

COLLECTE DE DONNEES BIOLOGIQUES A PARTIR DE SOURCES MULTIPLES ET HETEROGENES

Marie-Dominique DEVIGNES

Malika SMAIL

*Collaboration Langue et Dialogue LORIA - Nancy /
Genexpress CNRS-FRE2571 - Villejuif*

PLAN DE L'EXPOSE

Introduction : objectifs

I. Analyse du problème de la collecte de données biologiques

- I.1. Exemple de requête complexe
- I.2. Analyse du problème
- I.3. Propositions

II. Une solution dédiée : le projet Xmap

- II.1. Un assistant interactif
- II.2. Une version automatisée
- II.3. Exploitation des données collectées

III. Une solution générique : le projet Xprom

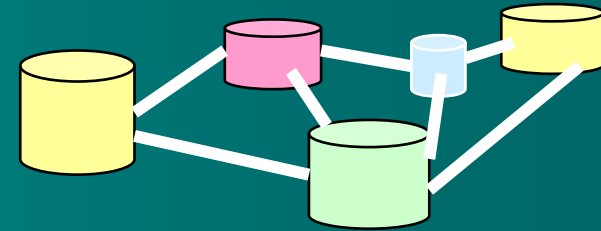
- III.1. Modélisation du scénario de collecte
- III.2. Représentation générique des données
- III.3. Mise en oeuvre

Conclusion et perspectives

- 1. Problèmes restés en suspens
- 2. Projet d'annuaire et de fédération de bases de données biologiques

Objectifs

Constat 1: *La réponse à une question donnée en biologie implique souvent l'interrogation d'une grande variété de sources.*



Constat 2: *Les sources de données biologiques et leurs contenus sont évolutifs.*



2. Automatiser la collecte des données

1. Aider l'utilisateur à accéder aux ressources pertinentes sur Internet.



3. Permettre l'organisation des données recueillies en vue du stockage, de l'analyse, de la visualisation ...

I. Analyse du problème de la collecte de données biologiques

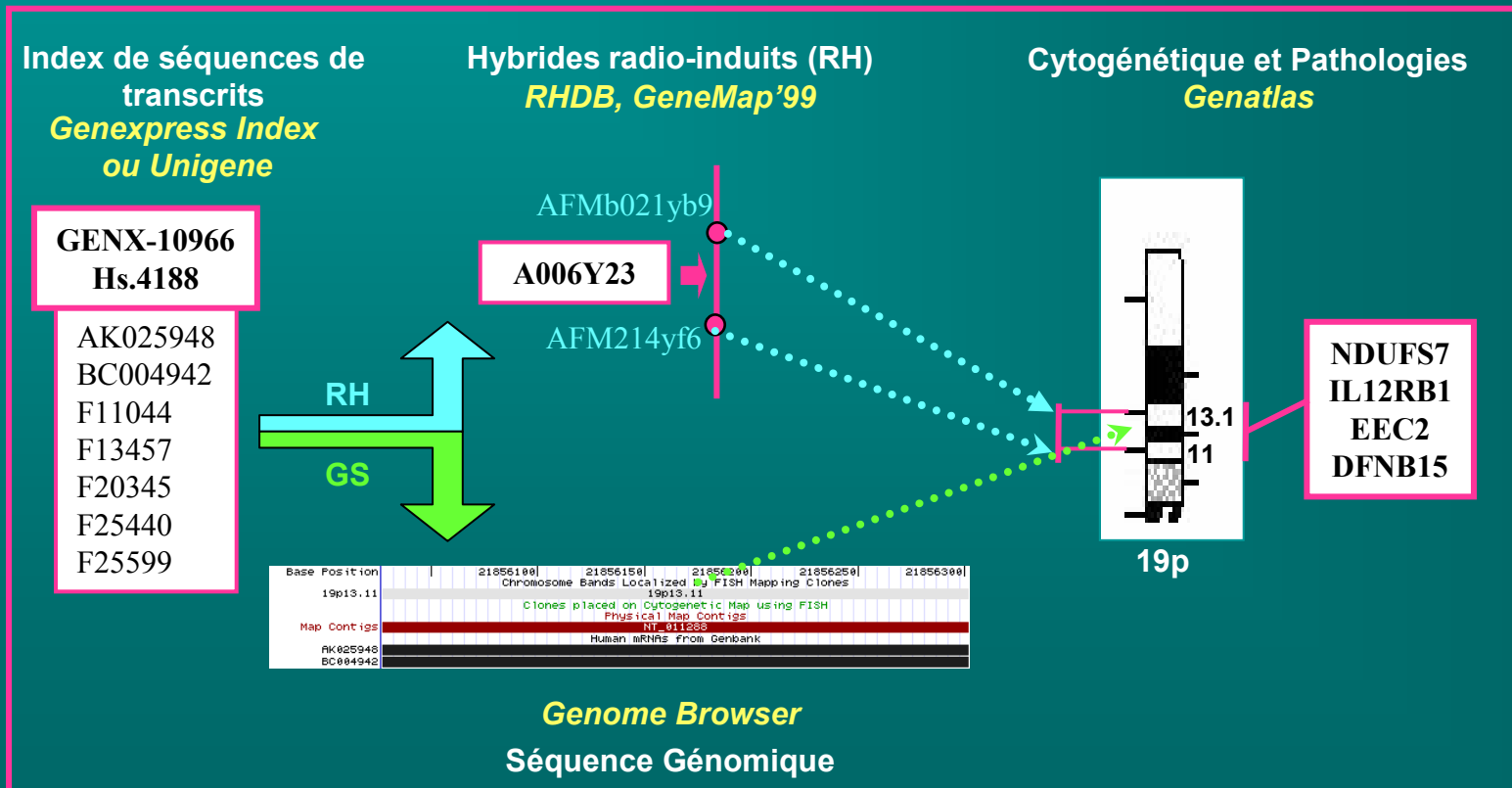
II. Une solution dédiée :
le projet Xmap

III. Une solution générique:
le projet Xprom

I. Analyse du problème de la collecte de données biologiques

I.1 Exemple de requête complexe

But : Mettre en relation de nouveaux gènes avec des pathologies orphelines sur la base de leur co-localisation



I. Analyse du problème de la collecte de données biologiques

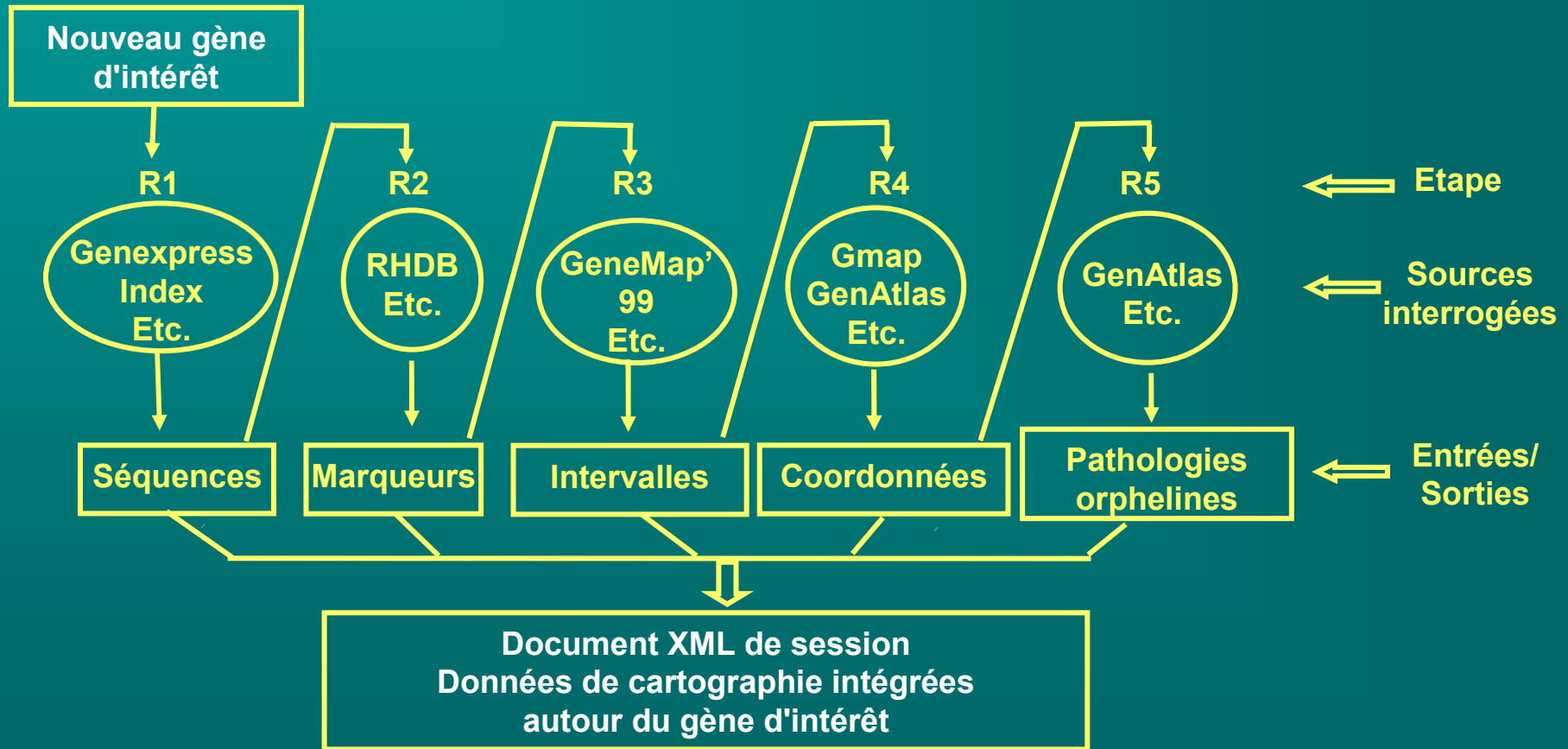
I.2 Analyse du problème

- **Données à collecter et ressources** : multiples, réparties, redondantes, évolutives, contradictoires, de fiabilité difficile à estimer...
- **Requêtes** : différentes selon les sites, filtrage des résultats, répétitives, enchaînées ...
- **Organisation des données collectées** : trier, hiérarchiser, mettre à jour ...
- **Formaliser le processus de collecte de données** : succession d'étapes constituées par l'interrogation d'une ressource distante ou locale
- **Intégration des résultats** : à l'aide de documents structurés

I. Analyse du problème de la collecte de données biologiques

I.3 Propositions (1)

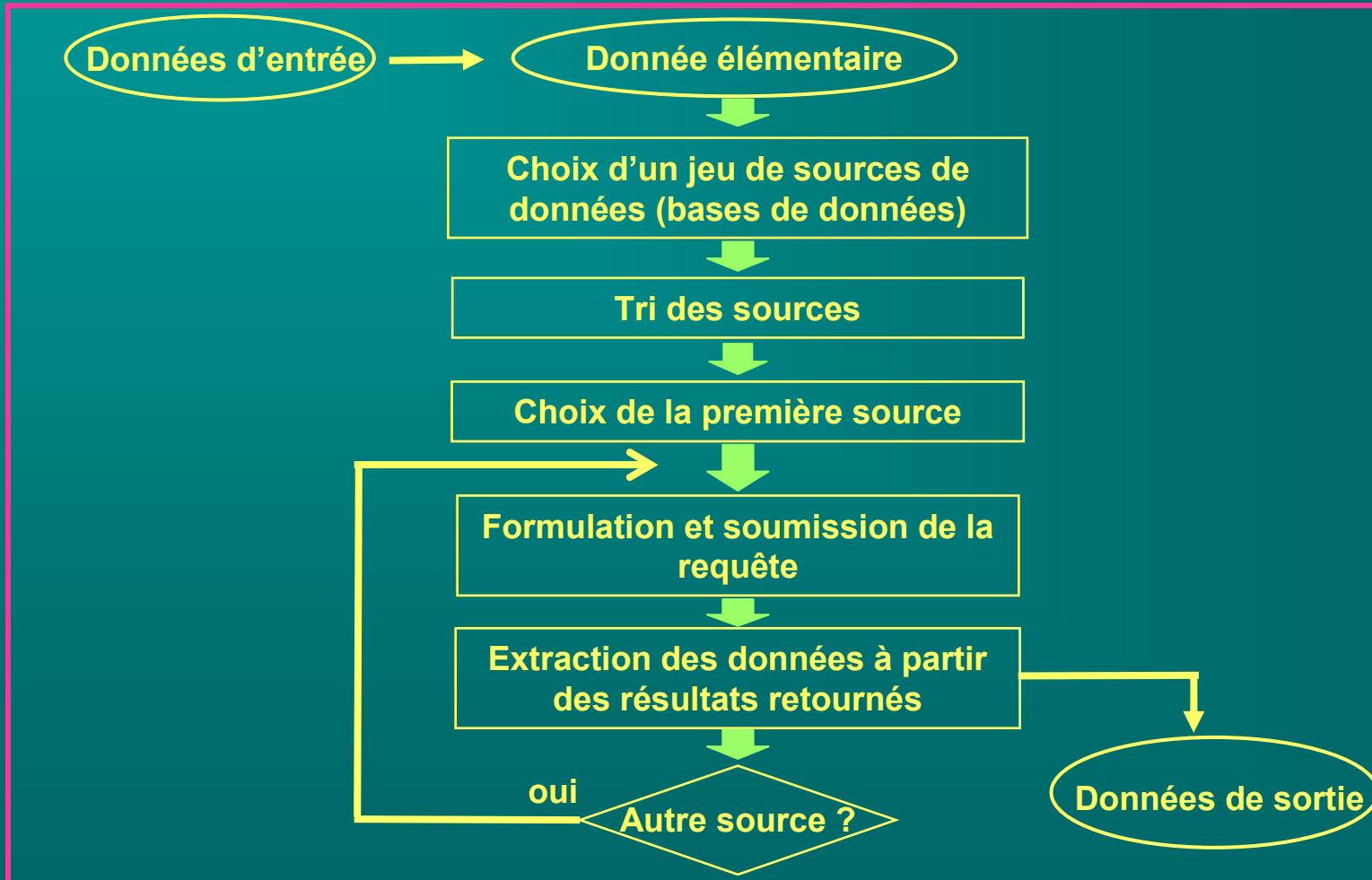
➤ Modélisation d'un scénario de collecte de données : exemple scénario RH



I. Analyse du problème de la collecte de données biologiques

I.3 Propositions (2)

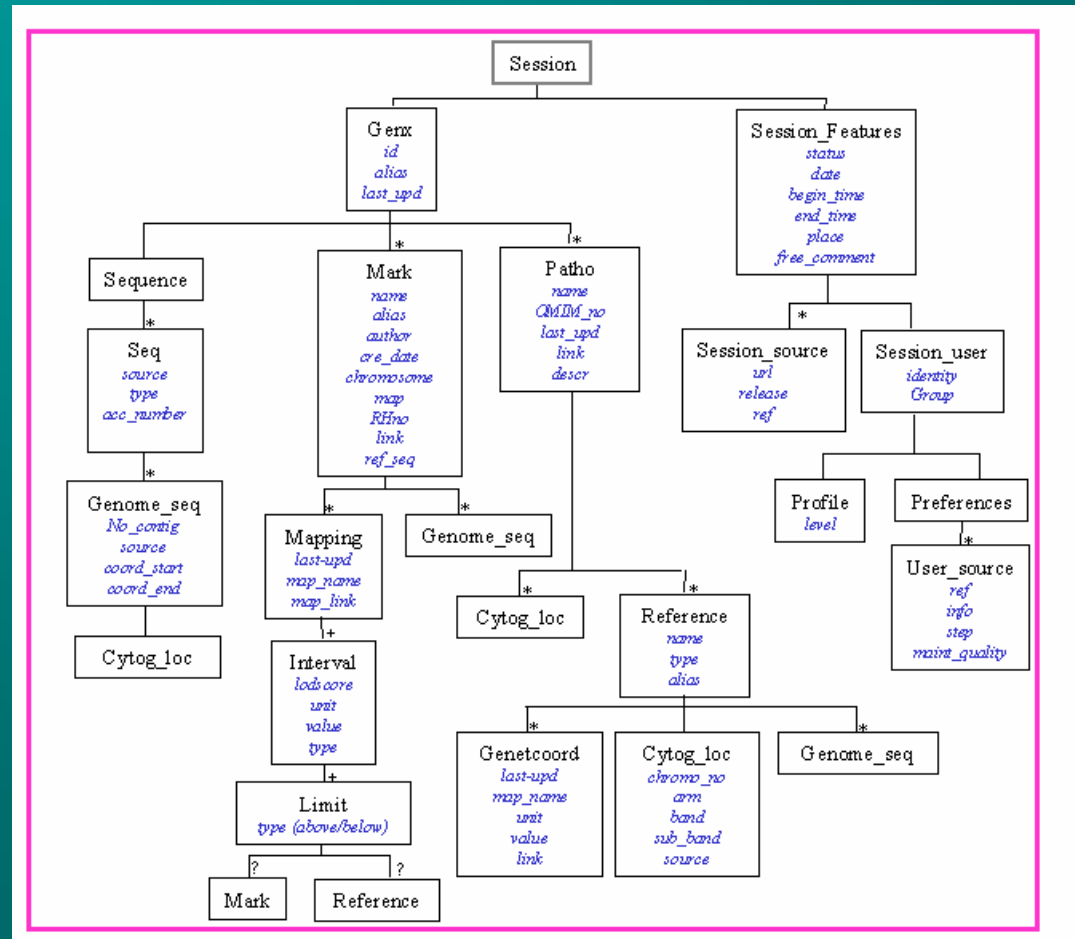
➤ Modélisation du processus d'interrogation des sources



I. Analyse du problème de la collecte de données biologiques

I.3 Propositions (3)

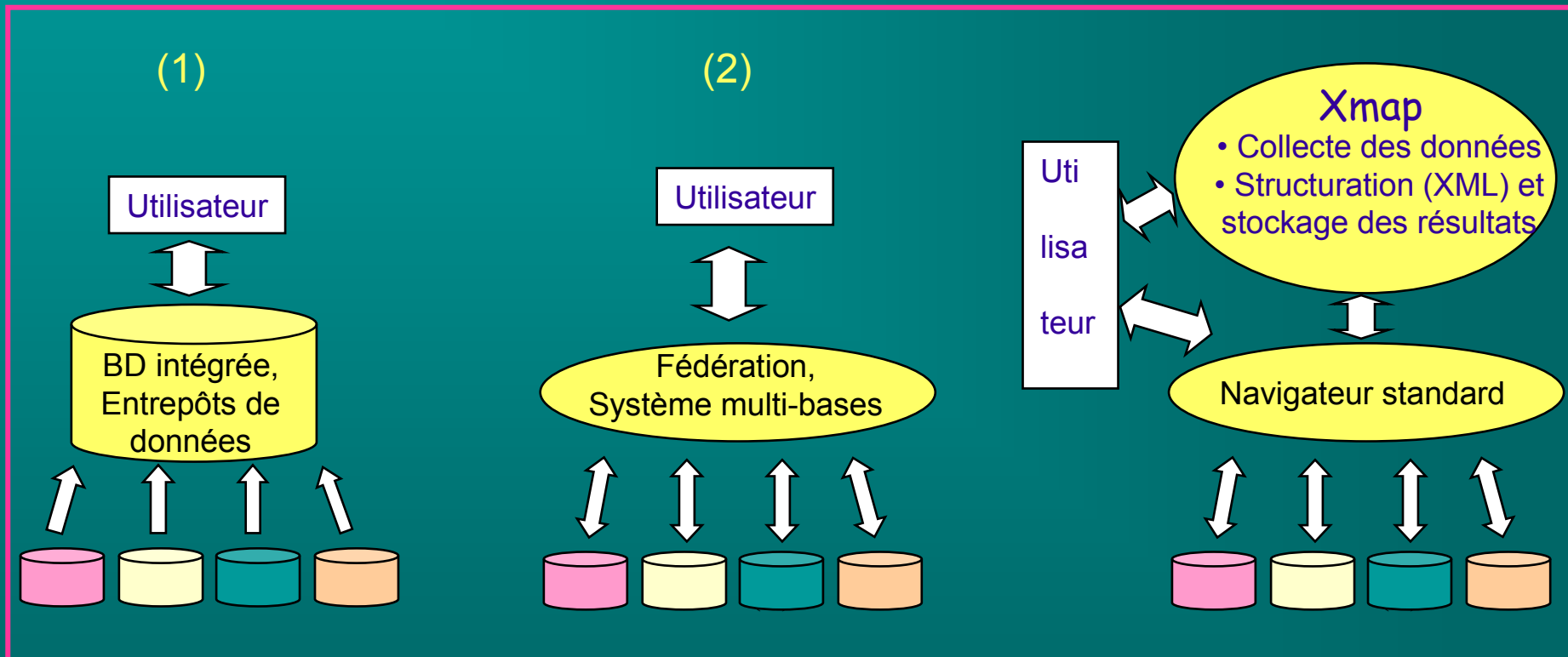
➤ Modélisation du document de session (DTD XML) : exemple Session Xmap



I. Analyse du problème de la collecte de données biologiques

I.3 Propositions (4)

- Une approche apparentée aux systèmes multi-bases



I. Analyse du problème de la collecte de données biologiques

II. Une solution dédiée : le projet Xmap

III. Une solution générique: le projet Xprom

II. Une solution dédiée : le projet Xmap

I.1. Un assistant interactif

The image displays two windows from the Xmap application. The left window, titled 'Pluxy beta-4', shows a multi-step data entry interface with tabs for 'R1', 'R2', 'R3', and 'R4'. The 'R4' tab is active, showing fields for 'Genetmark' (AFM164zb8), 'Map name' (GMAP), 'Unit' (cM), and 'Value' (19.1). A callout box points to the 'Genetmark' field with the text 'Saisie manuelle (copier-coller) des données'. Another callout points to the 'Map name' field with 'Chargement des données collectées à l'étape précédente'. A third callout points to the 'Value' field with 'Pour chaque étape : Interrogation, Filtrage, Visualisation'. A fourth callout points to the 'R4' tab with 'Succession des étapes'. A fifth callout points to a link at the bottom with 'Aide en ligne'. The right window, titled 'Entry Page - Netscape', shows the result of the query, displaying a detailed entry for 'GMAP-AFM164zb8' with fields for ID, OS, DT, RA, RL, DR, PS, PL, CH, and CM. A callout box points to the 'CM' field with the text 'Navigateur affichant la réponse à la requête'.

Pluxy © Dyade
INRIA Rocquencourt

Succession des étapes

Navigateur affichant la réponse à la requête

Pour chaque étape :
Interrogation,
Filtrage,
Visualisation

Chargement des données
collectées à
l'étape
précédente

Saisie manuelle
(copier-coller)
des données

Aide en ligne

Etape R4

aide sur l'étape R4...

Entry Page - Netscape

TOP PAGE QUERY RESULTS SESSIONS VIEWS DATABANKS

Reset View * Complete entries *

This entry is from: [GMAP-AFM164zb8](#)

[GMAP](#)

Save

Link

Printer Friendly

ID AFM164zb8
XX
OS Homo sapiens
XX
DT 27-OCT-1998
RN [1]
RA Dib C. et al.
RL A comprehensive genetic map of the human genome b
5264 microsatellites
RL Nature, 390, 152-154 1996).
RL ftp://ftp.genethon.fr/pub/Gmap
XX
DR EMBL: [216743](#)
DR GDB: [D19S216](#)
PS TCTTGCTACTCTAACCCTCCGC
PS GGCCCATGCTTTTTTAGGT
PL 179-191 bp (allele sizes range)
XX
CH 19
CM 19.1: 16.4; 23.3. (cM : Sex-averaged, female and
XX
AN allele_number size frequency
AL 1 191 0.241
AL 2 185 0.315
AL 3 179 0.241
AL 4 187 0.148
AL 5 189 0.056
//

Interface utilisateur de Xmap_INTERACTIVE (André Schaaff, 2000)

II. Une solution dédiée : le projet Xmap

I.1. Un assistant interactif (2)

Les « + » de Xmap_INTERACTIVE

- Assistance dans le choix des sources à interroger (utilisateur novice)
- Possibilité d'interroger de nouvelles sources (utilisateur expert)
- Respect de l'autonomie des sources (dernières mises à jour)
- Prise en charge possible de la formulation des requêtes
- Intégration et sauvegarde des données dans un document structuré (XML)
- Visualisation des données selon diverses feuilles de style

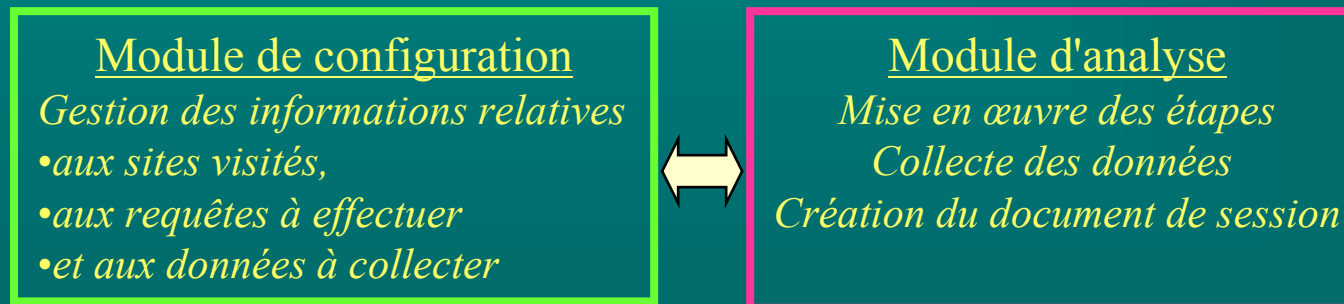
II. Une solution dédiée : le projet Xmap

II.2. Une version automatisée (1)

Xmap_AUTO1 (*Emmanuel Touzery, 2000; Yvan Norsa, 2001*)

But : Exécuter de façon automatique l'ensemble des étapes prévues par le scénario

- Scénario minimum : interrogation d'un seul site (le plus pertinent) pour chaque étape
- Permet de traiter une série d'entrées
- Extraction du passage recherché basé sur la reconnaissance d'expressions régulières (génération d'analyseurs lexicaux)
- Résultat de session au format XML compatible avec l'application Xmap_INTERACTIVE
=> affiner la session



II. Une solution dédiée : le projet Xmap

II.2. Une version automatisée (2)

Xmap_AUTO versus Xmap_INTERACTIF

Etape i	Xmap_INTERACTIF	Xmap_AUTO
SITE INTERROGE	Plusieurs sites proposés, ouvert	Un seul : le « meilleur »
INTERROGATION	Formulation par l'utilisateur, aide	Syntaxe + paramètres dans fichier de configuration
EXTRACTION	Copier-coller	Expressions régulières : avant / passage recherché / après --> génération d'analyseurs lexicaux
DOCUMENT DE SESSION	Selon DTD Xmap	Selon DTD Xmap

II. Une solution dédiée : le projet Xmap

II.2. Une version automatisée (3)

GENX XMap_Auto 4

/ Deroulement **/**

Choix du scenario :

Scenario RH Scenario GS Scenario RH + Scenario GS

Choix de la source de la derniere etape (recherche des pathologies) :

Locuslink Genatlas

/ Entree **/**

Saisie Unique :

Numero Genx Numero d'entree ou chemin du fichier d'entrees :

Numero Unigene

Numero d'accession de Sequence

Saisie Multiple :

Fichier (liste Genx)

Fichier (liste Unigene)

Fichier (liste de numeros d'accession)

/ Sortie **/**

Choix du repertoire des fichiers XML sortie

➤ Mise en œuvre de deux scénarios complémentaires

➤ Possibilité de choix sur les sources interrogées

➤ Tests et implication dans le Workshop d'Annotation des cDNA humains (Tokyo, Août 2002)

➤ Mise à disposition prévue sur la plateforme Bio-Info du Loria (*Laurent Pierron, 2002*)

Interface utilisateur de Xmap_AUTO4
(*Hervé de Palma, 2002*)

II. Une solution dédiée : le projet Xmap

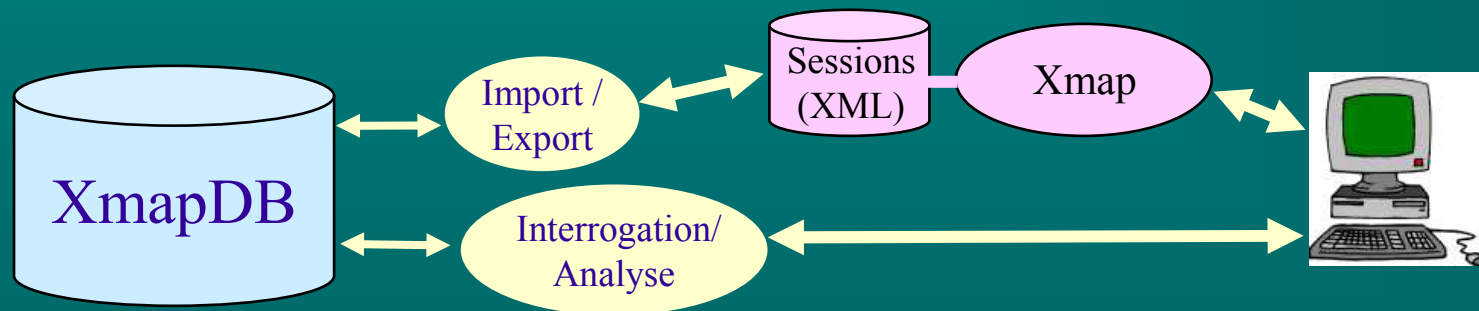
II.3 Exploitation des données collectées (1)

La base de données Xmap_DB

(Vincent Strohmenger, 2000-2001 ; Laurent Pierron, 2002)

But : Centraliser et exploiter les sessions Xmap

- Utiliser les avantages d'un système de gestion de base de données relationnelle
- Rendre accessibles les données de plusieurs sites (import-export)
- Permettre l'analyse de mémoire de sessions (statistiques)
- Maintenance unique



II. Une solution dédiée : le projet Xmap

II.3 Exploitation des données collectées (2)

Mise en œuvre de Xmap_DB

MAPPING

(XPath)

Modèle de données (DTD)

Schéma relationnel

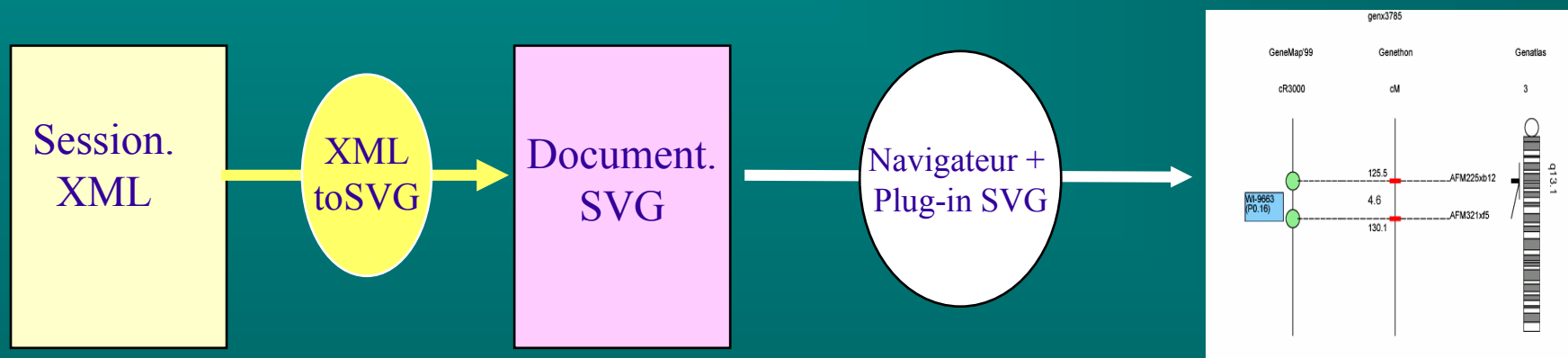
II. Une solution dédiée : le projet Xmap

II.3 Exploitation des données collectées (3)

Xmap_SHOW (*Yvan Norsa, 2001*)

But : Visualiser la position d'un gène et des pathologies co-localisées sur le chromosome

- Convertir le document de session XML en un format compatible avec des outils de visualisation
- Permettre la visualisation par le client



Utilisation de Scalable Vector Graphics : SVG (spécification du W3C)

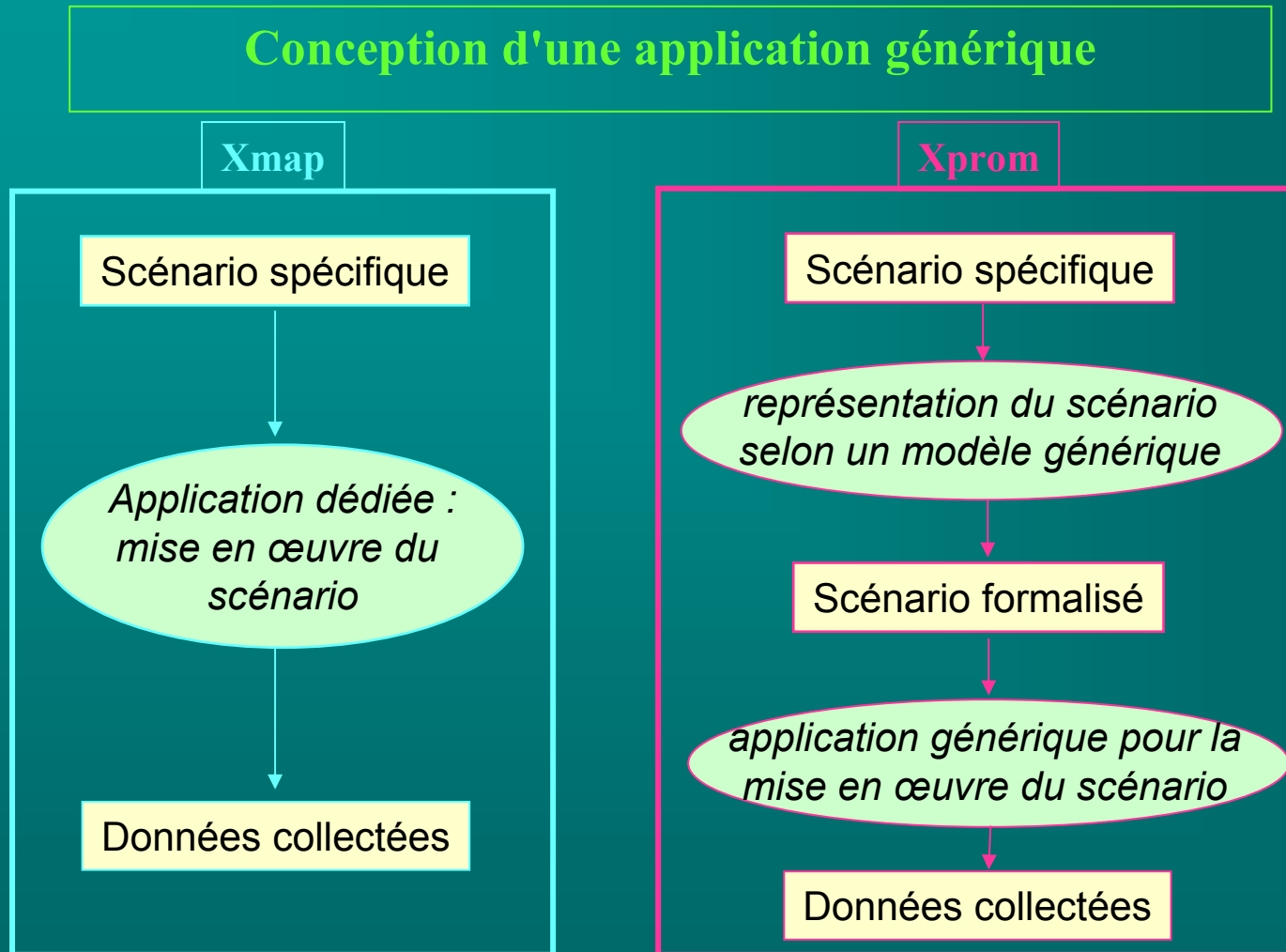
I. Analyse du problème de la collecte de données biologiques

**II. Une solution dédiée :
le projet Xmap**

**III. Une solution générique:
le projet Xprom**

III. Une solution générique : le projet Xprom

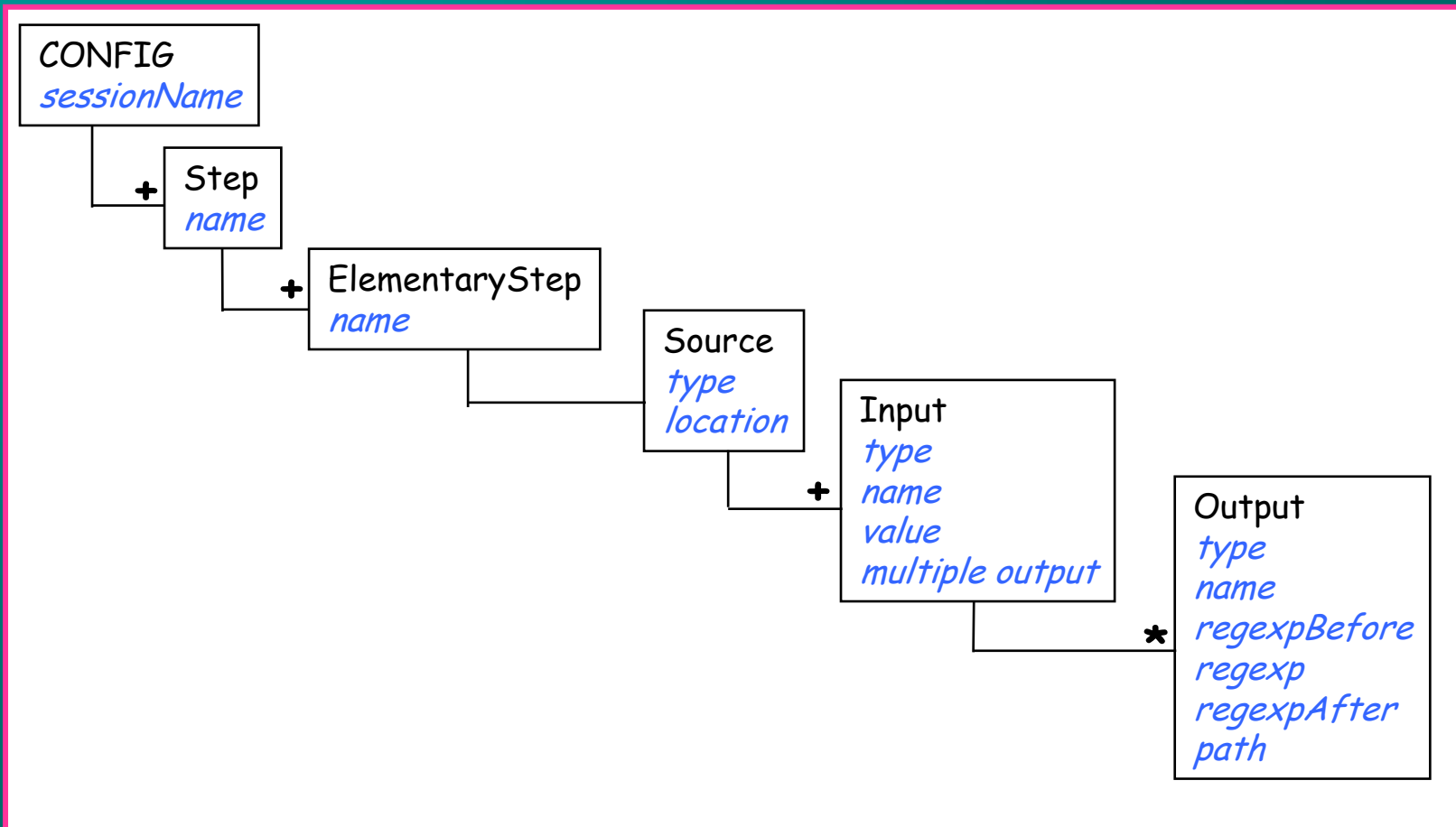
III.1. Modélisation du scénario de collecte (1)



III. Une solution générique : le projet Xprom

III.1. Modélisation du scénario de collecte (2)

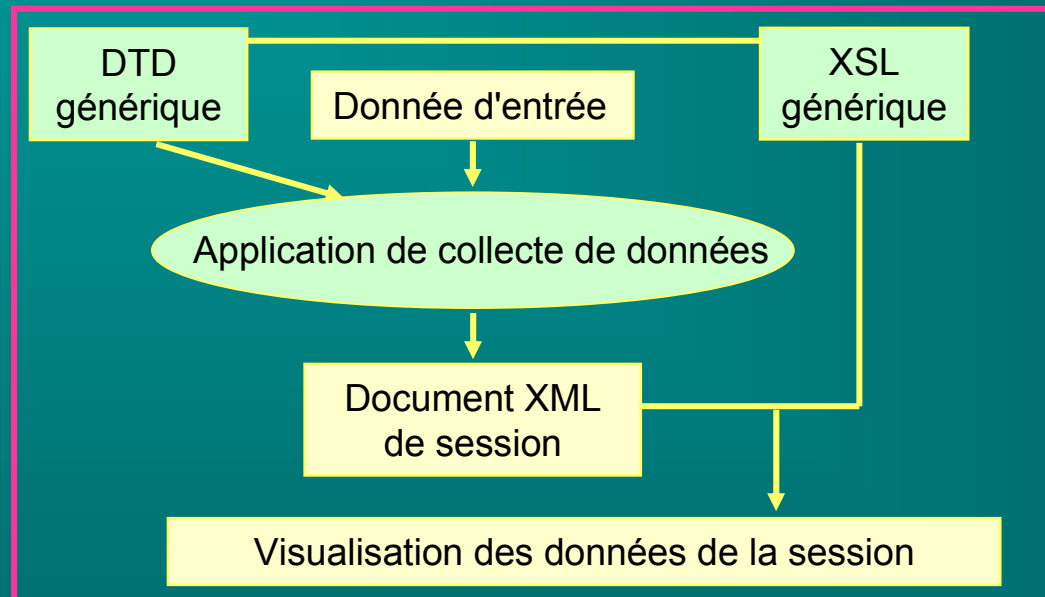
DTD de scénario générique



III. Une solution générique : le projet Xprom

III.2. Représentation générique des données (1)

Choix d'une DTD de session générique



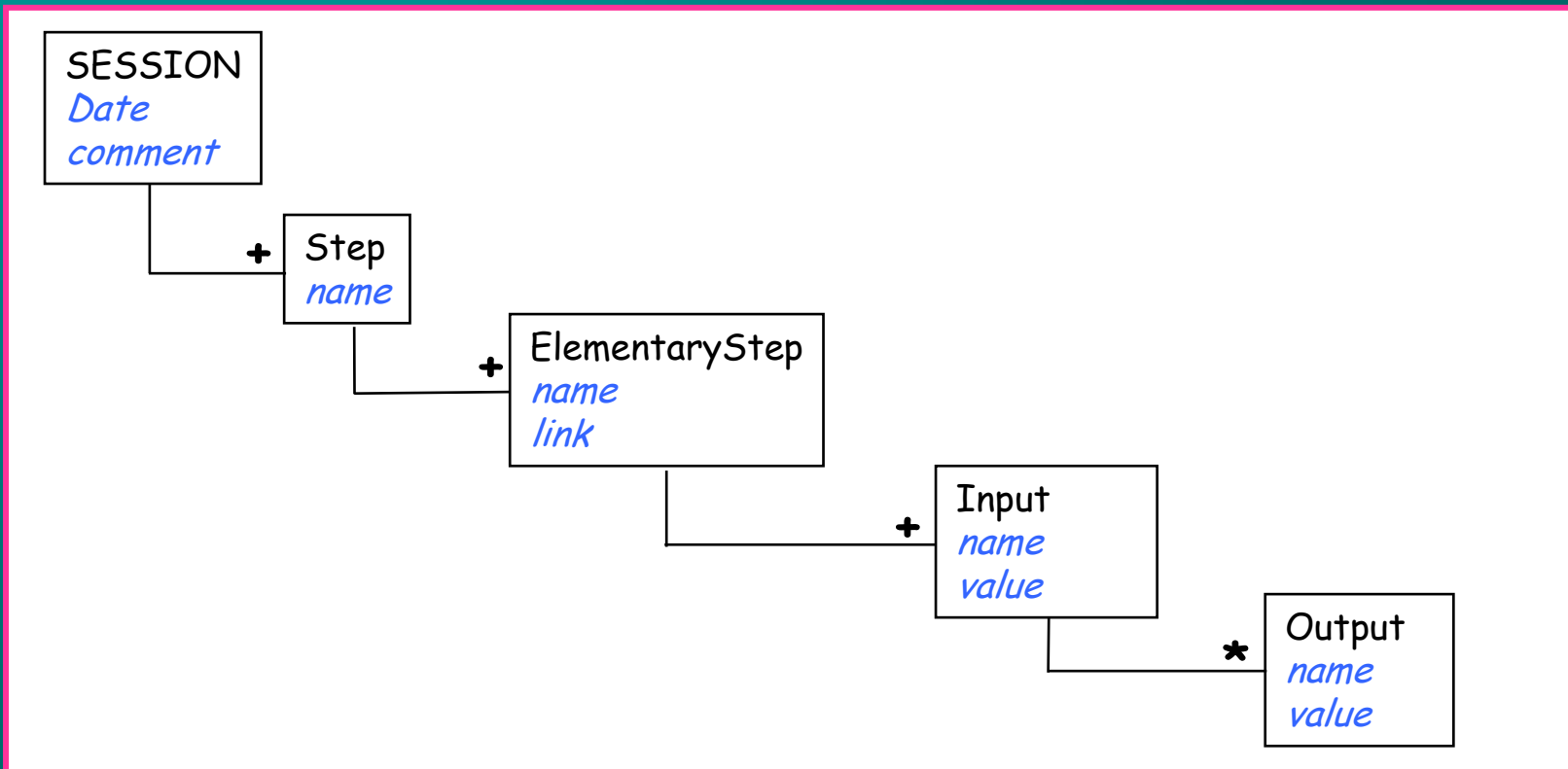
Structure du document de session orientée par le scénario

Possibilité de le convertir en une autre structure plus signifiante

III. Une solution générique : le projet Xprom

III.2. Représentation générique des données (2)

DTD de session générique



III. Une solution générique : le projet Xprom

III.3. Mise en œuvre (1)

Les deux modules de l'application Xprom (*Yvan Norsa, 2002*)

1. Module de Configuration

- Interface de saisie du scénario spécifique
- Conversion en document XML utilisable par le module de recherche

2. Module de Recherche

- Liste des étapes et sous-étapes
- Pour chaque sous-étape :
 - Formulation de la requête
 - Soumission
 - Récupération du document retourné
 - Extraction des données à collecter
 - Sauvegarde dans document XML de session
 - Enchaînement sous-étape suivante

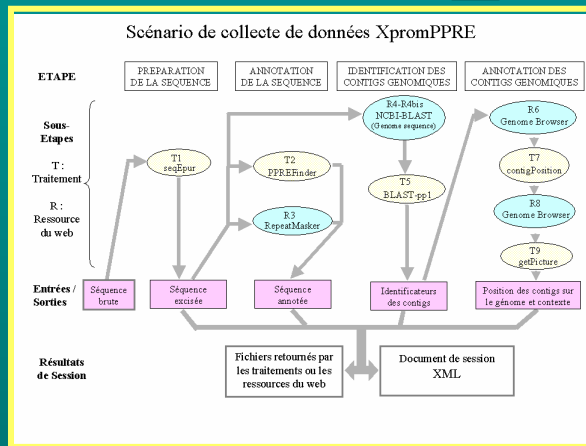
III. Une solution générique : le projet Xprom

III.3. Mise en œuvre (2)

Application : XpromPPRE

recherche du contexte d'une séquence dans le génome humain

Configuration



```
<?xml version="1.0"?>
<!DOCTYPE CONFIG SYSTEM "config.dtd">
<CONFIG sessionName="rawSeq">
  <Step name="eparation">
    <ElementaryStep name="seqPur">
      <Source type="Class" location="SeqPur.class">
        <Input type="String" name="path" multipleOutputs="false" value=""/>
        <Input type="String" name="motif" multipleOutputs="false" value=""/>
      </Input>
      <Output type="File" name="rawSeq" multipleOutputs="false" value=""/>
    </ElementaryStep>
  </Step>
  <Step name="RepeatMasker">
    <ElementaryStep name="RepeatMasker">
      <Source type="Class" location="RepeatMasker.class">
        <Input type="String" name="motif" multipleOutputs="false" value=""/>
      </Input>
      <Input type="File" name="excisedSeq" multipleOutputs="false" value=""/>
      <Output name="result" type="String" reseq_before="" reseq_after="" path=""/>
    </ElementaryStep>
  </Step>
  <Step name="BLAST pp1">
    <ElementaryStep name="BLAST pp1">
      <Source type="Class" location="BLAST.class">
        <Input type="String" name="motif" multipleOutputs="false" value=""/>
      </Input>
      <Input type="File" name="excisedSeq" multipleOutputs="false" value=""/>
      <Output name="result" type="String" reseq_before="" reseq_after="" path=""/>
    </ElementaryStep>
  </Step>
  <Step name="GenomE Browser">
    <ElementaryStep name="GenomE Browser">
      <Source type="Class" location="GenomE.class">
        <Input type="String" name="motif" multipleOutputs="false" value=""/>
      </Input>
      <Input type="File" name="excisedSeq" multipleOutputs="false" value=""/>
      <Output name="result" type="String" reseq_before="" reseq_after="" path=""/>
    </ElementaryStep>
  </Step>
  <Step name="getFichiers">
    <ElementaryStep name="getFichiers">
      <Source type="Class" location="getFichiers.class">
        <Input type="String" name="motif" multipleOutputs="false" value=""/>
      </Input>
      <Input type="File" name="excisedSeq" multipleOutputs="false" value=""/>
      <Output name="result" type="String" reseq_before="" reseq_after="" path=""/>
    </ElementaryStep>
  </Step>
</CONFIG>
```

Visualisation de Session Xprom
Session A402-B01_DOMENJOURD-1009-B01-T3-96-F1

```

Step execution
Sub_step seqPur
local
Input data :
rawSeq=1009-B01-T3-96-F1.seq
excisedSeq=1009-B01-T3-96-F1.seq_excised.fasta
Output data :
Step seq Annotation
Sub_step pprefinder
local
Input data :
excisedSeq=1009-B01-T3-96-F1.seq_excised.fasta
Output data :
??Pas de résultats ???
Sub_step repeatMasker
local
Input data :
sequence=>Temp_A402-B01_DOMENJOURD-1009-B01-T3-96-F1_50E01227_DSabl1ABX_Testing...no comment
%0aGCTACTAAACAGCGTCCCTGAGTGGAGCTTGTCTGCCTTGCCTGCTGACACAGCGGCGCTTGTAGGAA
%0aGCAATGGCGCTGATGCCAACCTTGTAGGATAGAA
GATGTACATG%0a
Output data :
There were no repetitive sequences detected in
repeatMasker/tmp/RM2seqp_bsd_1592"
Step genomicContig
Sub_step blastQuery
local
Input data :
QUERY=
>Temp_A402-B01_DOMENJOURD-1009-B01-T3-96-F1_50E01227_DSabl1ABX_Testing...no comment
%0aGCTACTAAACAGCGTCCCTGAGTGGAGCTTGTCTGCCTTGCCTGCACTAACTGCTGTAGGAGCAGCGCTTGTAGGAA
%0aGCAATGGCGCTGATGCCAACCTTGTAGGATAGAA
GATGTACATG%0a
DB=HTGc,_DATABASE=htg_blastSleep delay=100000
Output data :
Sub_step blastXML
local
Input data :
RID=1023104810.015681.5144
FORMAT_TYPE=XML
Output data :
hitID=AL589182.3, hitFrom=26065, hitTo=26193, hitAccession=AL589182
!!Free air press & hitz retenez tous les hits ??
Sub_step genomePos
local
Input data :
position=AL589182
Output data :
baseLocation=chr14:17157565-17319401
Sub_step contigPos
local????
Input data :
contigPosition=chr14:17157565-17319401
distance=-2000, hitFrom=26065, hitTo=26193
Output data :
chromosome=14, begin=17163630, end=17203630,
locatioString=chr14:17163630-17203630
Sub_step genomeBrowser
local
Input data :
genomePFile=genome_13842_1023104920.gff
Output data :
File=genome_Pfile.gff chr14:17163630-17203630
/trash/htg_genome_13842_1023104920.gff Last modified: Fri Jun 7 14:36:05
MET DST 2002
  
```

Scénario

Seq1009-B1

Donnée d'entrée

XpromPPRE_scenario.xml

Recherche

Seq1009-B1-visu.html

Seq1009-B1-session.xml

Conclusion et perspectives

1. Problèmes restés en suspens

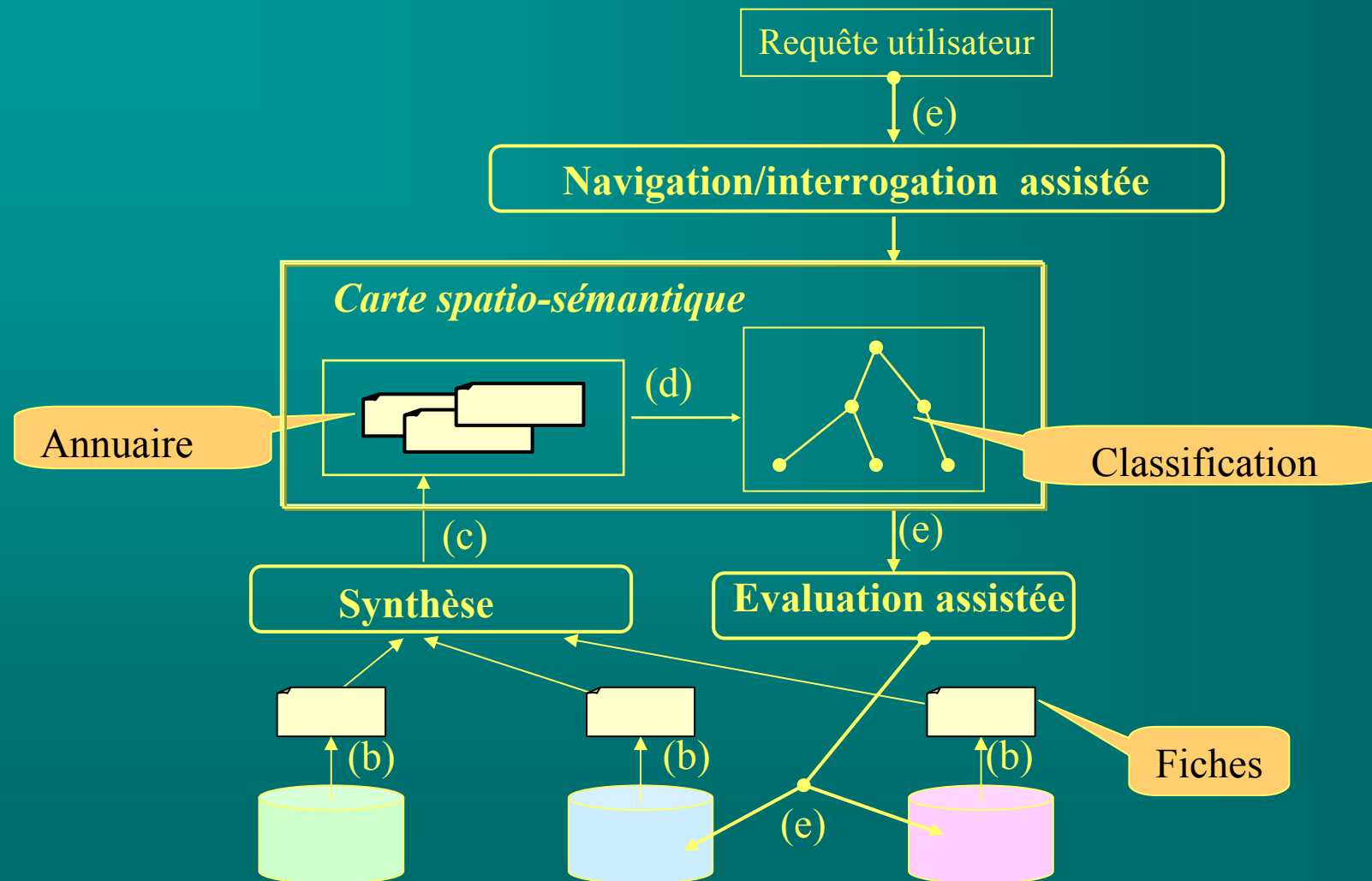
1.1. Gestion de sources multiples pour une étape donnée

- Dans version interactive, plusieurs sources sont interrogées
=> trier les réponses, gérer les conflits
- Dans version automatique, une seule source interrogée
=> choix manuel de la meilleure source
- Tri des sources
=> modifiable en fonction de résultats d'analyse de la mémoire de sessions

1.2. Spécification d'un nouveau scénario : identification des sources pertinentes

- Nécessité d'une approche orientée source

2. Approche orientée source : projet d'annuaire et de fédération de sources de données biologiques (1)



➤ **Mots-clés** : Intégration de données semi-structurées, web sémantique, « authentification » de données

2. Approche orientée source : projet d'annuaire et de fédération de sources de données biologiques (2)

But : proposer une interface de navigation et d'interrogation

☞ Requête simple :

- Sélection des sources pertinentes pour la requête
- En cas de sources multiples pertinentes, intégration et synthèse des résultats
 - ✓ Gestion des redondances, contradictions,...
 - ✓ Critères de qualité relative des données/sources

☞ Requête complexe : Scénario de collecte de données

- Modèle général de description de scénario
- Générateur de collecteur de données