Répondre à des questions à partir de textes ou de bases de connaissances

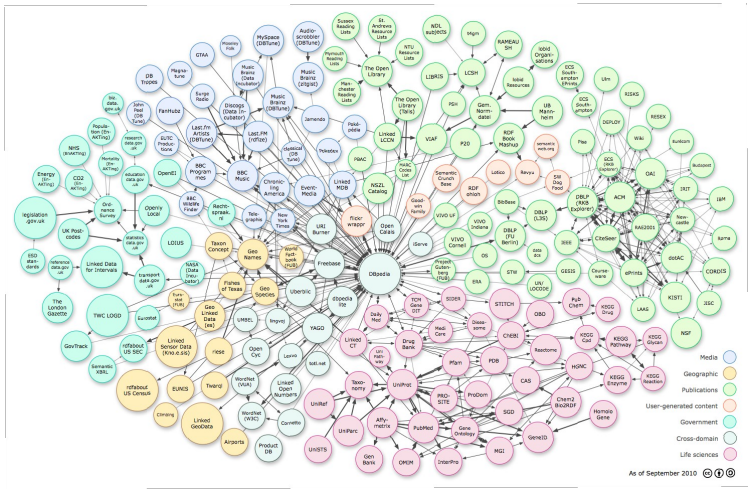
Présentation des problèmes

Brigitte Grau

LIMSI - Orsay

Recherche d'information précise

Quand Sangatte a-t-il été créé ?



http://www.trueknowledge.com/
http://www.wolframalpha.com/
Google.com

Google.com
Watson (IBM)
Synapse (Français)

Cadres d'application

- **Fouille de texte, veille**
  - Sur des produits, des entreprises
    - Qui achète le produit X ? Quelle progression a connu l'entreprise Y ?
  - En domaine bio-médical
    - Quelles sont les molécules qui traitent le diabète ?
    - Evaluation BioAsq
- **Questions posées sur des sites marchands**
  - Avatar
  - FAQ
- **Questions sur le Web**
  - Interrogation de texte, de bases de connaissances
-

Question sur texte

- **Question factuelle :** Quand Sangatte a-t-il été créé ?

… là), marche à pied (on a déjà surpris des étrangers marchant dans le tunnel). Les tentatives de traversées de la Manche sont à peine moins surveillées que les baignades. Quand le temps n'est pas mauvais, on perçoit les côtes anglaises distantes d'une quinzaine de kilomètres. Sur la mer, naviguent en permanence d'énormes ferries.

Le camp ouvert à Sangatte — on dira ici plutôt « camp » que « centre », à cause des conditions de vie qui y prévalent (voir ci-dessous) et de l'improbable statut juridique de cette « chose » sans précédent, sauf les camps des Républicains espagnols à la fin des années 30 — a été inauguré le 24 septembre 1999 dans un hangar où était installée, pendant le forage du tunnel sous la Manche, la logistique technique française.

L'ouvrage, qui appartenait à la société du tunnel, a été réquisitionné in extremis par les pouvoirs publics pour le transformer en lieu d'accueil, alors qu'il allait être vendu. Il a la …

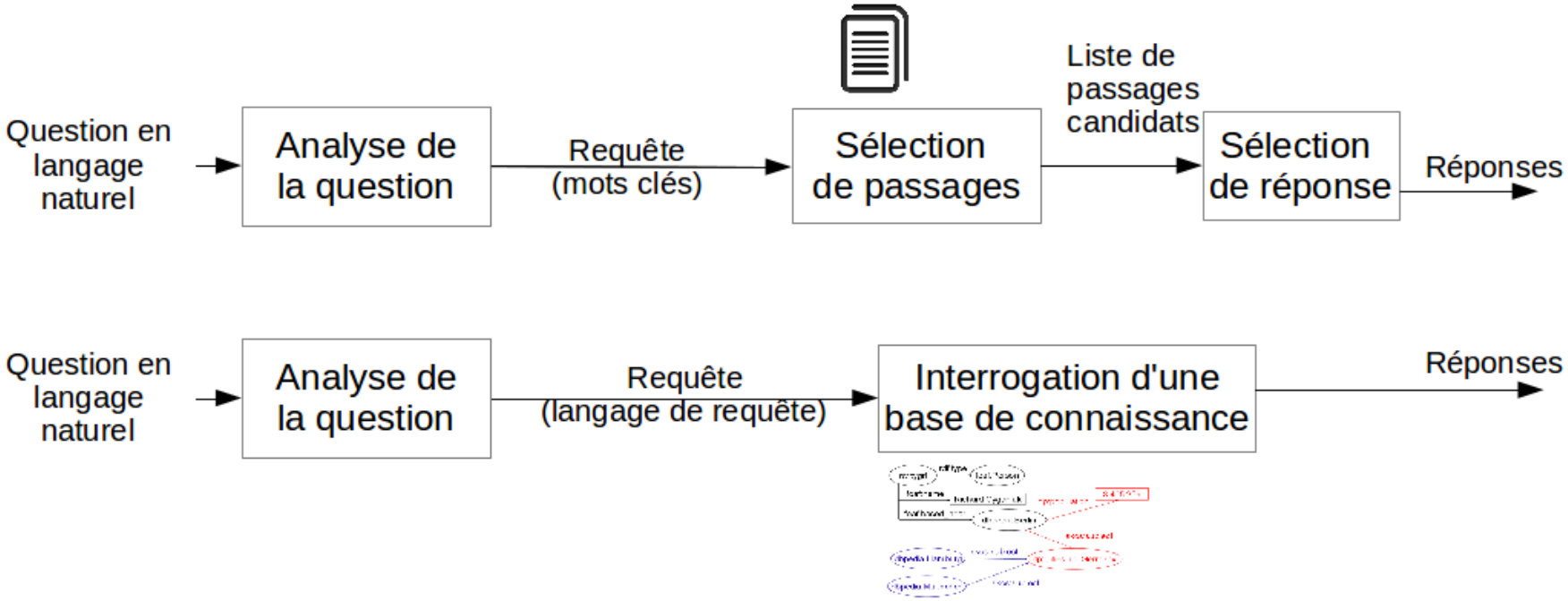
Question sur base de connaissances

- **Questions avec agrégat / fonction**
  - Combien de fois s'est mariée Jane Fonda?
    - SELECT COUNT(DISTINCT ?uri)
      WHERE { res:Jane_Fonda *dbo:spouse* ?uri }
- **Questions listes**
  - Quels livres de Kerouac ont été publiés par Viking Press?
    - SELECT DISTINCT ?uri
      WHERE { ?uri *rdf:type* dbo:Book .
      ?uri *dbo:publisher* res:Viking_Press .
      ?uri *dbo:author* res:Jack_Kerouac .}

Question sur bases de connaissances

- **Questions complexes**
  - Quel état des Etats-Unis a la plus haute densité de population?
    - SELECT DISTINCT ?uri
      WHERE { ?uri rdf:type *yago:StatesOfTheUnitedStates*
      ?uri *dbp:densityrank* ?rank}
      ORDER BY ASC (?rank)
      LIMIT 1

Systèmes de question-réponse



- **Texte**
  - ✗ Non structuré
  - ✗ Variations et ambiguïtés
  - ✔ Nombreuses informations
- **Base de connaissances**
  - ✔ Structurée
  - ✔ Normalisée, non ambiguë
  - ✗ Informations manquantes

Évolution de la tâche et évaluations

- **TREC puis TAC (de 1999 à 2008)**
  - Tâche Question-réponse
  - Tâche RTE (Recognizing Textual Entailment)
- **CLEF (2003-2009)**
  - Multilingue
- **NTCIR (2002-)**
  - Anglais, Japonais, Chinois
- **QA4MRE (2011-2015)@CLEF**
  - Réponse à des QCM
- **QALD@CLEF (2011-)**
  - Questions sur Dbpedia, *questions hybrides*

QA4MRE : examens d'entrée à l'université

The white-haired old man was sitting in his favorite chair, holding a thick book and rubbing his tired eyes. When his nineteen-year-old granddaughter, Valerie, came into the room, he looked up and smiled. His eyes instantly brightened with happiness to see her. ‘Hi, Grandpa. What are you reading?’ she asked, pulling up a chair beside him. ‘Oh, it's a book on the architecture of Spain. But I'm not really reading. Mostly I am just falling asleep over the pictures,’ he said, laughing. ‘Are you finished packing your bags yet?’ he asked. The following morning Valerie and two of her friends were flying to Europe for a two-week holiday. ‘Almost. I need to travel light, you see, so I can buy lots of new dresses and shoes in Paris and Barcelona.’ They both laughed because Valerie was not actually interested in fashion at all. She loved foreign languages, music, art, good food, and many other things - but not shopping for clothes. […]

<question q_id="1">
<q_str>Why did Valerie and Grandpa laugh?</q_str>
<answer a_id="1">Valerie had not finished her preparation.</answer>
<answer a_id="2">Valerie had too many things in her suitcase.</answer>
<answer a_id="3">They both knew that what Valerie said was not true.</answer>
<answer a_id="4">They both understood that Valerie had very little money.</answer>
</question>

Question sur texte

- **Coopération entre recherche d'information et traitement automatique de la langue**
  - Recherche des documents ou des passages pertinents
    - Indexation riche
    - Critère de densité
  - Analyse des questions
    - Type de réponse attendu
  - Analyse des passages pour en extraire la réponse
    - Appariement Question - Passages
    - Gérer la variation linguistique à différents niveaux

Question sur le texte : problèmes

- **Variations lexicales entre question et réponses**
  - Variations morphologiques
    - Who won the Nobel Prize in 1992 ?
    - Menchu, the winner of the 1992 Nobel Prize
  - Variations sémantiques
    - Utilisation d'un hyponyme dans la phrase réponse :
      - What was the name of the dog in the Thin Man movies?
      - The Thin Man (1934): Asta, the wire haired terrier …
    - Utilisation d'un hyperonyme dans la phrase réponse :
      - What is the name of the canopy at a Jewish wedding?
      - … was planning a full Jewish ceremony, complete with the traditional huppah canopy, …

Question sur le texte : problèmes

- **Variations syntaxiques**
  - Who is the daughter of Bill Clinton married to?
    - On 31 July 2010, 30-year-old Chelsea Clinton (the daughter of former U.S. president Bill Clinton and Secretary of State Hillary Clinton), who had recently received a master's degree from Columbia University's Joseph L. Mailman School of Public Health, married 32-year-old Marc Mezvinsky, an investment banker.
- **Combinaisons de variations**
  - What is the legal age to vote in Argentina ?
    - Voting is mandatory for all Argentines aged over 18.
- **Inférences**

Appariement question/passage

- **Problème d'implication textuelle**
  - T => H : un lecteur humain lisant T peut raisonnablement en déduire H (Glickman, 2006)
  - T : passage sélectionné
  - H : question sous forme déclarative (+ réponse candidate)
- **Techniques évaluées dans différentes tâches**
  - RTE (Recognizing Textual Entailment)
  - Reconnaissance de paraphrases
  - Ordonnancement de passages
  - Validation de réponses
  - Traduction

Appariement question/passage

- **Niveaux de représentation**
  - Sac ou séquence de mots/termes
    - Méthodes surfaciques : similarité ou alignements sur critères lexicaux
  - Représentation structurée
    - Syntaxique : exploitation des relations de dépendance
    - Sémantique :
      - Annotation en rôles sémantiques
      - Représentations conceptuelles
  - Représentations fondées sur la notion de relation
- **Caractérisation de la réponse exacte**
  - Pertinence du passage
  - Type attendu
  - Redondance
  - Position dans le passage
    - QAVAL (Grappy et al., 2011)

Résultats QAVAL

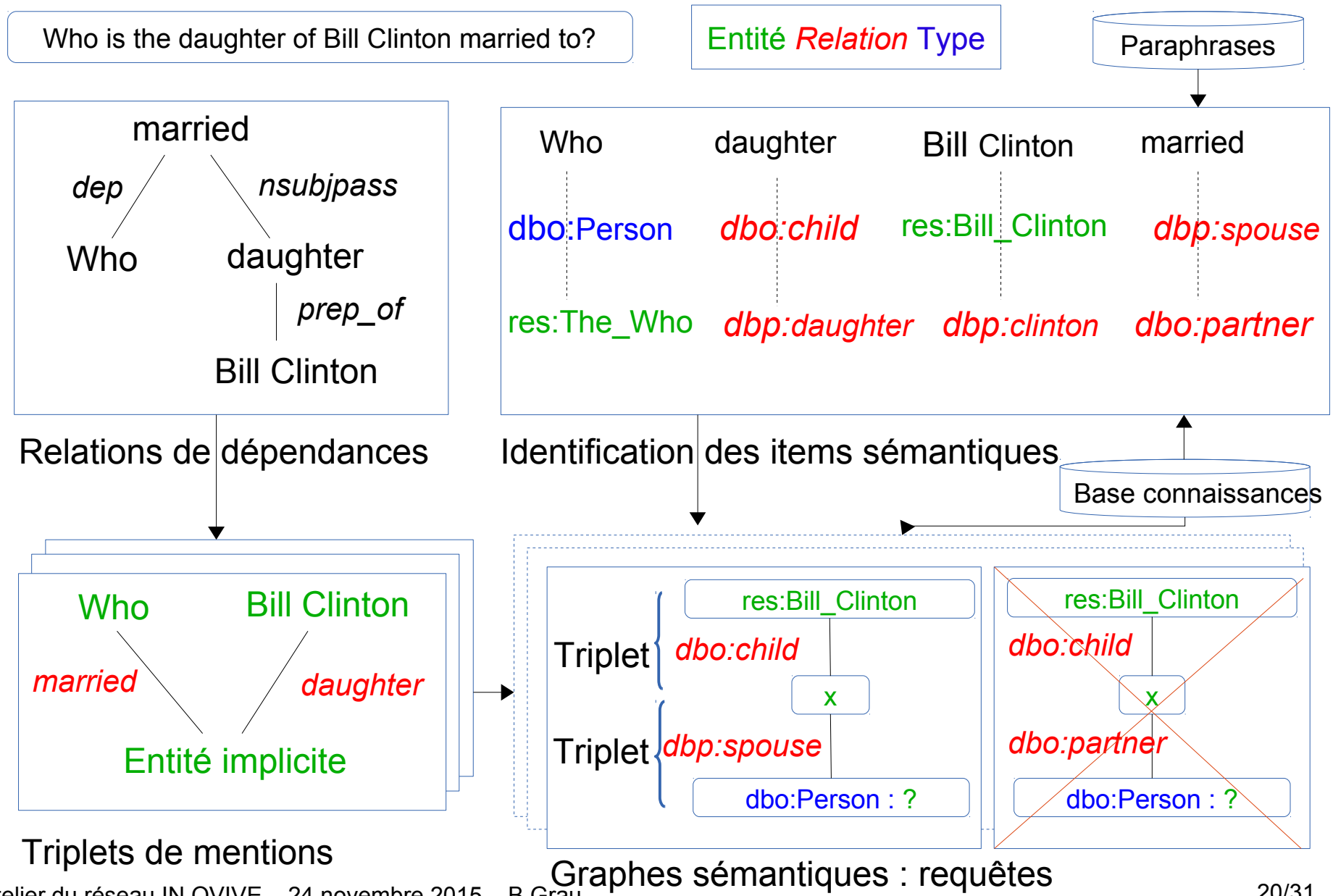
- **Evaluation de QAVAL**
  - Nombre de réponses correctes en rang 1 et dans les 5 premiers rangs
  - MRR (Mean Reciprocal Rank) sur les 5 premiers rangs
  - Baseline : the answer the closest to question words in the top five passages
- **Evaluation du module de validation**
  - Mêmes mesures sur les questions auxquelles il est encore possible de répondre après la sélection de passages

QAVAL	MRR		1st rank		Top 5 ranks	
Documents Web (QUAERO)	**0.43**		**34%**		**56%**	
Baseline Quaero	0.29		21%		43%	
Journaux (EQUER)	**0.47**		**39%**		**60%**	
Baseline EQUER	0.34		27%		47%	
ANSWER VALIDATION		baseline		baseline		baseline
Quaero (122 questions)	**0.49**	0.36	**40%**	28%	**67%**	49%
EQUER (113 questions)	**0.53**	0.38	**43%**	30%	**67%**	52%

Question sur base de connaissances : problèmes

- **Analyse de la question**
  - Déterminer les triplets = instances de relations
    - Termes désignant les mentions d'entités, de relations, de catégories
    - Identifier les éléments de la base de connaissance
    - Rattachement des arguments pour former les triplets
  - Problèmes
    - Variations linguistiques
    - Ambiguïtés
- **Appariement Question – Base → requête SPARQL**
  - Former un graphe de triplets
  - Opérateurs
  - Problèmes : ambiguïtés de structure

Exemple


Who is the daughter of Bill Clinton married to?
Entité Relation Type
Paraphrases
married
dep
nsubjpass
Who
daughter
prep_of
Bill Clinton
Who
daughter
Bill Clinton
married
dbo:Person
dbo:child
res:Bill_Clinton
dbp:spouse
res:The_Who
dbp:daughter
dbp:clinton
dbo:partner
Relations de dépendances
Identification des items sémantiques
Base connaissances
Who
Bill Clinton
married
daughter
Entité implicite
Triplet
res:Bill_Clinton
dbo:child
x
Triplet
dbp:spouse
dbo:Person : ?
res:Bill_Clinton
dbo:child
x
dbo:partner
dbo:Person : ?
Triplets de mentions
Graphes sémantiques : requêtes

Identification des éléments sémantiques

- **Identification des entités**
  - Tâches
    - Reconnaître les termes de la question : annotation
    - Associer les mentions à des entités de la base : désambiguïsation
  - Exemple d'outil pour DBPedia : Spotlight
- **Identification des relations**
  - Bases de paraphrases
    - Patty (Nakashole et al., 2012), WordNet
  - Similarités sur représentations continues par réseau de neurones
    - (Yih *et al.*, 2014), (Bordes *et al.*, 2014)

Appariement question – base de connaissance

- **Patrons de requêtes**
  - (Unger et al. 2012)
  - Peu flexible et dépendant de la base interrogée
- **Transformation de graphes**
  - Graphe de la question → graphe de la requête
    - (Zhou et al., 2014), (Beaumont et al. 2015)
  - 1 étape indépendante de la base, ambiguïtés conservées
- **Apprentissage supervisé de l'appariement**
  - (Yao et Van Durme, 2014) (Xu *et al.*, 2015)
  - A partir de phrases annotées en entités, relations (et réponses) de Freebase
  - Dépendance au schéma de la base
- **Contraintes provenant du schéma de la base**
  - Contraintes sémantiques / domaine et image des relations
  - Existence des triplets

Résultats

- **QALD@CLEF 2015**
  - Hybrid : 10 questions, max 2 réponses trouvées
  - Sur la base seulement

Total: 50

	Processed	Right	Partial	Recall	Precision	F-1	F-1 Global
Xser (en)	42	26	7	0.72	0.74	0.73	0.63
APEQ (en)	26	8	5	0.48	0.40	0.44	0.23
QAnswer (en)	37	9	4	0.35	0.46	0.40	0.30
SemGraphQA (en)	31	7	3	0.32	0.31	0.31	0.20
YodaQA (en)	33	8	2	0.25	0.28	0.26	0.18

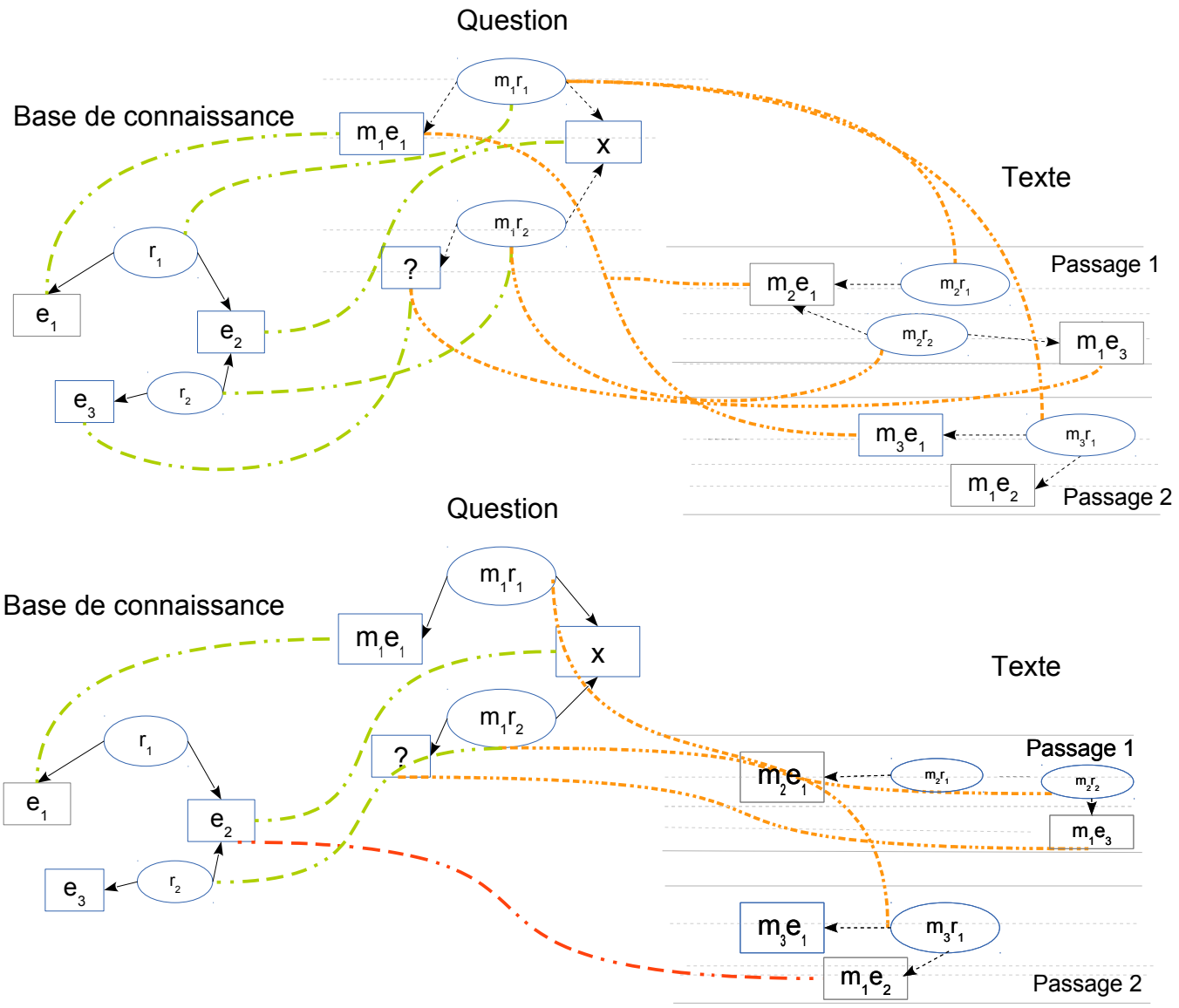
Résultats

- **Bonnes performances à QALD@CLEF**
  - Pour approche supervisée
- **Problèmes non traités**
  - Opérateurs
  - Relations implicites ou non lexicalisées (le film de X)
- **Différence de granularité entre schéma de la base et LN**
  - cosmonaute = astronaute de nationalité russe
  - Type complexe

Question sur ressource hybride

- **Pour les questions possédant ou portant sur une entité**
  - Complémentarité des informations : information absente de l'une des ressources
  - Information contextuelle
  - Niveau de granularité différent langue / schéma
- **Hybridation par fusion**
  - Recherche parallèle + Fusion de réponses
    - (Hildebrandt et al., 2004 ; Cucerzan et Agichtein, 2005)

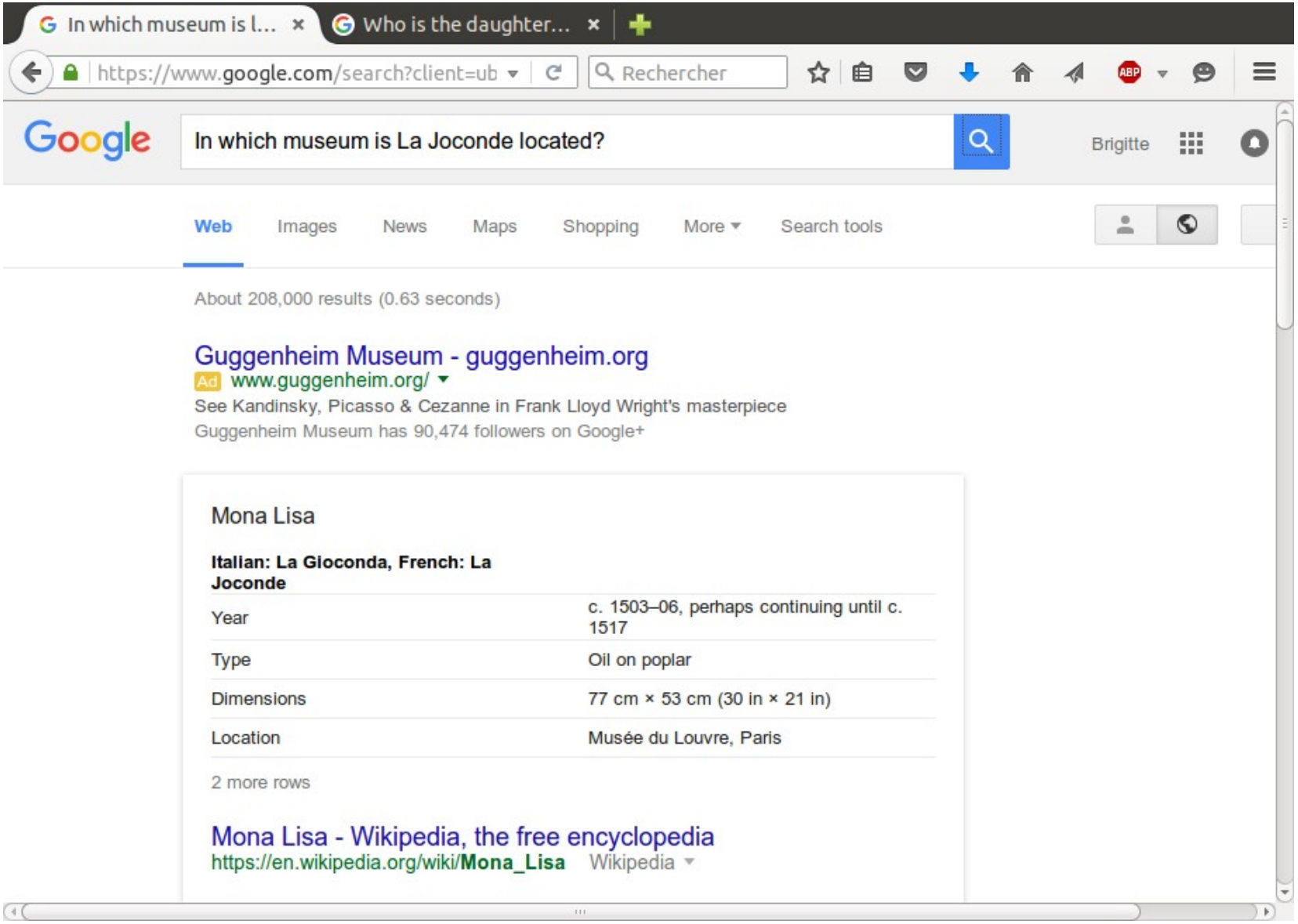
Question sur ressource hybride



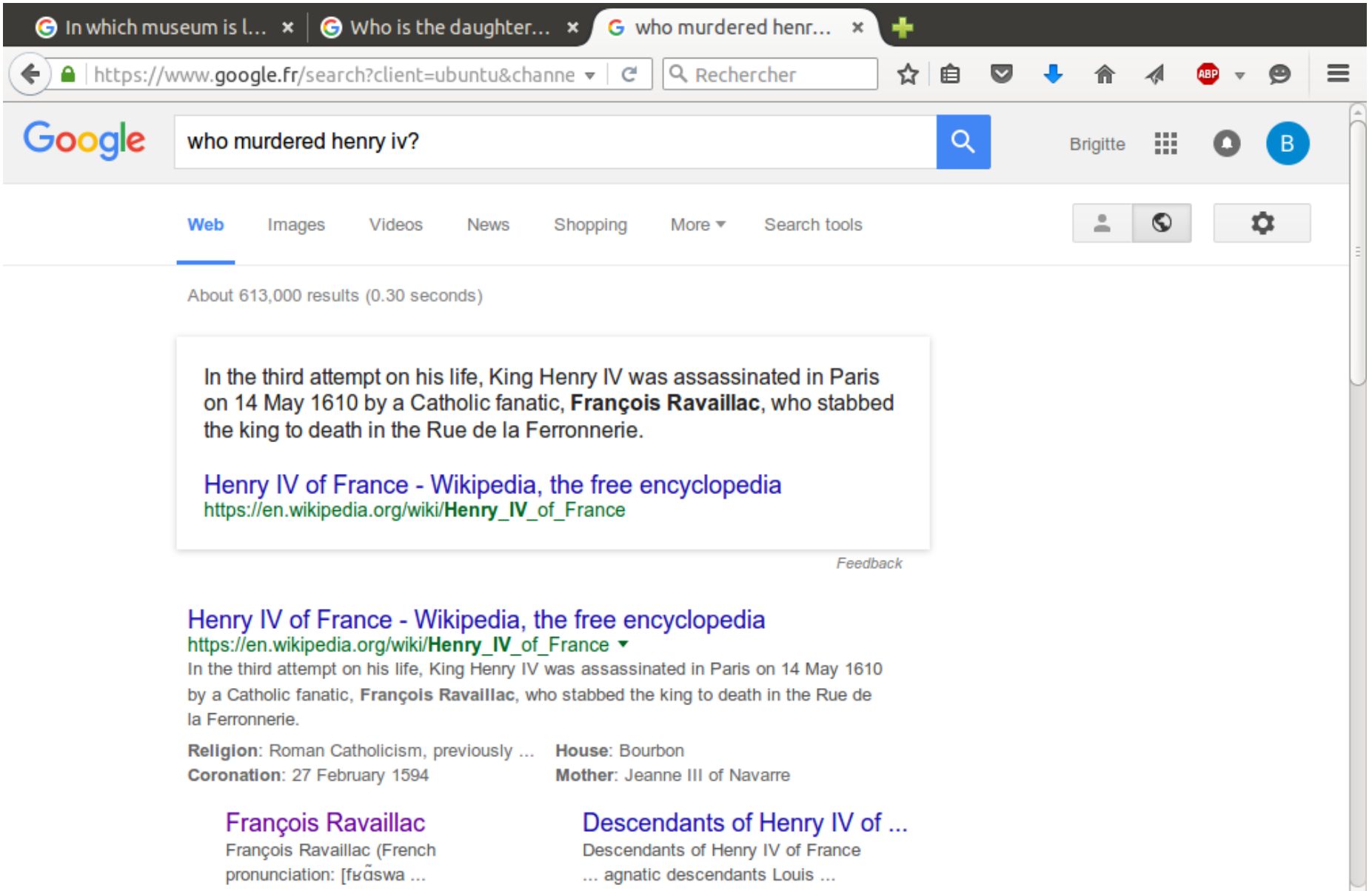
Question sur ressource hybride

- **Quelques travaux**
  - Production d'une représentation hybride
    - (Yahya et al. 2013)
    - Recherche dans la base + recherche dans le texte associé aux entités des contraintes textuelles pour ordonner les réponses
  - Décomposition en relations et recherche dans l'une ou l'autre ressource
    - Watson (IBM), Gagnant à Jeopardy (Chu-Carroll et al., 2012)

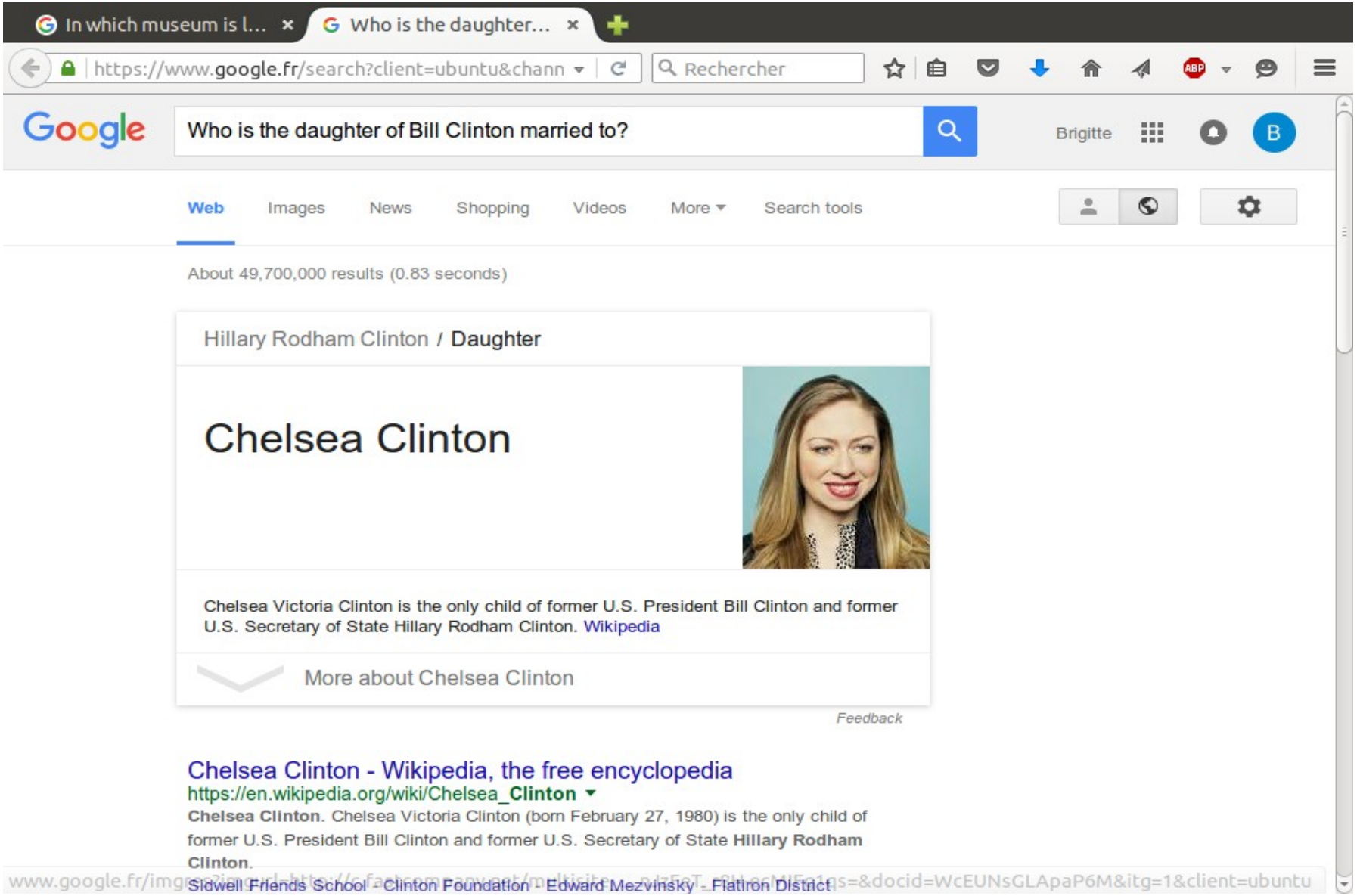
Exemples : Google.com



Exemples : Google.com



Exemples : Google.com



Conclusion

- **Problèmes ouverts**
  - Fusion de réponses
    - Les éléments de réponses sont répartis sur plusieurs documents
      - De simples listes à la constitution de modes d'emploi
      - Vérifier la cohérence de l'agrégation
  - Question – réponse interactif
    - Questions s'enchaînant dans un même contexte, relatif à la première question
  - Raisonnement, inférence
    - QA4MRE : question answering for reading evaluation
    - Recherche hybride