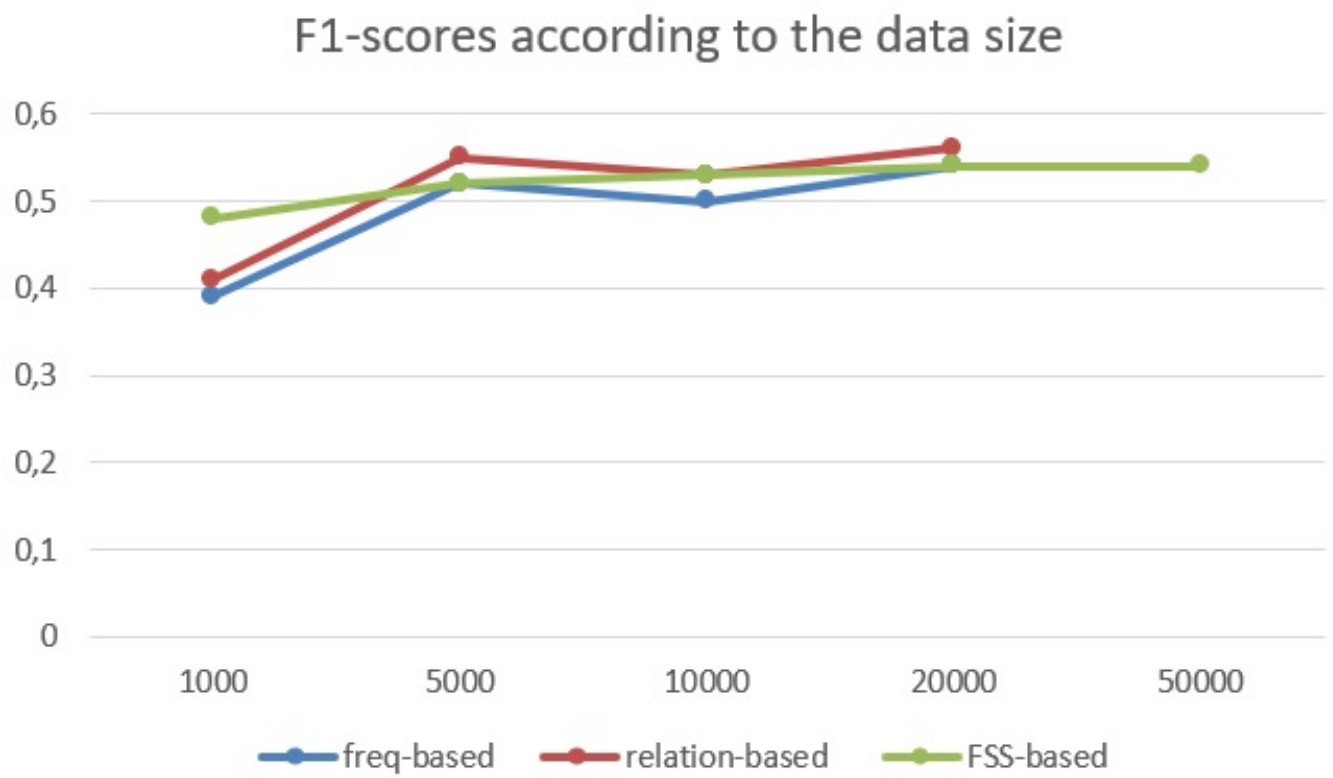
- Dividing all the users' queries into sessions.
- Defining a session as a navigation sharing the same IP address and does not exceed 60 minutes as

Utilisation pour analyser les usages de Gallica XXII

Total data duration	January 2016 - April 2017
Used data	April 2017
Number of queries	476,057,043
Number of queries after filtering	102,787,467
Number of traces	1,719,657

- Modeling such event logs is time consuming and sometimes out of reality.
- The clustering would propose a high number of clusters which is far from the easiness to be considered by DLs designers

Utilisation pour analyser les usages de Gallica XXIII

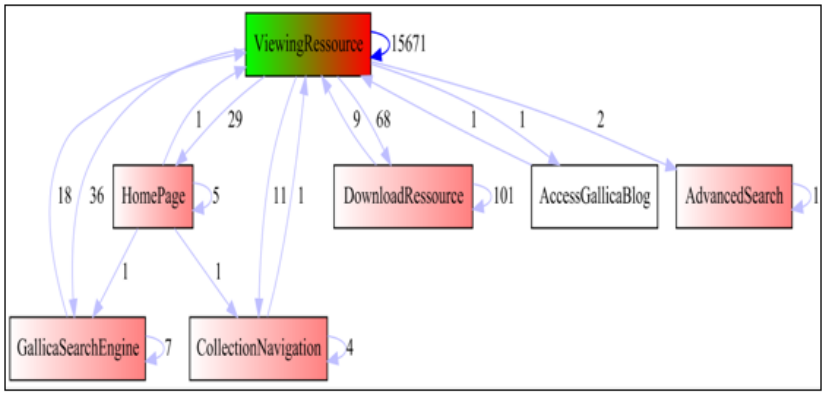


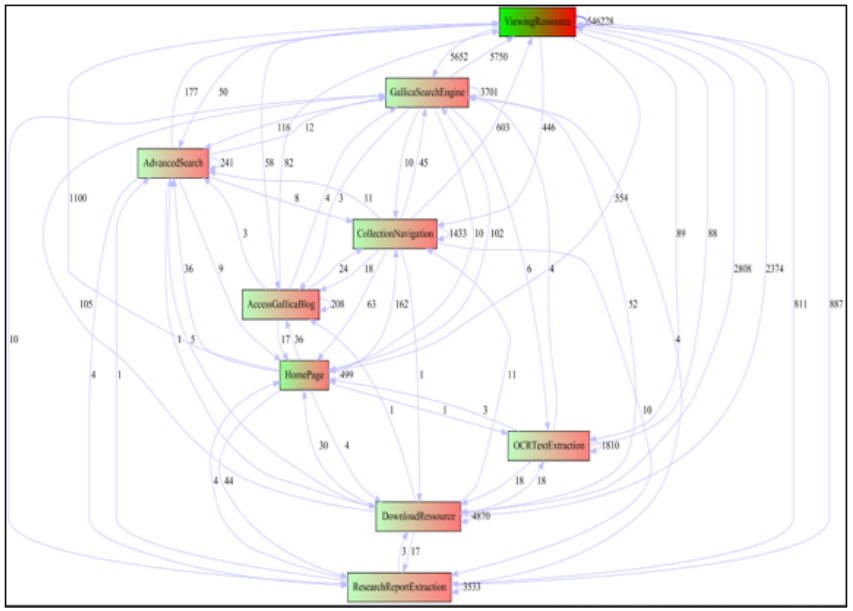
Utilisation pour analyser les usages de Gallica XXIV

		1000	5000	10000	20000	50000
freq-based	NB Clusters	2	3	2	2	–
	Silouhette	**0,564**	**0,547**	**0,547**	**0,551**	–
	Davies-bouldin	**0,495**	0,756	0.753	0.746	–
	Execution time	692 s	55248 s	316318 s	930880 s	–
	Execution time	11m32s	15h20m48s	87h51m58s	258h34m40s	–
relation-based	NB Clusters	2	2	3	2	–
	Silouhette	0,459	0,449	0,446	0,444	–
	Davies-bouldin	1,013	1,042	1,046	1,053	–
	Execution time	**347** s	16484 s	80299 s	420780 s	–
	Execution time	5m47s	4h34m44s	22h18m19s	116h53m	–
FSS-based	NB Clusters	2	2	2	2	2
	Silouhette	0,467	0,461	0,459	0,575	0,459
	Davies-bouldin	0,560	**0,573**	**0,579**	**0,582**	**0,581**
	Execution time	376 s	**15021** s	**66288** s	**278021** s	1582699 s
	Execution time	6m16s	4h10m21s	18h24m48s	77h13m41s	439h38m19s

TABLE 6.1 – Évaluation du *clustering* sur les n premières traces

Utilisation pour analyser les usages de Gallica XXV





Discovered process models using the FSS-based method for 20,000 traces with up to 1,000 events.

1. based on `https://projectblacklight.org/`

Conclusion

- Utilisation de la fouille de processus pour extraire les modèles de parcours d'usage des utilisateurs de Gallica
- Définition d'une nouvelle méthode de clustering
- Capable de traiter des gros volumes de traces
- Prochain problème : identifier des processus émergents