

# Représenter l'information historique

FREDERIC KAPLAN  
DIGITAL HUMANITIES LABORATORY



digital  
**humanities**  
Lausanne – Switzerland '14



**80 km d'archives  
documentant tous les  
détails de la vie  
quotidienne à Venise  
sur 1000 ans.**



**Nous nous sommes  
donné 10 ans pour  
transformer cette  
archive en un système  
d'information géant.**

**Rythme de  
numérisation :  
450 volumes / jour**

**100 Terabytes**



**Une fois les documents numérisés, ils doivent être transcrits. Les techniques d'OCR ne sont pas immédiatement fonctionnelles pour ce type de document.**

D V C A N T E S E R E N I S S I M O A C I N C L Y T O D V C E D  
LEONARDO LAUREDANO DEI G. V & C. S.

Die secunda Martij .1513.

Cum de mense Januarij procreto per quosdam incognitos facta per eos seditione  
et aduersionem plurimum barcharonum et hominum accesserint sepius nocturne  
ad vallem di Bombai ducatus venetiarum et uiolenter fractis clausuris dictae  
plurima damna et uiolencias in ipsa valle commiserunt: piscando per illam  
aufferendo: aggressum faciendo cum armis uersus custodes dictae uallis  
fugando: prorumpendo et in crudelissimas blasphemias omnipotentis dei  
se uirginis Maris eius Matris: cum domino et salubra plurima fructu  
uictoris dognolo di Iusti conductorum dictae uallis: & faciat pro debito  
uenire in lucem et notitiam predicatorum in alessandria: Idcirco uideat po  
quandorante huius consilij proclamet super scales primo alibi et alibi ubi  
int Aduocatoribus communi: & qui accusabit malefactores ipsos Aduocato  
sic per eius accusam ueritas habeat: Habeat ipse accusator libras mille  
nove secretus: et si per eius accusam malefactores uenerint in fortunas  
Huius habeat ipse accusator libras mille quingentas et teneat secretus  
ut dictum est fuerunt plures uno ad ipsos excessus committendos: & in  
Captum: & si unus eorum accusabit socios culpabiles Aduocatoribus co  
per eius accusam habeat ueritas: habeat ipse accusator predictas libras  
Teneat secretus et sit absintus non solum a parti delicto: uerum et  
alia delicto furti commisso per eum habitus non manifestato: et in lucem  
ternemo: et si per eius accusam socij culpabiles uenerint in fortunas

**Il y a de grandes  
variétés d'écritures et  
de langues sur 1000  
ans.**

**Plusieurs méthodes  
combinant la  
linguistique, les  
sources historiques  
peuvent être  
envisagées.**

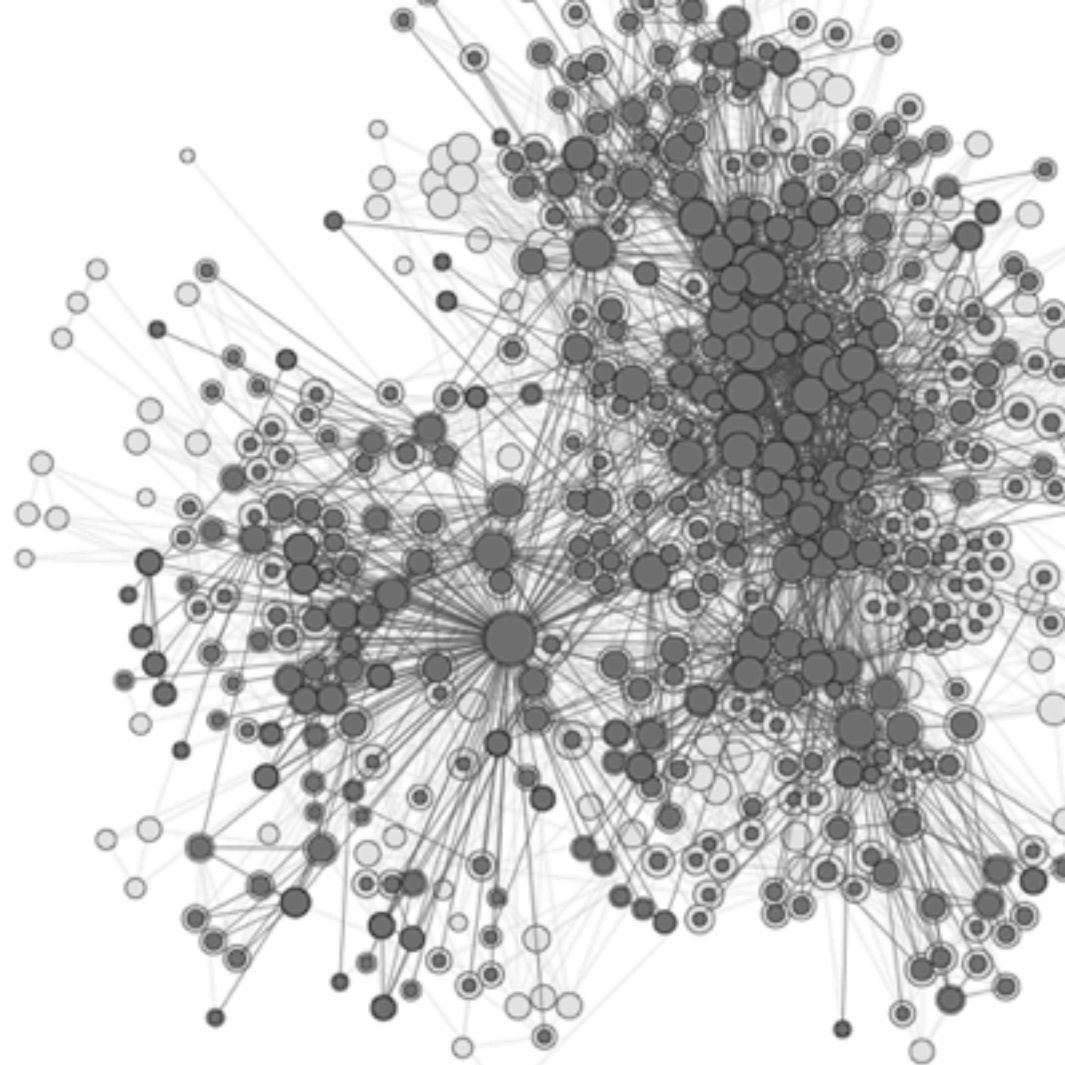
D V C A N T E S E R E N I S S I M O A C I N C L Y T O D V C E D  
L E O N A R D O L A V R E D A N O D E I . G . V & C E T

Die' secunda Martij . 1513.

Cum de mense Januarij preterito per quosdam incognitos facta per eos seditione  
et aduersionem plurimum barcharonum et hominum accesserint sepius nocturne  
ad vallem di Bombai ducatus venetiay et uiolenter fractis clausuris dis  
plurima damna et uiolencias in ipsa valle commiserunt; piscande per illam  
aufferendo aggressum faciendo cum armis uersus custodes dicty uallis  
fugando: prorumpendo et in crudelissimas blasfemias omnipotentis dei  
se virginis Mariæ eius Matris cum damno et salute plurima fratri  
uictoris dognolo di Iusti conductorum dicty uallis: & faciat pro debito  
uenire in lucem et notitiam predicatorum in alefactorum: Idcirco uideat po  
quandorunt huius consilij proclamet super scales primo alibi et alibi ubi  
int Aduocatoribus communi: & qui accusabit malefactores ipsos Aduocu  
sic p per eius accusam ueritas habeat: Habeat ipse accusator libras mille  
meat secretus: et si per eius accusam malefactores uenerint in fortius  
Huius habeat ipse accusator libras mille quingentas et teneat secretus  
ut delictum est fuerunt plures uno ad ipsos excessus committendos: & m  
Captum: & si unus eorum accusabit socios culpabiles Aduocatoribus co  
per eius accusam habeat ueritas: habeat ipse accusator predictas libras  
Teneat secretus et sit absintus non solum a parti delicto: uerum et  
alia delicto furti commisso per eum habentis non manifestato et in luc  
ternemo: et si per eius accusam socij culpabiles uenerint in fortius

**À partir des  
transcriptions, nous  
pouvons extraire des  
informations  
sémantiques sur les  
personnes, les lieux,  
etc.**

**10 milliards de relations**

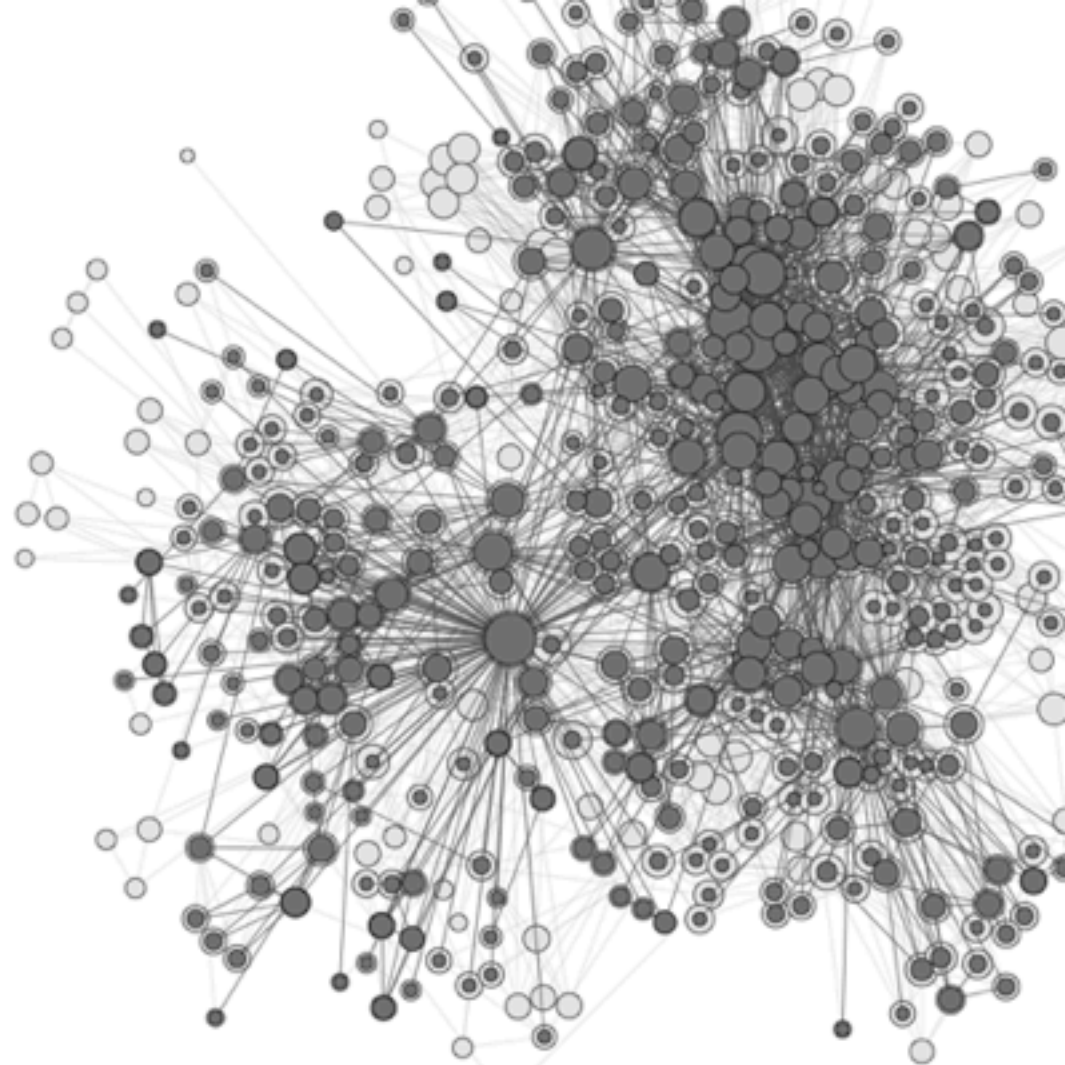




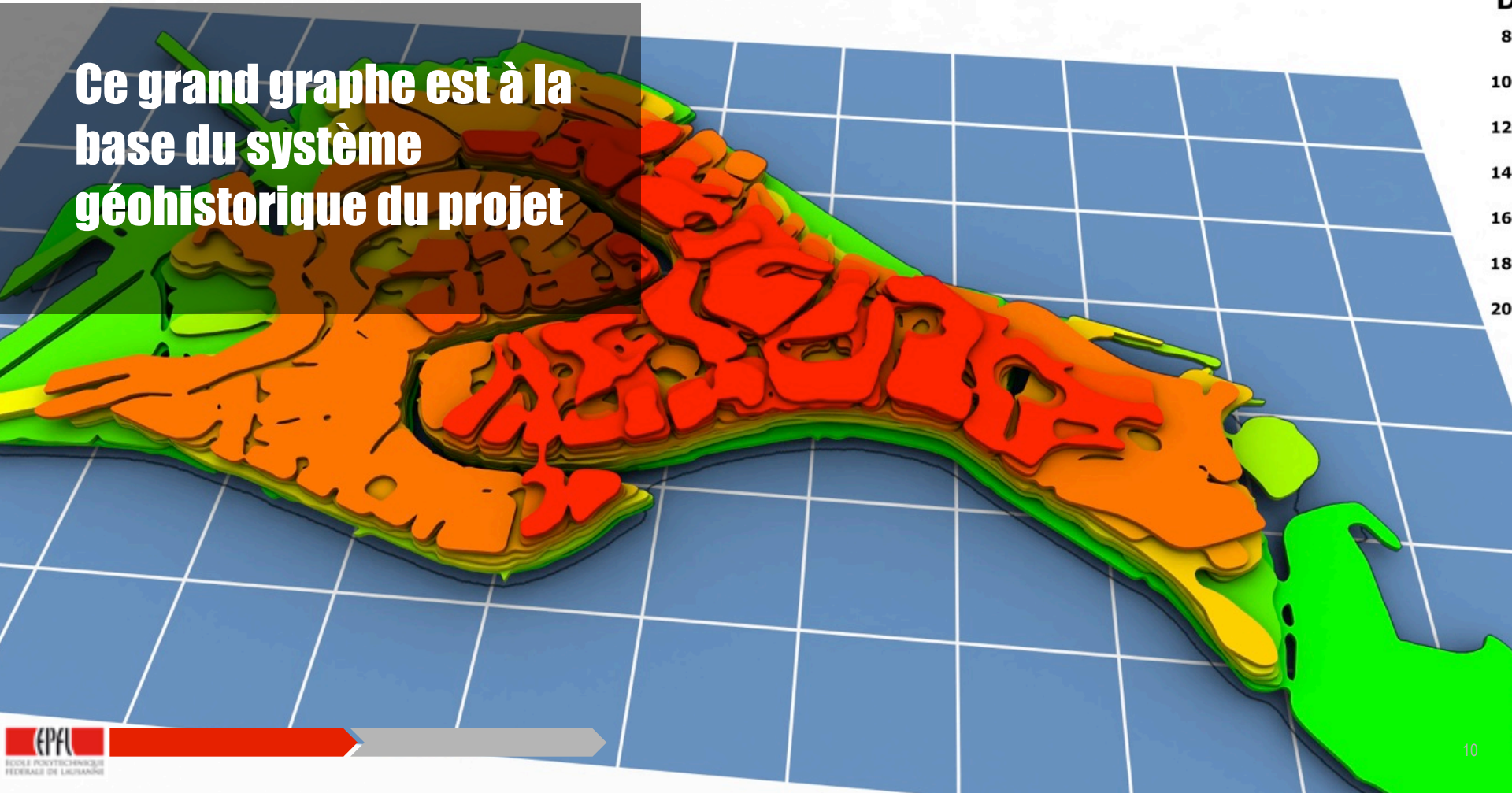
**Ce système  
d'information pourra  
être interrogé de  
multiples manières**

**Qui habitait dans ce  
palazzo en 1325 ?**

**Combien coutait la  
dorade au Rialto en  
1433 ?**



**Ce grand graphe est à la base du système géohistorique du projet**



Nous construisons avec Marc-Antoine Nüssli, un nouveau langage formel de description sémantique adapté à la recherche en histoire. Le nom provisoire de ce langage est **metaRDF**.

La connaissance historique est fondamentalement **incertaine**.  
Une vision particulière d'un événement historique se base sur un certain nombre de **sources** ainsi que sur des **interprétations** faites à partir de ces sources.

Généralement, le produit d'une recherche historique se présente sous la forme d'une **synthèse**, par exemple un récit ou une carte, et ne rend pas bien compte du processus intellectuel ayant conduit à sa création.

Notre approche de **coder la connaissance historique** tout en **documentant** l'ensemble des processus intellectuels qui conduisent des sources jusqu'à à cette connaissance.

Les langages de représentations de connaissance du web sémantique, **RDF** et **OWL**, ont un défaut majeur qui limite leur utilisation dans notre cadre.

Les connaissances exprimées  
(triplet RDF) ne sont pas des  
objets du même ordre que le  
contenu de ces connaissances  
(ressources RDF)



Ceci implique qu'il n'est pas directement possible de représenter des méta-connaissances, telles que la **provenance** ou **l'incertitude** d'une information.

Afin de pallier à ce problème,  
nous avons conçu un système à  
**deux niveaux** de connaissances.

Le premier niveau documente la provenance, la nature et la formalisation des connaissances historiques, le second niveau, les connaissances elles-mêmes.

Dans notre système, les informations de second niveau sont exprimées sous forme de triplets **RDF réifiés** alors que les informations de premier niveau sont modélisées sous forme de triplets **RDF normaux**.

# Expressed RDF

(RialtoReconstruction hasTimeSpan 1588-1591)

*subject*

*predicate*

*object*

# Reified RDF

(Statement134 rdf:subject RialtoReconstruction)

(Statement134 rdf:predicate hasTimeSpan)

(Statement134 rdf:object 1588-1591)

# Expressed RDF

(RialtoReconstruction hasTimeSpan 1588-1591)

*subject*

*predicate*

*object*

# Reified RDF

(Statement134 rdf:subject RialtoReconstruction)

(Statement134 rdf:predicate hasTimeSpan)

(Statement134 rdf:object 1588-1591)

(Statement134 metardf:reliability 0.8)

(Statement134 metardf:creator FrederickKaplan)

La base de connaissances ainsi créée contient donc des meta-connaissances historiques, c'est-à-dire **des connaissances sur la création de connaissances historiques.**

Ces meta-connaissances peuvent documenter le choix des sources, les phases de transcription, codage, interprétation que ces opérations soient réalisées par des **humains** ou des **machines**.



# Nous documentons ici comment le mot “Rialto” a été associé à RialtoBridge par un algorithme d'extraction sémantique

```
(Statement22 rdf:subject “Rialto”)
```

```
(Statement22 rdf:predicate Standsfor)
```

```
(Statement22 rdf:object RialtoBridge)
```

```
(Statement22 metardf:reliability 0.9)
```

```
(Statement22 metardf:method Wikipediacomparison)
```

Cette approche permet de créer des **espaces de connaissances** cohérents localement basés par exemple sur les mêmes sources ou les mêmes méthodes.

Par exemple, un **espace documentaire** décrivant des ressources bibliographiques utilisées pour un projet.

{d}

et un **espace fictionnel** associé décrivant les connaissances extraites de ces ressources bibliographiques.

$$\{d\} \rightarrow \{f\}$$

# Voici un exemple simple

```
(Mydocumentspace rdf:type metardf:Knowledgespace)
```

```
(Mydocumentspace metardf:creator FrederickKaplan)
```

```
(Mydocumentspace metardf:useOntology DublinCore)
```

{d}

# Créons un document dans cet espace documentaire

```
(Statement1 rdf:subject HistoryofVenice)
```

```
(Statement1 rdf:predicate rdf:type)
```

```
(Statement1 rdf:object Book)
```

```
(Mydocumentspace metardf:hasStatement Statement1)
```

{d}

# Indiquons que ce document peut être une **source de connaissance**

(KnowledgeSource1 metardf:SourceInstance HistoryofVenice)

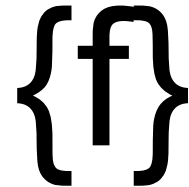
(Mydocumentspace metardf:hasKnowledgeSource KnowledgeSource1)

{d}

# Créons maintenant un espace fictionnel associé à ce livre.

```
(HistoryofVeniceSpace rdf:type metardf:Knowledgespace)
```

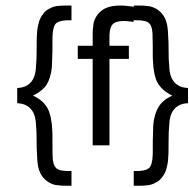
```
(HistoryofVeniceSpace metardf:creator FredericKaplan)
```





# Cet espace utilise une ontologie propre au chercheur qui le code.

(HistoryofVeniceSpace metardf:useOntology MyOntology)



# Lions cet espace fictionnel à l'espace documentaire

(HistoryofVeniceSpace metardf:importKnowledgeSource DocumentSpace)

{d} → {f}

# Créons deux triplets réifiés dans cet espace.

```
(Statement111 rdf:subject RialtoReconstruction)
```

```
(Statement111 rdf:predicate affected)
```

```
(Statement111 rdf:object RialtoBridge)
```

```
(Statement112 rdf:subject RialtoReconstruction)
```

```
(Statement112 rdf:predicate hasTimeSpan)
```

```
(Statement112 rdf:object 1588-1591)
```

```
(HistoryofVeniceSpace metardf:hasStatement Statement111)
```

```
(HistoryofVeniceSpace metardf:hasStatement Statement112)
```

# Lions ces triplets à la source de l'espace documentaire

(Statement111 metardf:hasSource KnowledgeSource1)

(Statement112 metardf:hasSource KnowledgeSource1)

$\{d\} \rightarrow \{f\}$

Avec cette approche, nous pouvons extraire à tout moment la connaissance historique correspondant à un certain contexte de provenance et donc, **à une réalité historique possible.**

Par exemple : reconstituer  
l'histoire de Venise en ne prenant  
en compte que les documents  
écrits par des Vénitiens.

Le système permet de faire la “**jointure**” entre plusieurs espaces fictionnels et détecter automatiquement les possibles incohérences.

$$\{f_1\} \cup \{f_2\} \Rightarrow \{f_3\}$$

# Cette jointure procède en deux étapes

1. La réécriture des deux espaces vers un ontologie pivot

2. La détection des entités et relations identiques.



Pour les espaces fictionnels,  
l'ontologie pivot ISO CIDOC-CRM  
est une bonne option mais ce  
n'est pas la seule.

Les règles de transformation sont  
elles-même décrites dans les  
standards du web sémantique  
(SWRL)

Cette extension permettant une meta-connaissance historique est donc entièrement compatible avec les standards existants.

Pour forcer cette structure particulière des données, les utilisateurs interagissent avec le système au travers d'une interface spécifique.

Cette interface se décline, soit sous la forme d'une interface de programmation (API), soit sous la forme d'une interface utilisateur graphique (GUI).

Les opérations effectuées sont **standardisées** et peuvent donc être optimisées et parallélisées.

Nous travaillons à une architecture dédiée pour ce type de traitement.

Nous comptons constituer un groupe de travail autour de ce modèle de description sémantique. N'hésitez pas à nous contacter si vous souhaitez nous rejoindre.

[dhlab.epfl.ch](http://dhlab.epfl.ch)

[frederic.kaplan@epfl.ch](mailto:frederic.kaplan@epfl.ch)  
[@frederickaplan](https://twitter.com/frederickaplan)

