The AMARYLLIS task

Information Retrieval for Scientific Collections

Patrick Kremer Laurent Schmitt
INstitut de l'Information Scientifique et Technique NANCY, France

History (1)

- ✓ Amaryllis project : an information retrieval task since 1996
- ✓ created by the french Ministry of Research and AUF (ex AUPELF-UREF)
- ✓ IR in scientific documents, texts, newspapers, …
- ✓ 2 evaluation campaigns
- ✓ campaign of the third millenium included in CLEF

History (2)

- ✓ first campaign
- ✓ 1996/1997
- ✓ classical test
- ✓ 130 Mb
- ✓ 30 topics
- ✓ 8 teams / 12

- ✓ second campaign
- ✓ 1998/1999
- ✓ classical test
- ✓ internet test
- ✓ interlingual test
- ✓ 100 Mb
- ✓ 70 topics
- ✓ 11 teams / 16

Main objectives

- ✓ information retrieval
- ✓ access to scientific documents
- ✓ specialized vocabulary

Characteristics

- ✓ short documents
- ✓ strongly structured documents (ti, ab, kw)
- ✓ writing style
- ✓ specific terminology
- ✓ keywords
- ✓ thesaurus

Different kind of tests

- ✓ on free text : title and abstract
- ✓ with controlled vocabulary (keywords)
- ✓ both text and controlled vocabulary

- ✓ with or without help of a thesaurus

2002 collection

- ✓ 150 000 documents from the PASCAL and FRANCIS databases (years 1998, 1999, 2000)
- ✓ about 2O3 Mb
- ✓ documents in french language
- ✓ 25 topics in french and english

Document example

<doc>

<docno>AM-004987</docno>

<text><ti>Le tubercule choroïdien de Bouchut révélateur d'une méningite tuberculeuse chez un enfant immunocompétent</ti>

<ab>Les auteurs rapportent un cas de papillite bilatérale associée à des tubercules de Bouchut ayant permis le diagnostic d'une méningite tuberculeuse infraclinique chez un enfant de 13 ans. Ils rappellent l'intérêt diagnostique de l'examen ophtalmologique chez les malades suspects de tuberculose.</ab></text>

<mc>Méningite, Tuberculose, Mycobacterium tuberculosis, Tubercule, Choroïde, Etude cas, Diagnostic, Homme</mc>

<kw>Meningitis, Tuberculosis, Mycobacterium tuberculosis, Tuber, Choroid, Case study, Diagnosis, Human</kw>

</doc>

Topic example

<top>
<num>020</num>
<EN-title>Biotechnologies: transgenic animals and plants</EN-title>
<EN-desc>Genetic transformation, genetic manipulation and genetic engineering are relevant subjects: genetically modified bacteria, animal and plant have been excluded.</EN-desc>
<EN-narr>
<c>Transgenic plant</c><c>Transgenic animal</c><c>Genetic transformation</c><c>Genetic engineering</c><c>Genetically modified organism</c><c>Transfection</c>
</EN-narr>
</top>

Same topic in french

<top>

<num>020</num>

<FR-title>Biotechnologies: Les animaux et plantes transgéniques</FR-title>
<FR-desc>Tout ce qui est transformations génétiques, manipulations génét
ques, génie génétique, organismes génétiquement modifiés a été considéré comme pertinent. Les bactéries, animaux et plantes inférieures génétiquement modifiées ont été exclus</FR-desc>

<FR-narr>

<c>Plante transgénique</c> <c>Animal transgénique</c> <c>Transformation génétique</c> <c>Génie génétique</c> <c>Manipulation génétique</c> <c>Organisme génétiquement modifié</c> <c>Transfection</c>

</FR-narr>

</top>

Thesaurus

- ✓ 174 000 entries
- ✓ controlled vocabulary with some thesaurus relationships

Vocabulary example

```
<RECORD>
   <TERMFR>Comté Gaspe-Ouest Québec</TERMFR>
   <TRADENG>Gaspe-Ouest County Quebec</TRADENG>
   <SYNOFRE>
      <SYNOFRE1>Gaspe Ouest</SYNOFRE1>
   </SYNOFRE>
   <AUTOP>
      <AUTOP1>Québec</AUTOP1>
   </AUTOP>
</RECORD>
```

Relevance assessments

- ✓ pooling method
- ✓ revision by a team of 10 assessors
- ✓ each specialized in the main domain (e.g. biologist for biology , etc…)

Retrieval results

- ✓ Pool of documents : 6950 (278 / topic)
- ✓ Relevant documents : 2018 (80.72 / topic (6% to 78% / topic)
- ✓ Irrelevant documents : 4932 (197.28 /topic)

Participation

- ✓ 16 teams
- ✓ 3 teams at least

Why ?

Conclusion

- ✓ AMARYLLIS was created to be the first quantitative evaluation action on French language
- ✓ now part of european CLEF on scientific texts
- ✓ loss of participants

Thank you