Recherche d'Information (RI)

Mathias Géry
Mathias.Gery@univ-st-etienne.fr
Laboratoire Hubert Curien, UMR CNRS 5516
Université Jean Monnet Saint-Étienne

Laboratoire Hubert Curien

- Effectifs : 240, dont 130 enseignants-chercheurs / chercheurs.
- Département Optique-Photonique.
- Département Informatique-Télécom-Image :
  - Thématique « Formation, compréhension et analyse de l'image ».
  - Thématique « Systèmes embarqués sécurisés et architectures matérielles ».
  - Thématique « Connected Intelligence ».
  - **Thématique « Data Intelligence » (Amaury Habrard).**

Laboratoire Hubert Curien

- Thématique « Data Intelligence » (Amaury Habrard) :
  - Projet « Machine Learning » (E. Fromont).
  - **Projet « Data Mining / Information Retrieval » (C. Largeron) :**
    - 8 enseignants-chercheurs, 5 doctorants / post-docs.
    - Thèmes de recherche :
      - Fouille et Recherche d'Information (RI) Structurée (XML).
      - RI et réseaux sociaux.
      - Fouille de données.
      - Fouille de réseaux sociaux, de graphes.
      - RI Multimédia (images).
      - Etc.
    - Prototypes, expérimentations, participation à des compétitions internationales.
    - Relations entreprises (projets, thèses CIFRE).

Recherche d'Information (RI)

- Plan du cours « RI » :

Michel Beigbeder

  - 5. RI et structure intra-document (ex. XML)
  - 6. RI et structure inter-documents (ex. les liens hypertextes)
  - 7. Projet : de la théorie à la pratique, mise en œuvre d'un SRI sur des données structurées du Web, participation à une mini-compétition de RIS.

Recherche d'Information (RI)

- Plan du cours « RI Structurée » (RIS) :
  - 5. RI et structure intra-document (ex. XML).
    - Introduction : RI classique.
    - Web (HTML, XML) : Document Structuré ou Hypertexte ?
    - RI Structurée :
      - Langages de requêtes structurés.
      - Indexation en RIS.
      - Passage Retrieval, modèle à proximité.
      - Pondération en RIS : normalisation, types d'éléments, propagation.
    - Évaluation en RI : la compétition INEX.
  - 6. RI et structure inter-documents (ex. les liens hypertextes).
  - 7. Projet : participation à une mini-compétition de RIS.

Recherche d'Information (RI)

- Cours (diapos sur l'ENT).
- TDs : algorithmes, articles de recherche.
- Projet : Moteur de recherche de documents structurés.
  - À réaliser entre les séances,
  - Discussions en séance,
  - Étapes intermédiaires,
  - Rendu : 25 janvier.
  - ➔ note « pratique » du cours de Recherche d'Information.

Recherche d'Information (RI)

- Planning : attention aux modifications de dernière minute !
- ➔ vérifiez ADE !
  - 1) Vendredi 21/10, 14h-16h, salle B10 Carnot
  - 2) Vendredi 18/11, 14h-16h, salle B10 Carnot
  - 3) Mardi 06/12, 13h30-15h30, salle A13 Métare
  - 4) Mardi 13/12, 13h30-15h30, salle A13 Métare
  - 5) Jeudi 5/01/2017, 14h-16h, salle B10 Carnot
  - 6) Mardi 17/01, 13h30-15h30, salle A13 Métare

Cours en ligne : plateforme Claroline

Cours en ligne : RI2016
https://ead-sciences.univ-st-etienne.fr/claroline/course/index.php?cid=RI2016

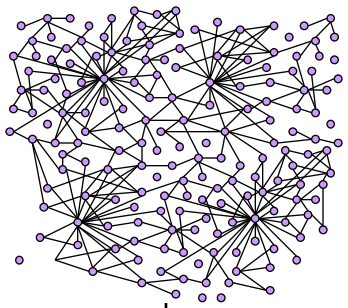
Baaziz	Hamza	Escalle	Dimitri
Bah	Alhassane	Forestier	Fabien
Bendari	Yassine	Gourseaud	Pierre
Berlande	Jacomo	Khalifa	Nizar
Bezrhoud	Ibrahim	Mersel	Dali
Boualem	Oussama	Ramirez	Diana
Bouassida	Haithem tahar	Rouatbi	Nour el houda
Cretel	Lea	Tichit	Ludovic
Da costa vaz	Julien	Wang	Jing
Diallo	Marlyatou	Wang	Zun zun

Recherche d'Information (RI)

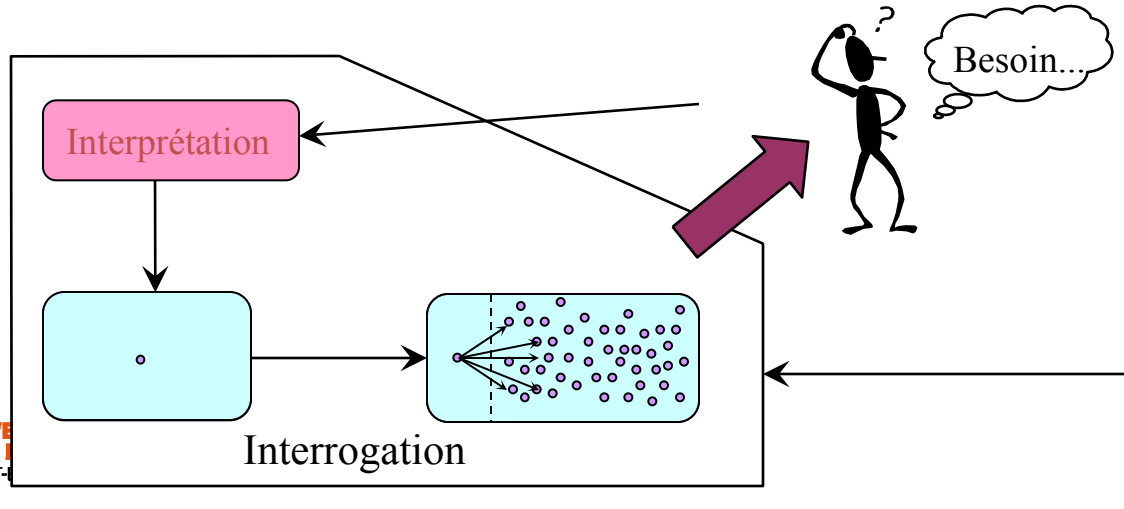
- Plan du cours « RI Structurée » (RIS) :
  - 5. RI et structure intra-document (ex. XML).
    - Introduction : RI classique.
    - Web (HTML, XML) : Document Structuré ou Hypertexte ?
    - RI Structurée :
      - Langages de requêtes structurés.
      - Indexation en RIS.
      - Passage Retrieval, modèle à proximité.
      - Pondération en RIS : normalisation, types d'éléments, propagation.
    - Évaluation en RI : la compétition INEX.
  - 6. RI et structure inter-documents (ex. les liens hypertextes).
  - 7. Projet : participation à une mini-compétition de RIS.

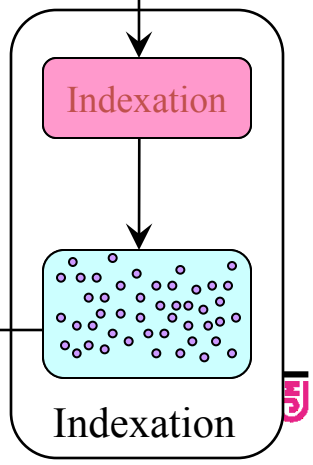
Recherche d'Information « Classique » (rappel)

- « Recherche d'Information », cf. cours M. Beigbeder.
- Modèles classiques de RI (booléen, vectoriel) :
  - L'unité d'indexation (document) est la page HTML.
  - Correspondance : document ←→ requête.
  - ➔ documents atomiques et indépendants.



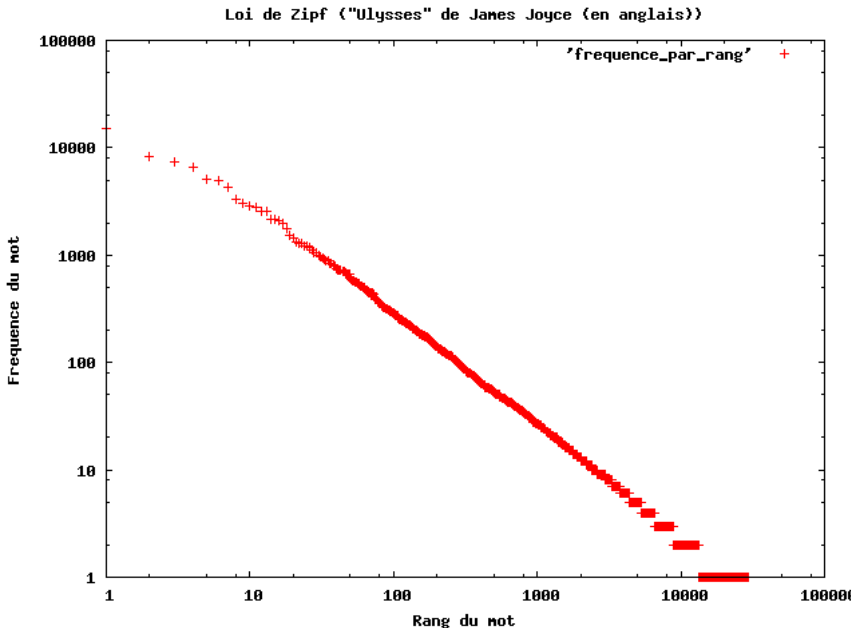
Web





Modèles classiques de RI (rappel)

- 50 ans de recherche en RI !
- Exemple : modèles « bag of words » (« sac de mots »).
  - Documents atomiques et indépendants.
  - Modéliser la distribution des termes :
    - Loi de Zipf (fréquence du terme $f(n) = \frac{K}{n}, n = rang\ du\ terme$).
    - Une formule « magique » : pondération *tf.idf*
    - Représentativité.
    - Discriminance.



Pondération en RI classique (rappel)

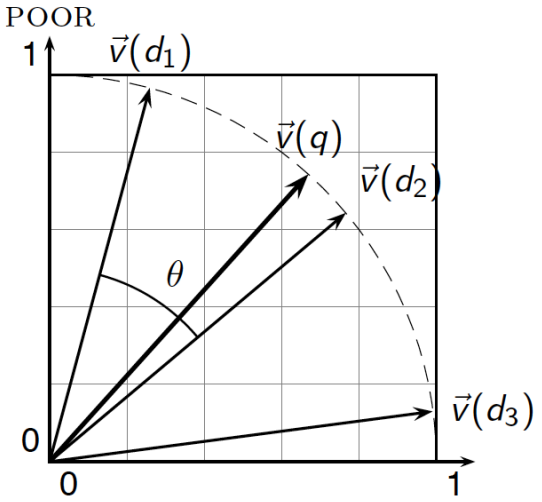
- Pondération *tf.idf* classique ?
- Exemple SMART *ltn* :
  - Poids $w_{t,d}$ du terme t dans le document d :

$$w_{t,d} = TF_{t,d} * IDF_t = (1 + \log(tf_{t,d})) * \log\left(\frac{N}{df_t}\right)$$

  - Avec :
    - $TF_{t,d}$ = **représentativité** de t dans d
      = par ex. : nombre d'occurrences de t dans d
    - IDF_t = **discriminance** de t dans le corpus
      = par ex. : nombre de documents contenant t
    - N = nombre de documents dans le corpus.

Modèles classiques de RI (rappel)

- Exemple : modèle vectoriel [Salton83]
  - Document représenté par un vecteur dans l'espace vectoriel des termes.
  - Index = matrice termes x documents.



$\vec{d_1} = (w_{11}, w_{12}, \ldots, w_{1j}, \ldots, w_{1n})$
$\vec{d_2} = (w_{21}, w_{22}, \ldots, w_{2j}, \ldots, w_{2n})$
$\vec{d_3} = (w_{31}, w_{32}, \ldots, w_{3j}, \ldots, w_{3n})$

$\vec{q} = (w_{q1}, w_{q2}, \ldots, w_{qj}, \ldots, w_{qn})$

d/t	RICH	POOR
d_1	0.25	0.97
d_2	0.75	0.63
d_3	0.98	0.14
q	0.65	0.75

Matrice termes-documents

Source : P. Nayak, P. Raghavan, Standford

Et la structure dans tout ça ? Quelle structure ?

- Plan du cours « RI Structurée » (RIS) :
  - 5. RI et structure intra-document (ex. XML).
    - Introduction : RI classique.
    - Web (HTML, XML) : Document Structuré ou Hypertexte ?
    - RI Structurée :
      - Langages de requêtes structurés.
      - Indexation en RIS.
      - Passage Retrieval, modèle à proximité.
      - Pondération en RIS : normalisation, types d'éléments, propagation.
    - Évaluation en RI : la compétition INEX.
  - 6. RI et structure inter-documents (ex. les liens hypertextes).
  - 7. Projet : participation à une mini-compétition de RIS.

Structure du Web et RIS

- **<u>Objectif :</u> RI Structurée (RIS) pour le Web.**

- RI « classique » :
  - Ensemble de documents atomiques et indépendants.
  - ➔ Modèles classiques à repenser pour la RI Structurée.

- Quels documents sur le Web ?
  - On oublie le Web 2.0. Ici : Web documentaire, « Web 1.0 » ou « Web 1.5 ».
  - Client / serveur : 1 requête HTTP ➔ 1 document HTML/XML.
  - Documents Structurés « classiques » (arborescence).
  - Dynamique mais pas trop (CMS, blog, etc.).

- Simplification : Web = Documents Structurés + Graphe.

Structure du Web et RIS

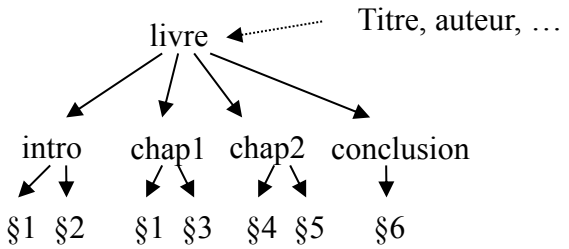
- Qu'est-ce qu'un document sur le Web ?
- Un document structuré arborescent ?
  - HTML, XML, etc.
  - <h1>, <h2>, <p>, , , <table>, etc.
- Un hypertexte ?
  - Page HTML = nœud du graphe.
  - Lien hypertexte = arcs du graphe.
  - <a href="http://www.univ-st-etienne.fr">
  - <a href="rubriques/actus.html">
  - <a href="#introduction">
- Une base de données (data-centric) ?

Web : Documents Structurés (DS) ?

- Document Structuré (DS) :
  - contenu + structure + sens de lecture
- Le Web contient des documents structurés :
  - Structure implicite ou formats HTML, XML, MPEG, etc.
  - Un livre est composé de : titre, introduction, chapitre 1, chapitre 2, conclusion, etc.
  - Méta-données (auteur, mots clés, propriétaire, date, etc.)

```
<html>
<head>
<title> Mon titre </title>
<meta name="author"
        content="Mathias">
<meta name="keywords"
      content="AAW Web">
</head>
<body>
```

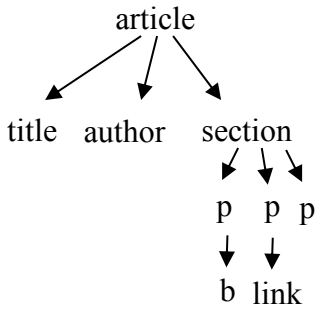
```
<h1>Introduction</h1>
  bla bla
<h1>Chapitre 1</h1>
  bla bla
<h1>Chapitre 2</h1>
  bla bla
<h1>Conclusion</h1>
…….
</body></html>
```



XML : eXtensible Markup Language

- XML :
  - Version simplifiée de SGML.
  - Standard W3C (XML 1.0, 1998).
- Concepts XML :
  - Balises (tags).
  - Balises ouvrantes / fermantes.
  - Eléments :
    - Feuilles ou nœuds.
    - Composants logiques d'un document XML.
  - Attributs.
  - Pas de chevauchement.

```
<article>
  <title>RI Structurée</title>
  <author id="443">Mathias Géry</author>
  <section>
    <p>La RI <b>structurée</b>
              … bla bla…</p>
    <p>Pour aller <link href="plusloin.xml">
                plus loin</link></p>
    <p>Références…</p>
  </section>
</article>
```



XML : eXtensible Markup Language

- Comparable à HTML, mais :
  - Extensible (balises / tags).
  - DTD (Document Type Definition).
  - Les balises doivent être fermées.

- Différents structures :
  - Logique,
  - physique,
  - de mise en forme,
  - de navigation, etc.

```
<!DOCTYPE article [
  <!ELEMENT article (title, author, section+)>
  <!ELEMENT title (#PCDATA)>
  <!ELEMENT author (#PCDATA)>
  <!ATTLIST author id CDATA  #REQUIRED>
  <!ELEMENT section (p+)>
  <!ELEMENT p (b, link, #PCDATA)*>
  <!ELEMENT b (#PCDATA)>
  <!ELEMENT link (#PCDATA)>
  <!ATTLIST link href (#CDATA) #REQUIRED>
]>
```

- XML utilisé par LA compétition de RIS :
  - INEX (**IN**itiative for the **E**valuation of **X**ML Retrieval), 2002.

Exemple de document structuré en XML

```
<article xmlns:xlink="http://www.w3.org/1999/xlink">
  <header>
    <title>Handel House Museum</title>
    <id>1707709</id>
  </header>
  <bdy>
    <image width="150px" src="London_Handel_House.jpg" type="thumb">
      <caption>Handel House. Note the <link xlink:href="310649.xml">blue plaque</link>
      </caption>
    </image>
    <p>
      The <b>Handel House Museum</b> at 25 <link xlink:href="2599649.xml">Brook
Street</link>, in the exclusive central <link xlink:href="17867.xml">London</link>
district of <link xlink:href="94167.xml">Mayfair</link> was the home of the <link
xlink:href="11867.xml">German</link> born <link xlink:href="4500.xml">baroque</link>
composer <link xlink:href="12775.xml">George Frideric Handel</link>
from 1723 until his death at the house in 1759. He composed works such as
<it><link xlink:href="149131.xml">The Messiah</link></it>, <it><link
xlink:href="811987.xml">Zadok the Priest</link></it> and the <it><link
xlink:href="1246814.xml">Fireworks Music</link></it> there.
    </p>
    <sec>
      <st>The museum</st>
      <p>
        The house has been restored to look as it did during Handel's occupancy. A typical
early 18th century London terrace house, it comprises a basement, three main storeys and an
attic, and Handel was the first occupant. The attic was later converted into a fourth full
floor. The ground floor is now a music and gift shop and the upper floors are leased to a
charity called the Handel House Trust, and have been open to the public since 8 November
2001. The interiors have been restored to the somewhat spartan style of Georgian era,
using mostly architectural elements from elsewhere, as other than the staircase, few of
the original interior features survived. The Handel House Collection Trust has assembled a
collection of Handel memorabilia, including the Byrne Collection of several hundred items,
which was acquired in 1998.
      </p>
    </sec>
  </bdy>
</article>
```

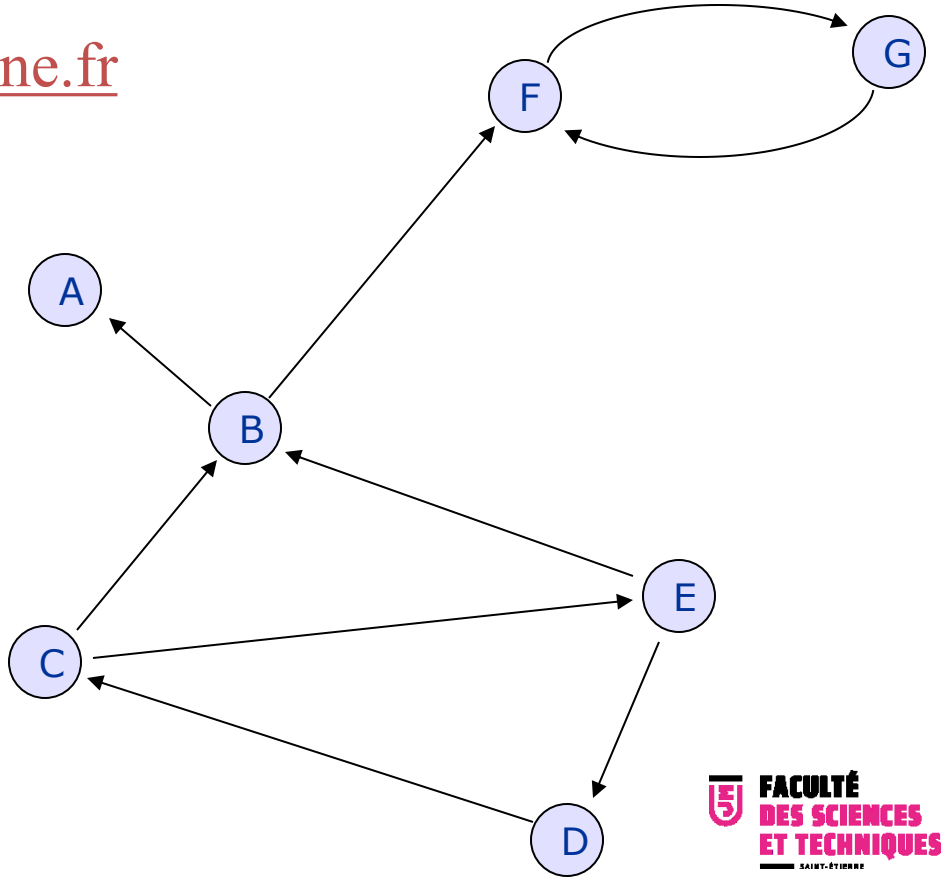
Document « The Haendel House »
(Collection « INEX Wikipédia »)

Compétition de RI XML « INEX »

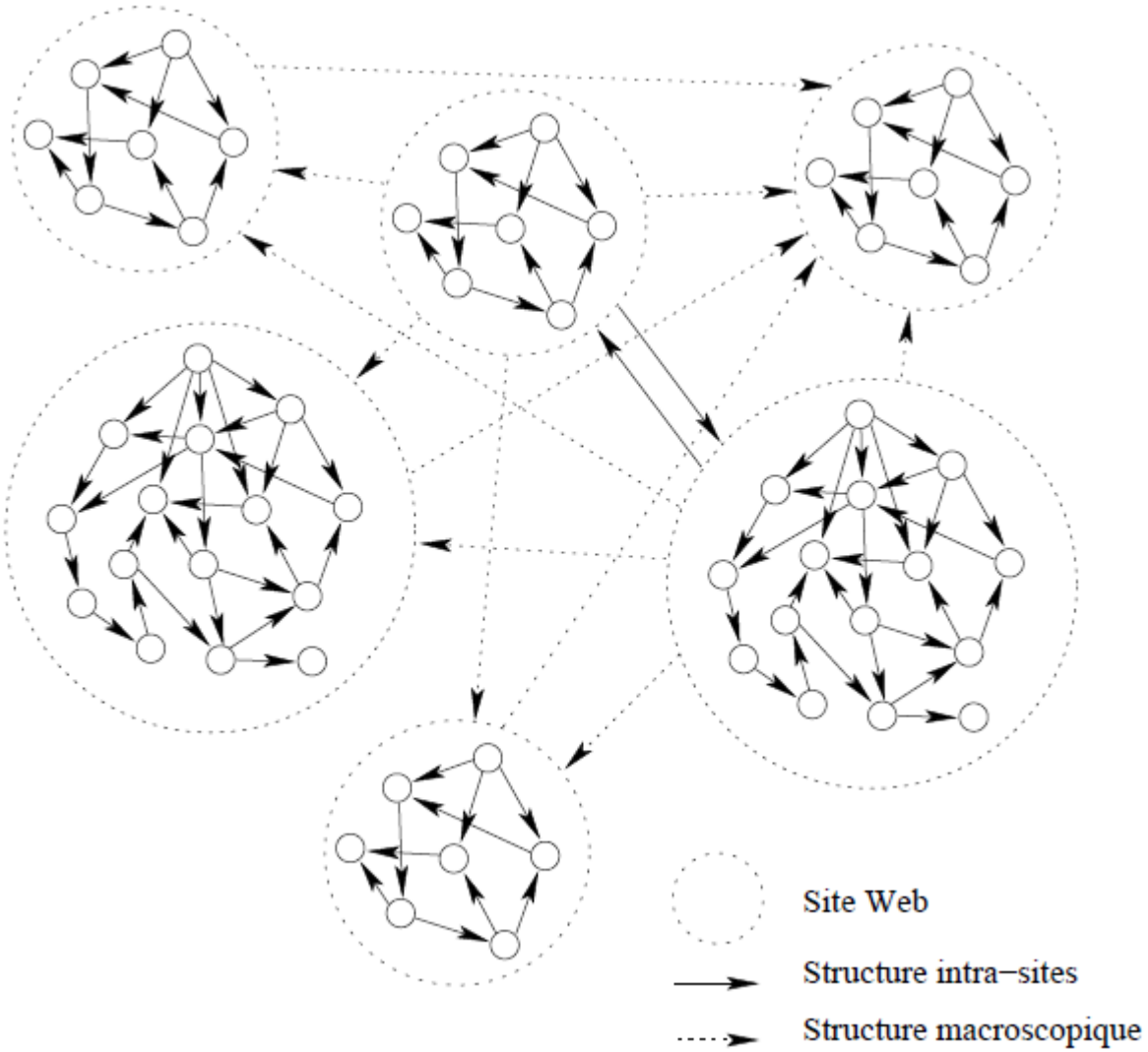
Balises (tags) : article, bdy, header, title, image, caption, sec, p, etc.

Web : Hypertexte ?

- Hypertexte : représentation non linéaire d'une information textuelle sous la forme d'un graphe de nœuds connectés par des liens.
- URL : http://www.univ-st-etienne.fr
- Le Web est est un hypertexte :
  - Liens hypertextes, URLs
  - Structure de graphe !
- Graphe interne à un site :
  - Nœuds = pages Web,
  - Arcs = liens hypertextes.
- Graphe inter-sites :
  - Nœuds = sites Web,
  - Arcs = liens hypertextes.

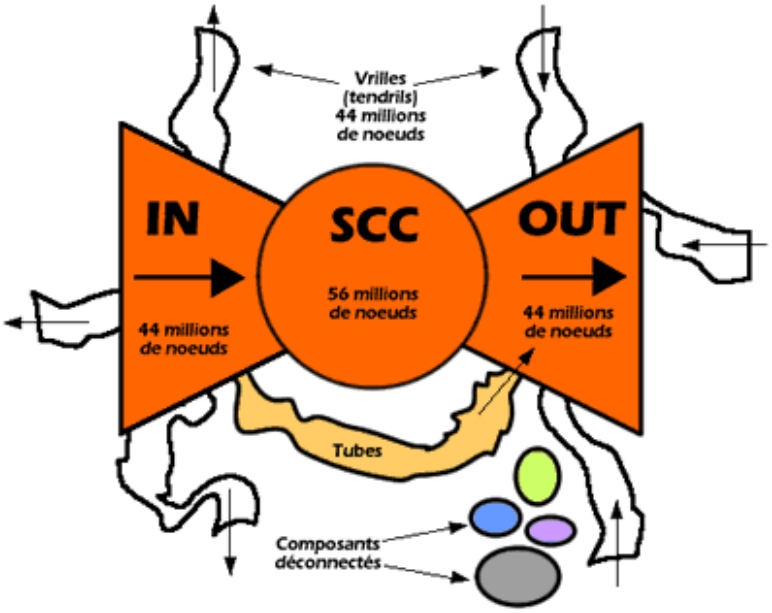


Web : Hypertexte ?



Le graphe du Web

- Graphe orienté :
  - Nœud = pages Web, Arcs = liens hypertexte.
  - Degré entrant/sortant d'un nœud.
  - Chemin : suite de nœuds reliés par des arcs.
- Calculs « classiques » dans le Web vu comme un graphe :
  - Distance : longueur du plus petit chemin (ppc) entre 2 nœuds.
  - Distribution des degrés.
  - Diamètre : moyenne ppc entre toutes les paires de nœuds (19 clics [Albert 99]).
  - Connectivité, ex. : composantes fortement connexes (« nœud papillon » [IBM 99]).



UNIVERSITÉ JEAN MONNET SAINT-ÉTIENNE

Hypertexte en XML

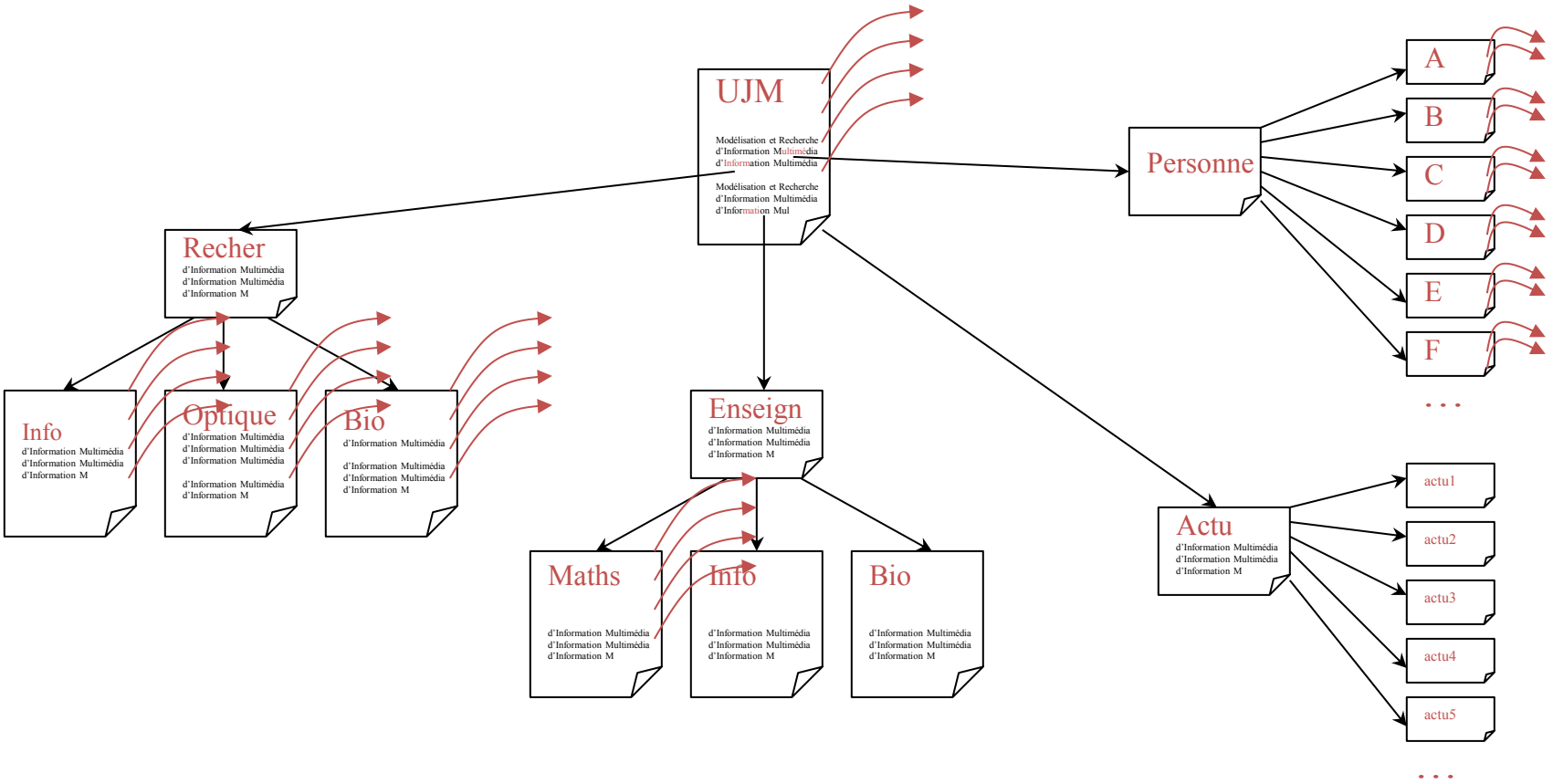
```
<article xmlns:xlink="http://www.w3.org/1999/xlink">
  <header>
    <title>Handel House Museum</title>
    <id>1707709</id>
  </header>
  <bdy>
    <image width="150px" src="London_Handel_House.jpg" type="thumb">
      <caption>Handel House. Note the <link xlink:href="310649.xml">blue plaque</link>
      </caption>
    </image>
    <p>
      The <b>Handel House Museum</b> at 25 <link xlink:href="2599649.xml">Brook
Street</link>, in the exclusive central <link xlink:href="17867.xml">London</link>
district of <link xlink:href="94167.xml">Mayfair</link> was the home of the <link
xlink:href="11867.xml">German</link> born <link xlink:href="4500.xml">baroque</link>
composer <link xlink:href="12775.xml">George Frideric Handel</link>
from 1723 until his death at the house in 1759. He composed works such as
<it><link xlink:href="149131.xml">The Messiah</link></it>, <it><link
xlink:href="811987.xml">Zadok the Priest</link></it> and the <it><link
xlink:href="1246814.xml">Fireworks Music</link></it> there.
    </p>
    <sec>
      <st>The museum</st>
      <p>
        The house has been restored to look as it did during Handel's occupancy. A typical
early 18th century London terrace house, it comprises a basement, three main storeys and an
attic, and Handel was the first occupant. The attic was later converted into a fourth full
floor. The ground floor is now a music and gift shop and the upper floors are leased to a
charity called the Handel House Trust, and have been open to the public since 8 November
2001. The interiors have been restored to the somewhat spartan style of Georgian era,
using mostly architectural elements from elsewhere, as other than the staircase, few of
the original interior features survived. The Handel House Collection Trust has assembled a
collection of Handel memorabilia, including the Byrne Collection of several hundred items,
which was acquired in 1998.
      </p>
    </sec>
  </bdy>
</article>
```

Document « The Haendel House »
(Collection « INEX Wikipédia »)

Document « The Haendel House »

Balise : link !

DS ou hypertexte ?



Relations de référence

Relations de composition

Web : base de données ?

- Niveaux de structure fine.
- XML : Data-centric vs Document-centric.
- Import/export BD.
- Exemple data-centric :

```
<article>
  <title>RI Structurée</title>
  <author idAut="443">Mathias Géry</author>
  <editor idEdt="121">Springer</author>
  <nbpages>120</nbpages>
  <prix>30</prix>
</article>
```

- Cf. Web Sémantique…

Documents du Web ?

- En RI, un « document » :
  - Ce qu'on indexe,
  - Ce qu'on retrouve ➔ granularité de la réponse.

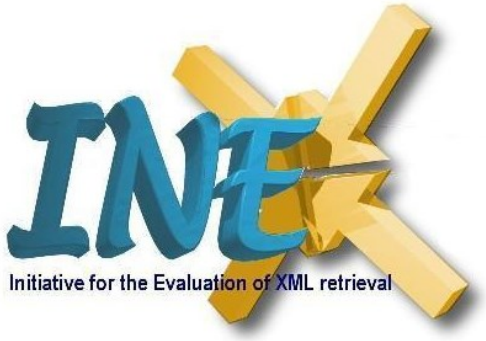
- Qu'est-ce qu'un document sur le Web ?
  - Un fichier HTML ? URL : http://www.univ-st-etienne.fr
  - Un morceau de fichier HTML ? Paragraphe ? Phrase ?
  - Une suite de fichiers HTML ? Avec les liens ?
  - Un site Web ? Une partie d'un site Web ?
  - Plusieurs sites Web ?
  - Un flux RSS ?
  - Un document virtuel (fragments) personnalisé ?

- Plan du cours « RI Structurée » (RIS) :
  - 5. RI et structure intra-document (ex. XML).
    - Introduction : RI classique.
    - Web (HTML, XML) : Document Structuré ou Hypertexte ?
    - RI Structurée :
      - Langages de requêtes structurés.
      - Indexation en RIS.
      - Passage Retrieval, modèle à proximité.
      - Pondération en RIS : normalisation, types d'éléments, propagation.
    - Évaluation en RI : la compétition INEX.
  - 6. RI et structure inter-documents (ex. les liens hypertextes).
  - 7. Projet : participation à une mini-compétition de RIS.

Evaluation de la RIS : la compétition INEX

- Évaluation en RI : cf. cours M. Beigbeder

- **IN**itiative for the **E**valuation of **X**ML Retrieval.
- 1ère compétition en RIS (RI XML), 2002.
- Collection de test :
  - Collection de documents.
  - Tâches de RI.
  - Jeu de requêtes.
  - Jugements de pertinence.
  - Mesures d'évaluation.
- Adaptés à la RIS !



Compétition INEX : collection de documents

- Plusieurs collections depuis 2002.
- Collection de documents issus de Wikipédia (2006) :
  - 659'388 articles Wikipedia, 52 millions éléments XML,
  - 1'244 balises différentes,
  - 4.5 Go dont 1,6 Go de texte,
  - 114 requêtes (2008) avec jugements de pertinence.
- Collection de documents issus de Wikipédia (2009) :
  - 2,6 millions articles Wikipedia, 400 millions éléments XML,
  - >30'000 balises différentes,
  - 50 Go de XML,
  - 68 + 52 requêtes (2009+2010) avec jugements de pertinence.

Compétition INEX : jeu de requêtes

- Exemple de requête (« topic » n°14, INEX 2010) :

```
<topic id="2010014" ct_no="329">
  <title>composer museum</title>
  <description>Documents or parts of documents that describe or identify a
    museum dedicated or which has a significant section dedicated to a
    composer</description>
  <narrative>I want to know the museum that are dedicated to or which have
    a significant section dedicated to a composer. Other music related museum
    which are  not dedicated to a composer are not relevant (for instance a
    museum of musical instruments)</narrative>
</topic>
```

- Seule l'utilisation du champ « *title* » est autorisée.

Compétition INEX : tâches de RI

- Exemple : « track » (piste) RI « ad-hoc » :
  - RI XML, avec ou sans contraintes de structure.
  - Résultat = liste classée d'éléments ou de passages pertinents, à la fois spécifiques et exhaustifs.

- Exemples de « sous-tâches » de RI Ad-Hoc :
  - Relevant in Context (RIC) :
    - Liste d'éléments (ou de passages) groupés par article.
    - Au sein d'un même article : pas forcément dans l'ordre d'origine.
    - Pas de recouvrement.

Cf. par ex. INEX 2010 : http://www.inex.otago.ac.nz/tracks/adhoc/runsubmission.asp?action=specification

Compétition INEX : tâches de RI (suite)

- Exemples de « sous-tâches » de RI Ad-Hoc :
  - Relevant in Context (RIC).
  - Restricted RIC : idem, maxi 500 caractères / articles (smartphones).
  - Restricted Focused :
    - Liste d'éléments (ou de passages), non groupés.
    - Pas de recouvrement.
    - Maxi 1000 caractères / requête.
  - Efficiency : 15, 150 ou 1500 éléments par requête, en précisant : temps de calcul, coût d'E/S, caractéristiques des machines, etc.

Cf. par ex. INEX 2010 : http://www.inex.otago.ac.nz/tracks/adhoc/runsubmission.asp?action=specification

Compétition INEX : tâches de RI

- Mais aussi d'autres « pistes » (tracks) :
  - Book,
  - Data-centric.
  - Interactive.
  - Link-the-wiki.
  - Q/A.
  - Relevance Feedback.
  - Web Service Discovery.
  - XML Mining.

INEX : tâche RIC

- Consulter :
  - http://www.inex.otago.ac.nz/tracks/adhoc/runsubmission.asp?action=specification

1) Relevant in Context Task

Motivation for the Task

The scenario underlying the Relevant in Context Task is to return the relevant information (captured by a set of elements or passages) within the context of the full article. The task makes a number of assumptions: Display results will be grouped per article, ~~in their original document order,~~ providing access through further navigational means; Users consider the article as the most natural unit, and prefer an overview of relevance in their context.

Evaluation

Task is as before, but viewed as a form of snippet retrieval. We use as main measure the proposal from the Univ. Tampere that takes the reading length into account -- probably the T2I(300) measure from http://dx.doi.org/10.1007/s10791-010-9133-9 which strongly penalizes the retrieval of non-relevant text. The new measure takes the suggested reading order within each article into account, but almost all earlier RiC submissions contained this ranking.

Results to Return

The aim of the Relevant in Context Task is to first identify relevant articles (the fetching phase), and then to identify the relevant results within the fetched articles (the browsing phase). The /article[1] element itself need not be returned, but is implied by any result from a given article. Since the content of an element is fully contained in its parent element and ascendants, the set may not contain overlapping elements. Also passage results may not be overlapping.

Summarizing: The Relevant in Context Task requires a ranked list of articles, and for each article a ranked list results covering the relevant material in the article. Overlap is not permitted.

Compétition INEX : « runs »

- « Run » : liste de résultats retournée par un SRI :
  - Par défaut : maxi 1500 éléments XML par requête.
  - Grouper par n° de requête, et pour chaque requête : trier par ordre décroissant de score.
  - Selon les tâches, restrictions de taille, d'ordre des éléments, de recouvrement, etc.
- Exemple (avec 7 requêtes) : 1 run = 1 fichier = 7*1500 lignes.

- Chaque participant peut soumettre un nombre limité de runs à chaque tâche.

Format des « runs » INEX

2010001 Q0 364275 12 0.9765 Emse514 /article[1]/bdy[1]/sec[6]/p[1]

- Colonnes séparées par des espaces :
  - Colonne 1 : requête n°2010001.
  - Colonne 2 : Q0 (inutilisé)
  - Colonne 3 : article n° 364275 de la collection INEX.
  - Colonne 4 : cet élément est classé au 12ème rang dans ce run.
  - Colonne 5 : le score de cet élément est 0.9765.
  - Colonne 6 : nom de l'équipe.
  - Colonne 7 : le chemin de l'élément dans l'article 364275.

    Remarque : ce chemin peut être : /article[1]

Cf. pour INEX 2010 : http://www.inex.otago.ac.nz/tracks/adhoc/runsubmission.asp?action=specification