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

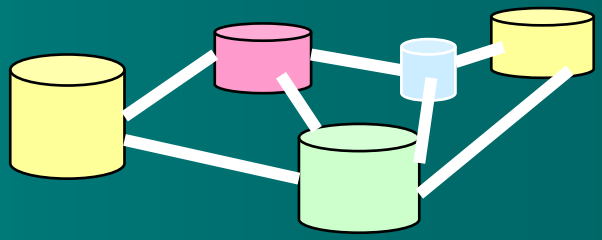
- III.1. Modélisation du scénario de collecte
- III.2. Représentation générique des données
- III.3. Mise en oeuvre

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.***

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

2. Automatiser la collecte des données

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

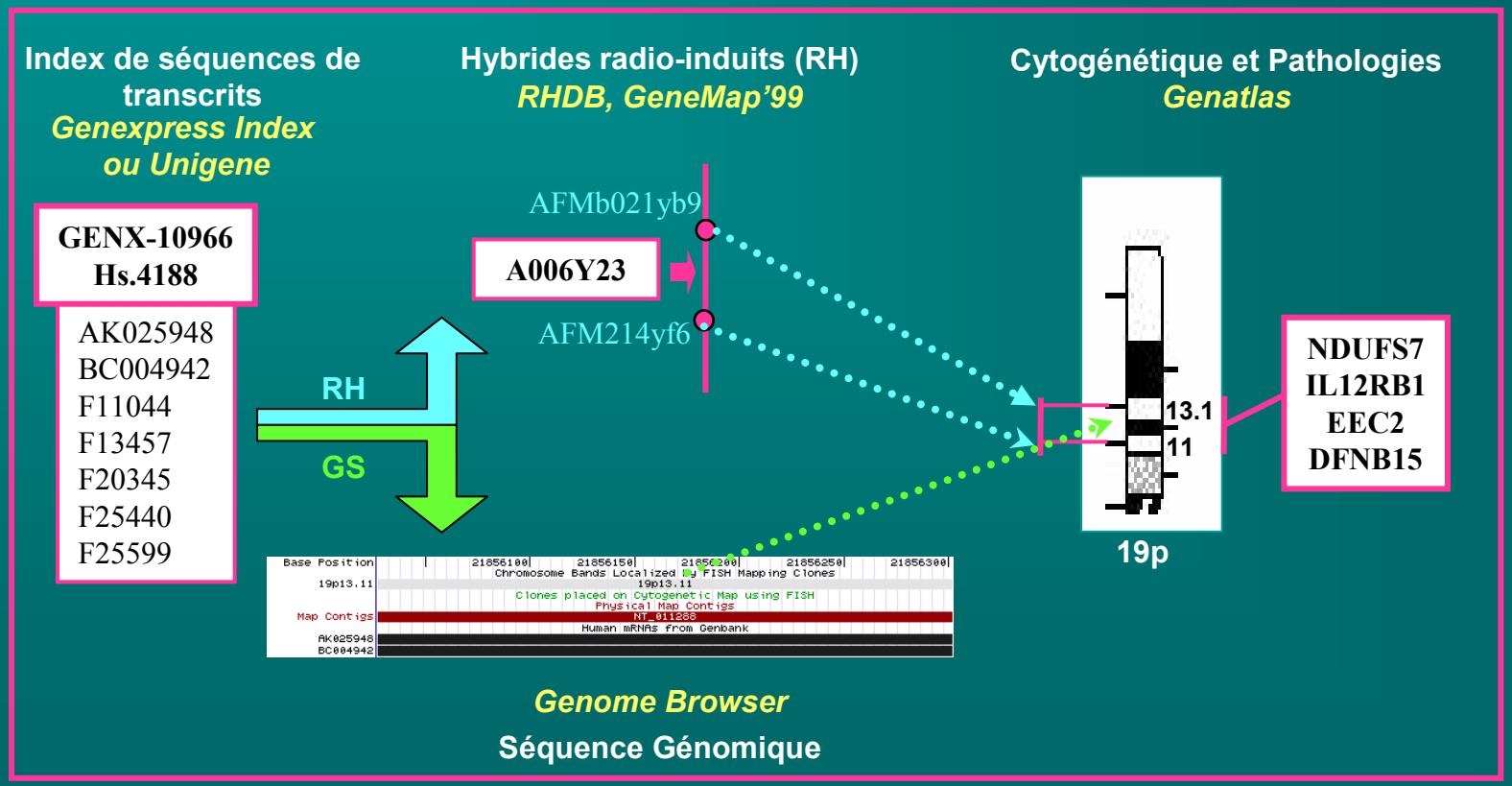
I. Analyse du problème de la collecte de données biologiques

II. Une solution dédiée : le projet Xmap

III. Une solution générique: le projet Xprom

I.1 Exemple de requête complexe

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

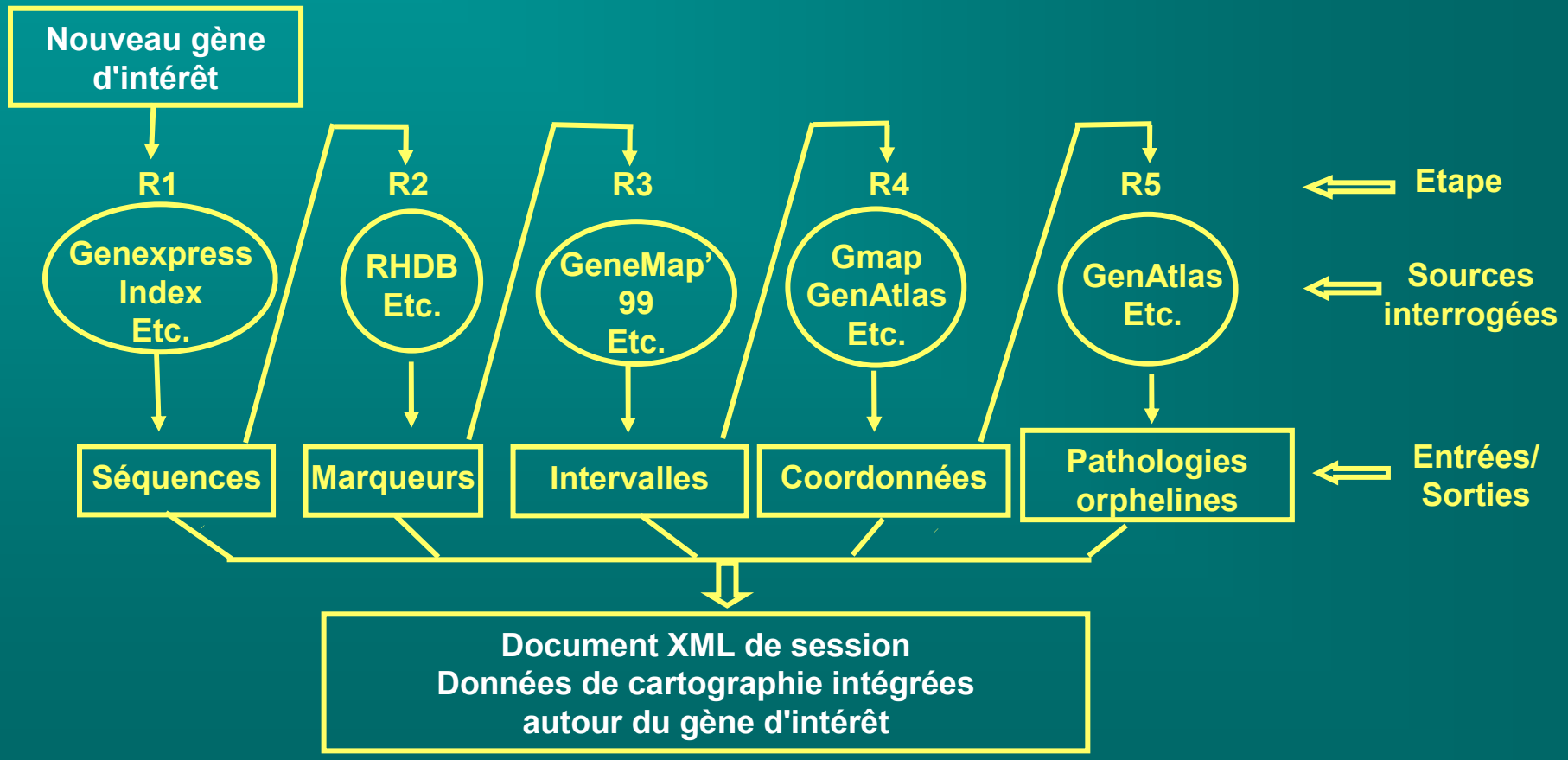


I.2 Analyse du problème

- **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

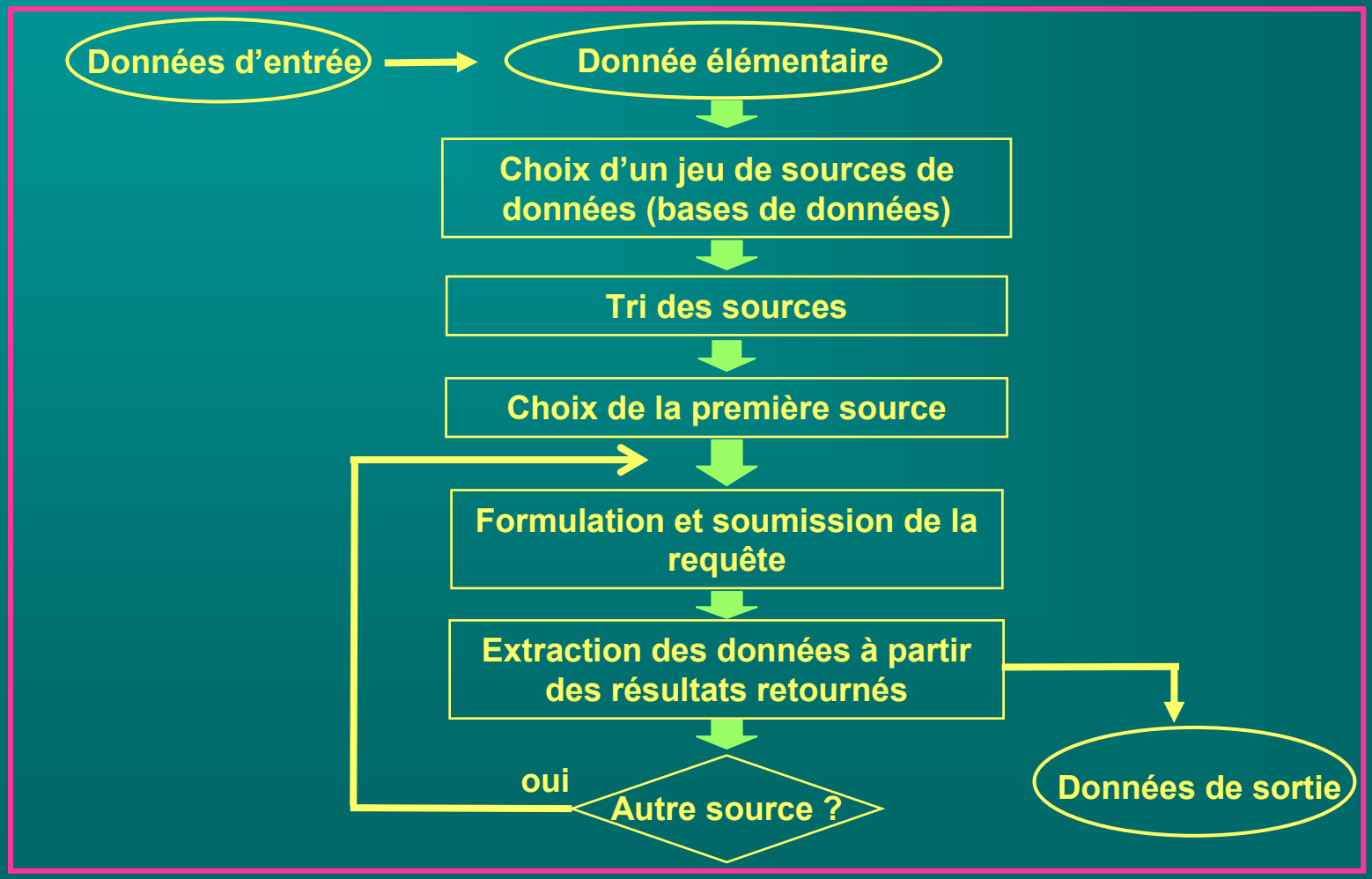
I.3 Propositions (1)

- **Modélisation d'un scénario de collecte de données : exemple scénario RH**



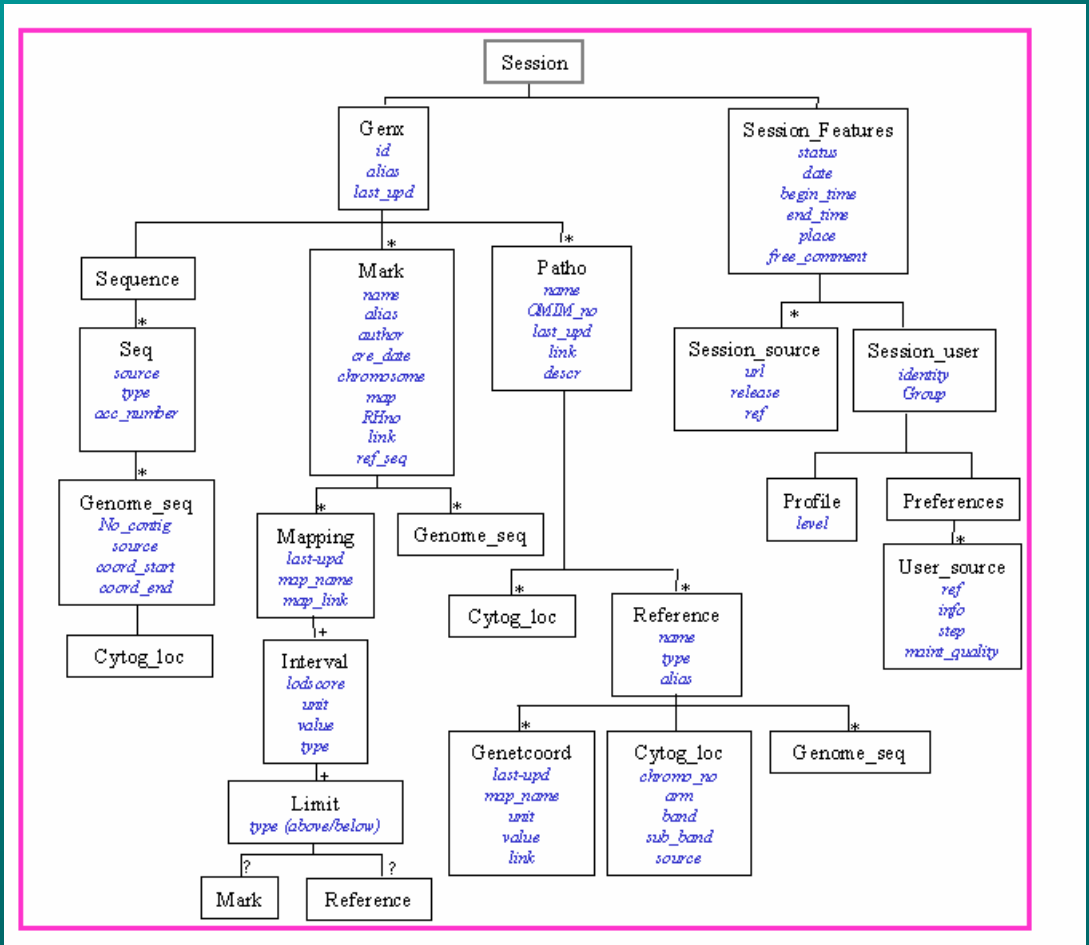
I.3 Propositions (2)

- **Modélisation du processus d'interrogation des sources**



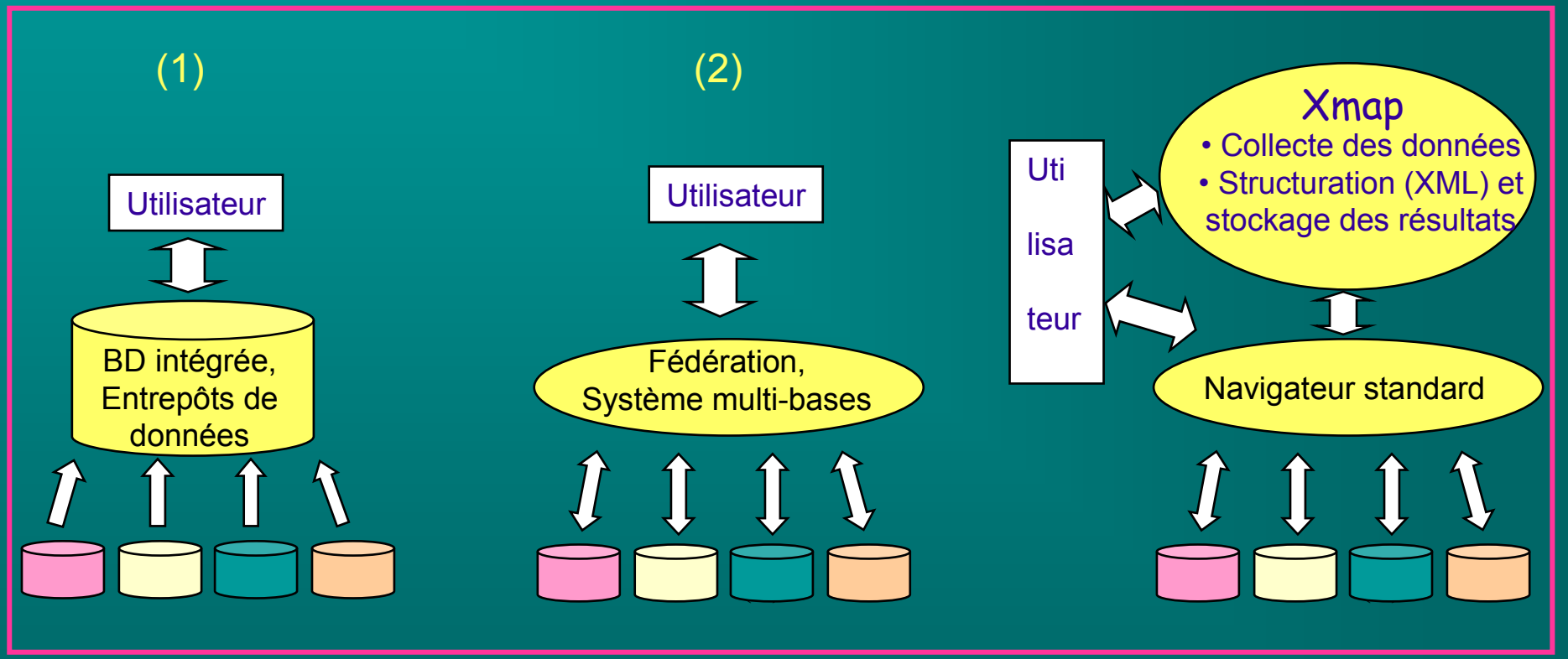
I.3 Propositions (3)

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



I.3 Propositions (4)

- Une approche apparentée aux systèmes multi-bases

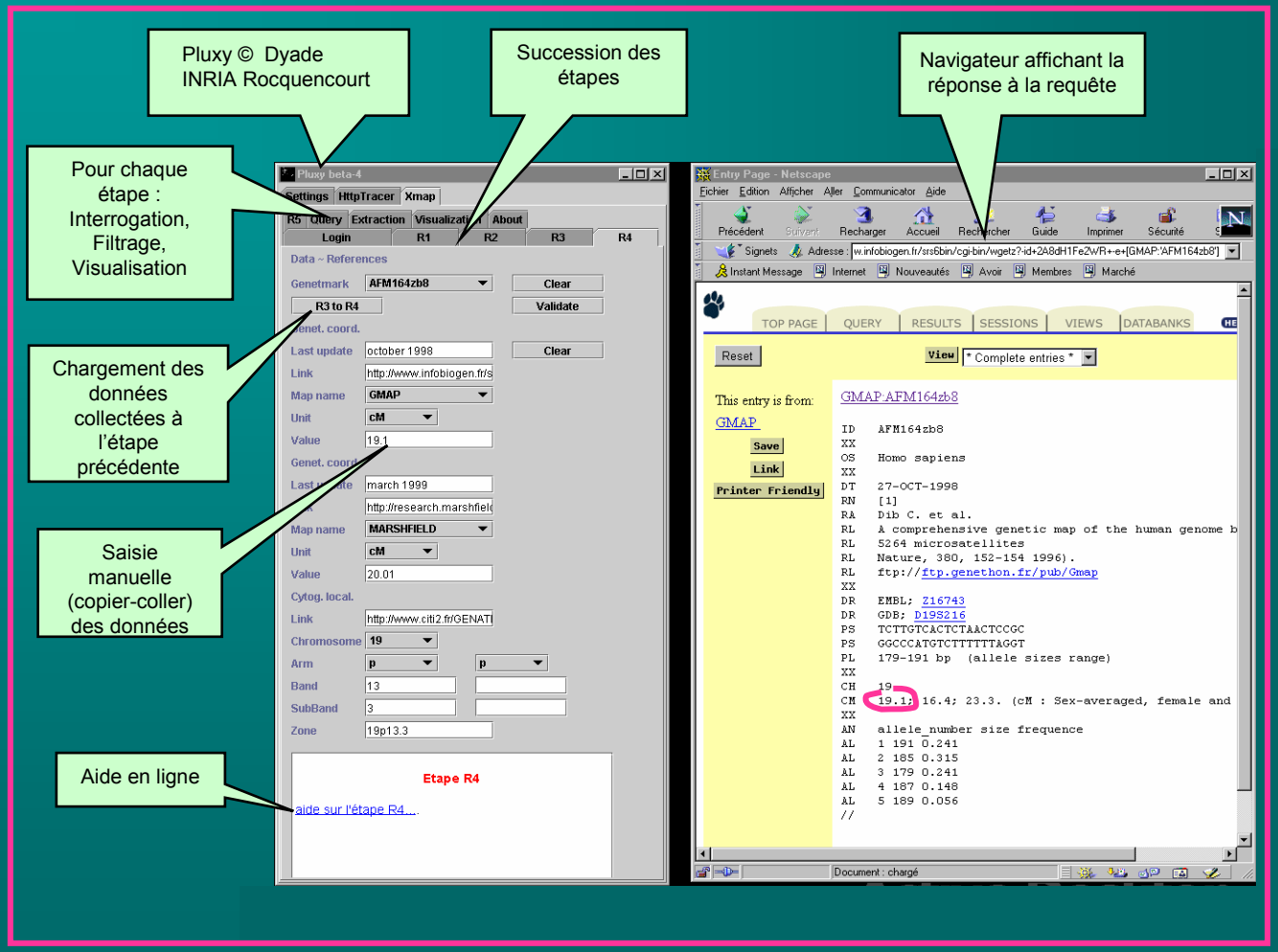


I. Analyse du problème de la collecte de données biologiques

II. Une solution dédiée : le projet Xmap

III. Une solution générique: le projet Xprom

I.1. Un assistant interactif



Interface utilisateur de Xmap_INTERACTIVE (*André Schaaff, 2000*)

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.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 reconnaissances d'expressions régulières (génération d'analyseurs lexicaux)
- Résultat de session au format XML compatible avec l'application Xmap_INTERACTIVE

=> affiner la session

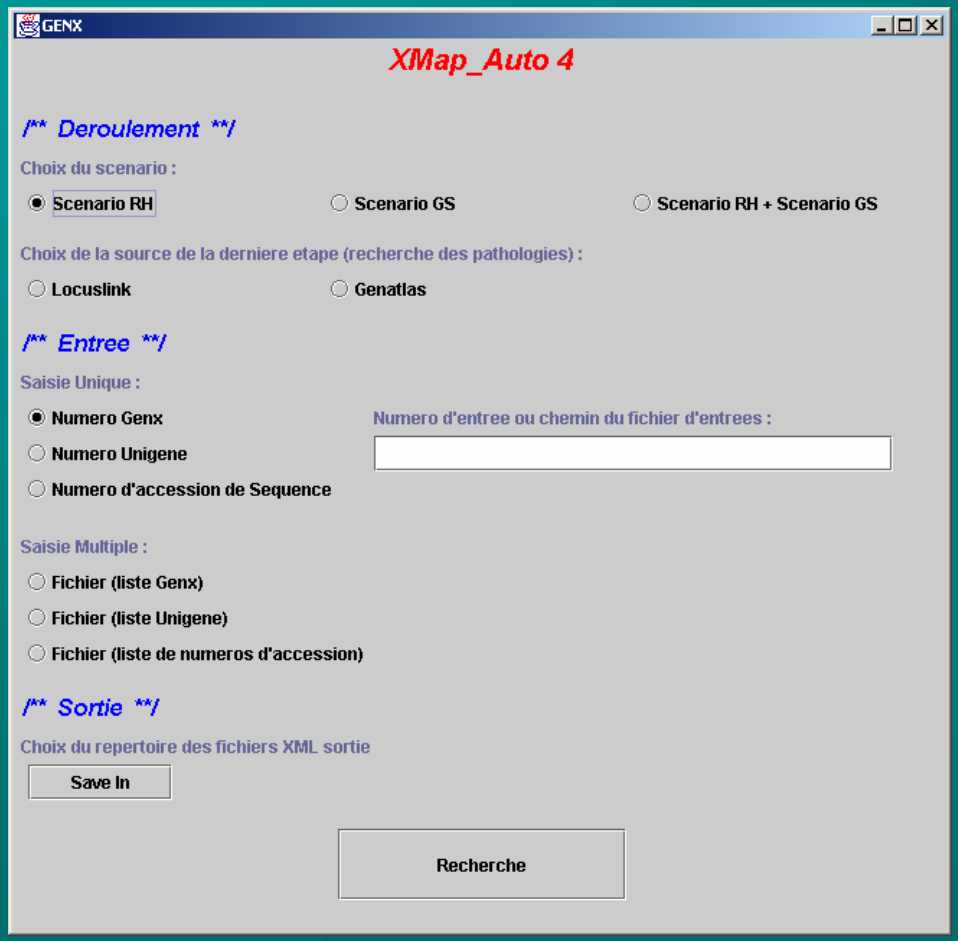
Module de configuration	⇔	Module d'analyse
Gestion des informations relatives *•aux sites visités,* *•aux requêtes à effectuer* *•et aux données à collecter*		*Mise en œuvre des étapes* *Collecte des données* *Création du document de session*

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.2. Une version automatisée (3)



Interface utilisateur de Xmap_AUTO4
(Hervé de Palma, 2002)

- 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)*

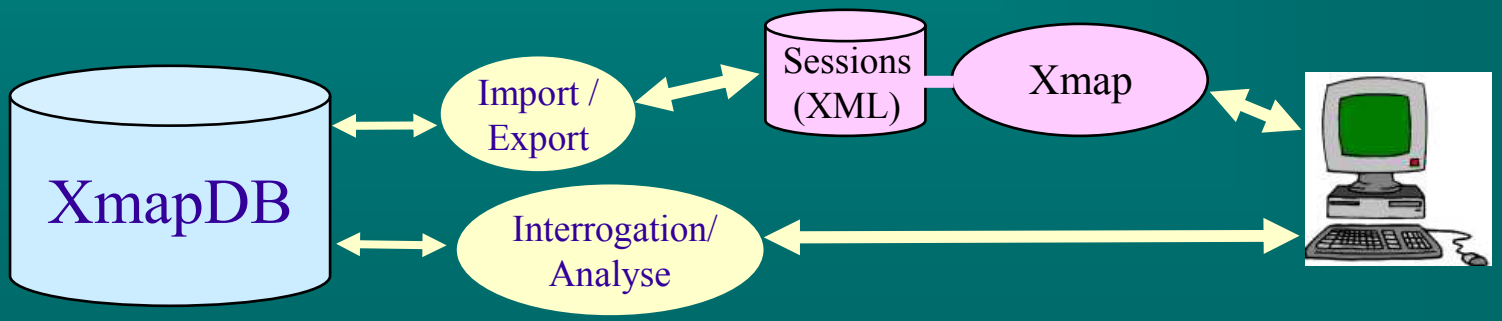
II.3 Exploitation des données collectées (1)

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.3 Exploitation des données collectées (2)

Mise en œuvre de Xmap_DB

Modèle de données (DTD)

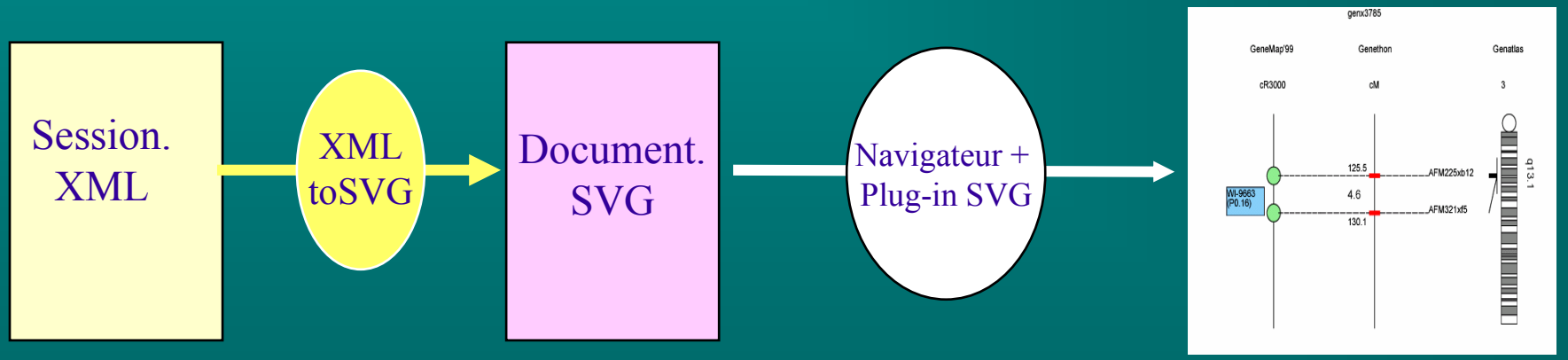
Schéma relationnel

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



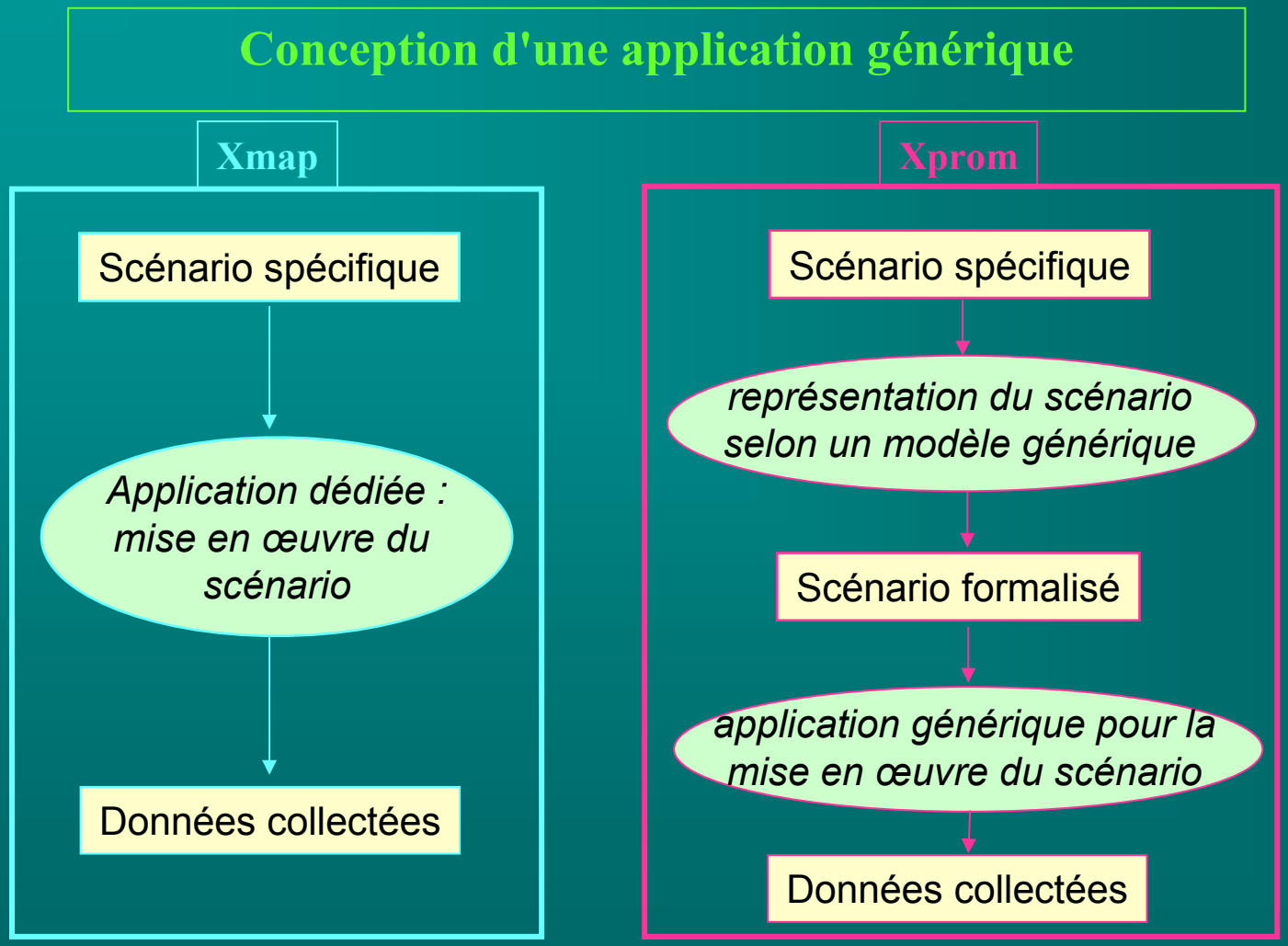
Utilisation de Scalable Vector Graphics : SVG (spécification du W3C)

I. Analyse du problème de la collecte de données biologiques

II. Une solution dédiée : le projet Xmap

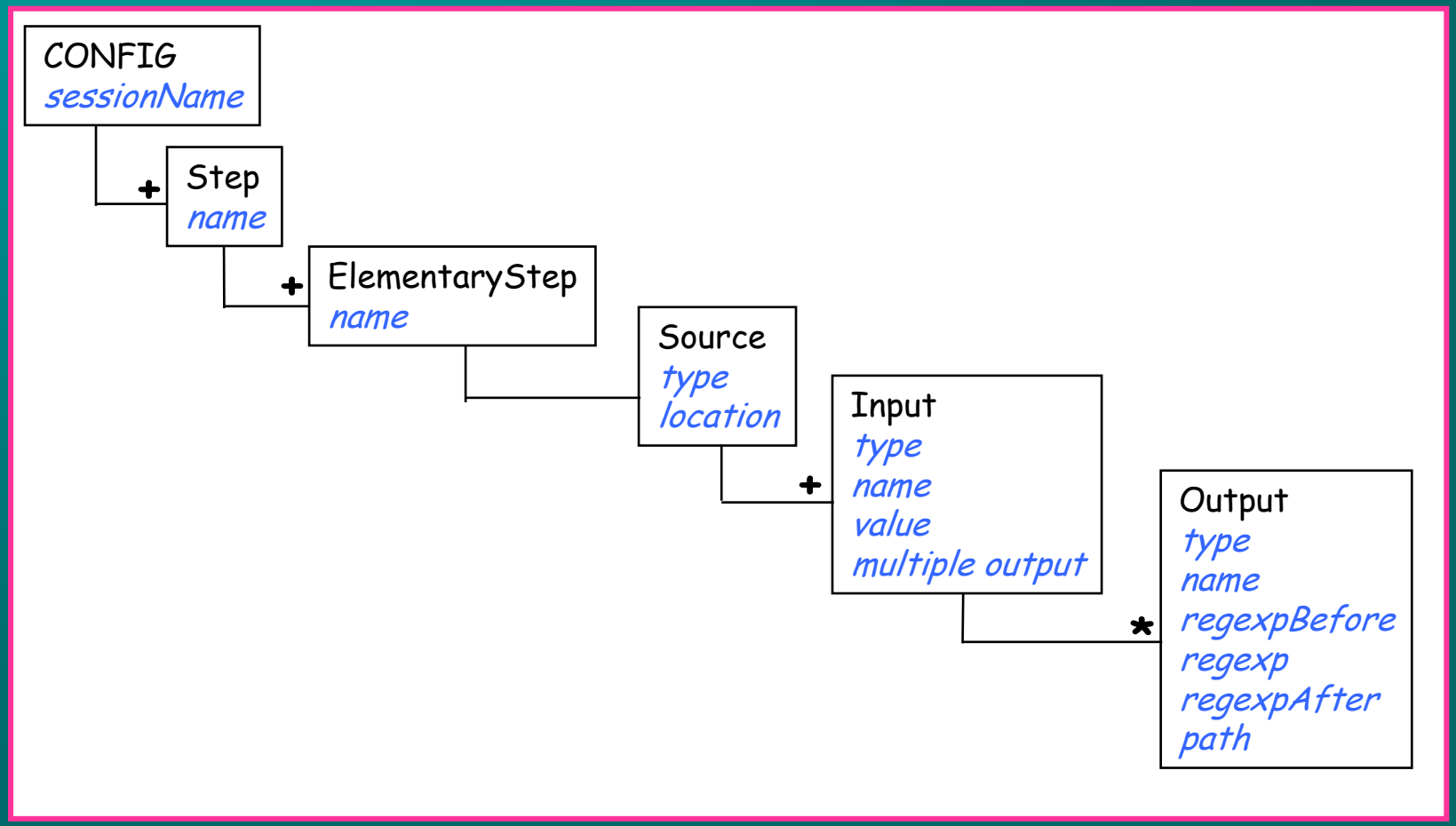
III. Une solution générique: le projet Xprom

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

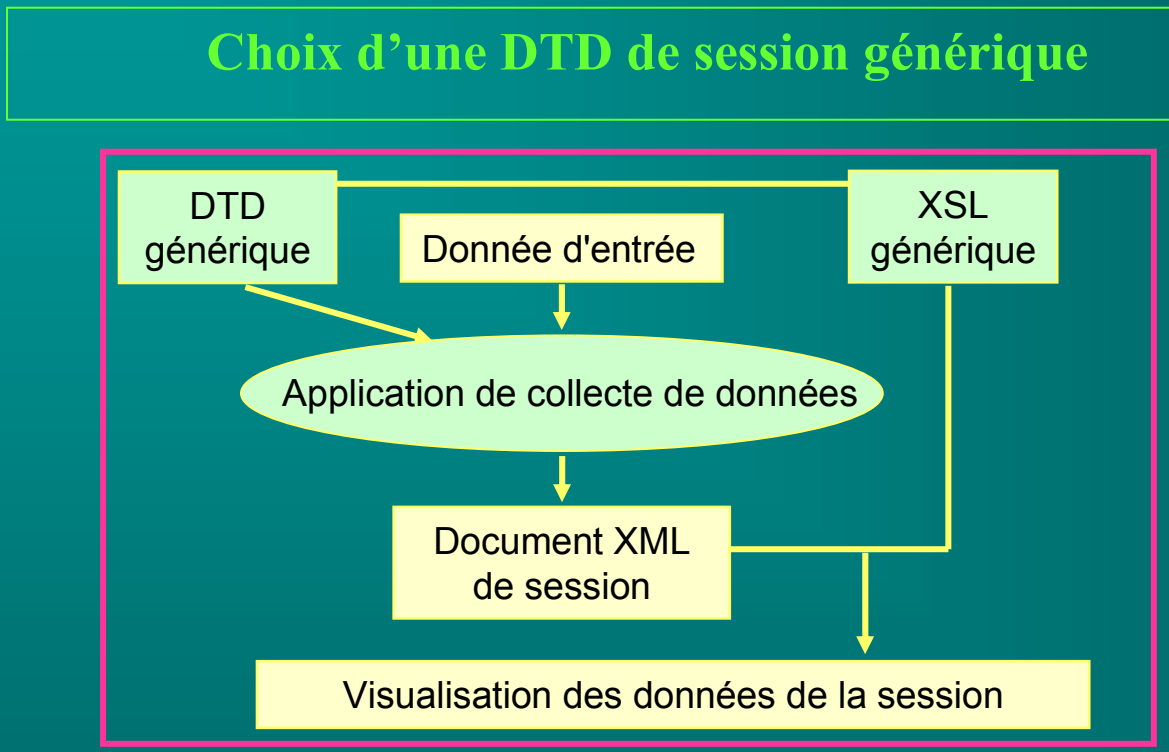


III.1. Modélisation du scénario de collecte (2)

DTD de scénario générique



III.2. Représentation générique des données (1)

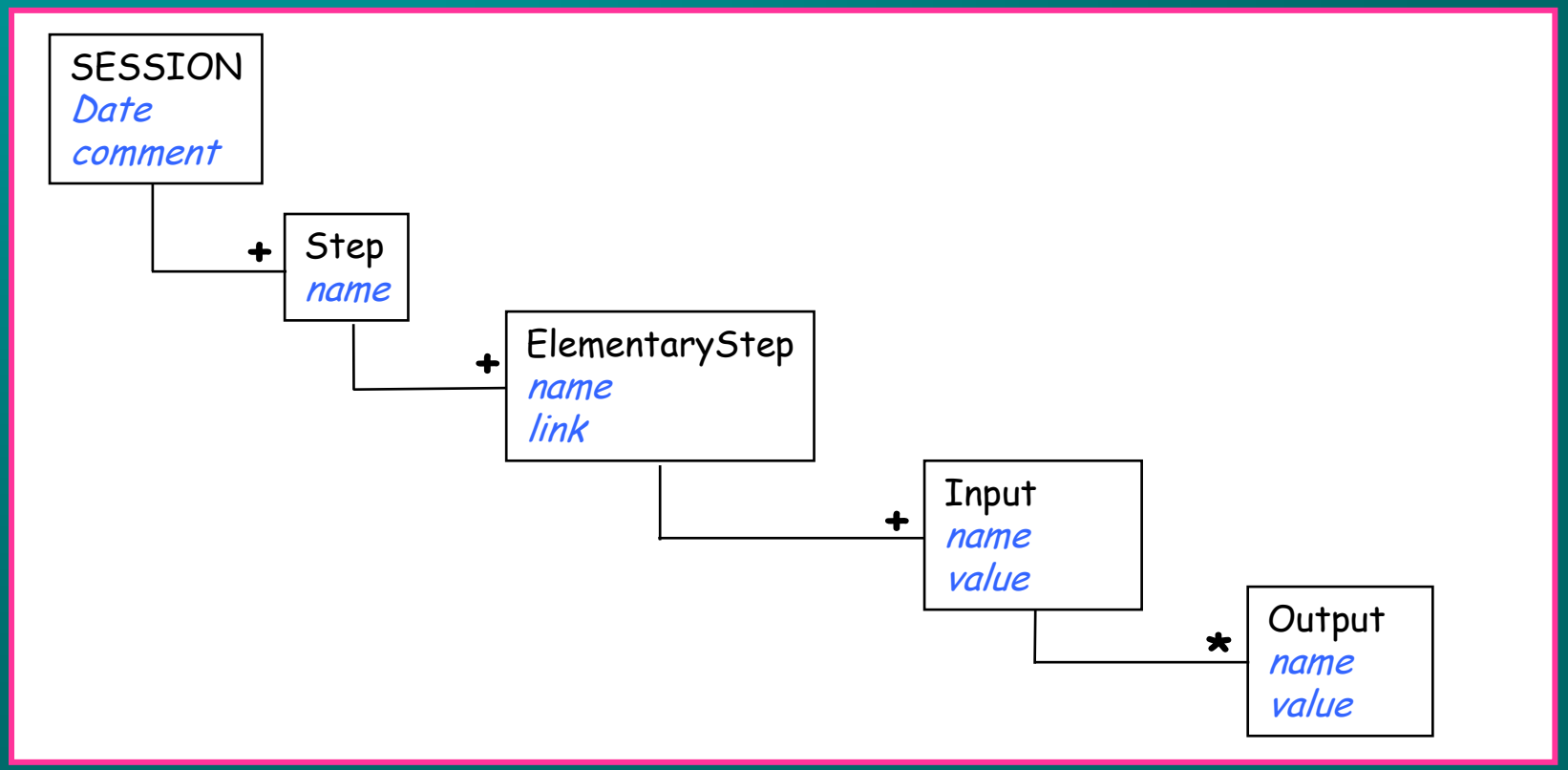


Structure du document de session orientée par le scénario

Possibilité de le convertir en une autre structure plus signifiante

III.2. Représentation générique des données (2)

DTD de session générique



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

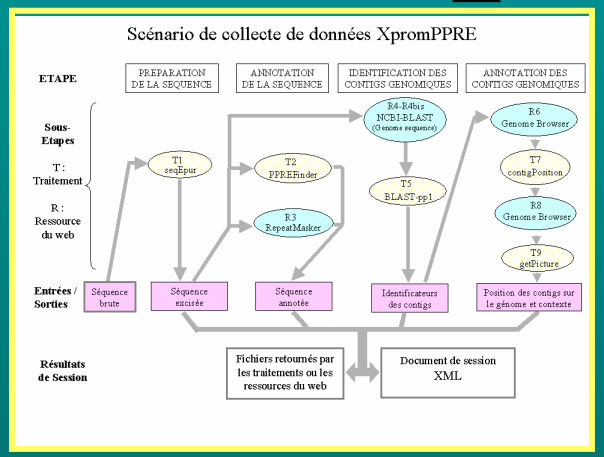
- Liste des étapes et sous-étapes
- Pour chaque sous-étape :
  - Formulation de la requête
  - Soumission
  - Récupération du document retourné
  - Extraction des données à collecter
  - Sauvegarde dans document XML de session
  - Enchaînement sous-étape suivante

III.3. Mise en œuvre (2)

Application : XpromPPRE

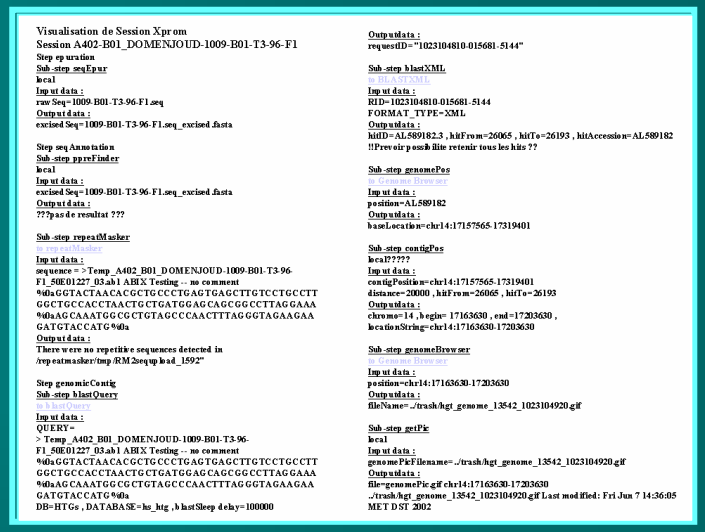
recherche du contexte d'une séquence dans le génome humain

Configuration



Scénario

XpromPPRE_scenario.xml



Seq1009-B1-visu.html

Seq1009-B1-session.xml

Seq1009-B1

Donnée d'entrée

Recherche

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

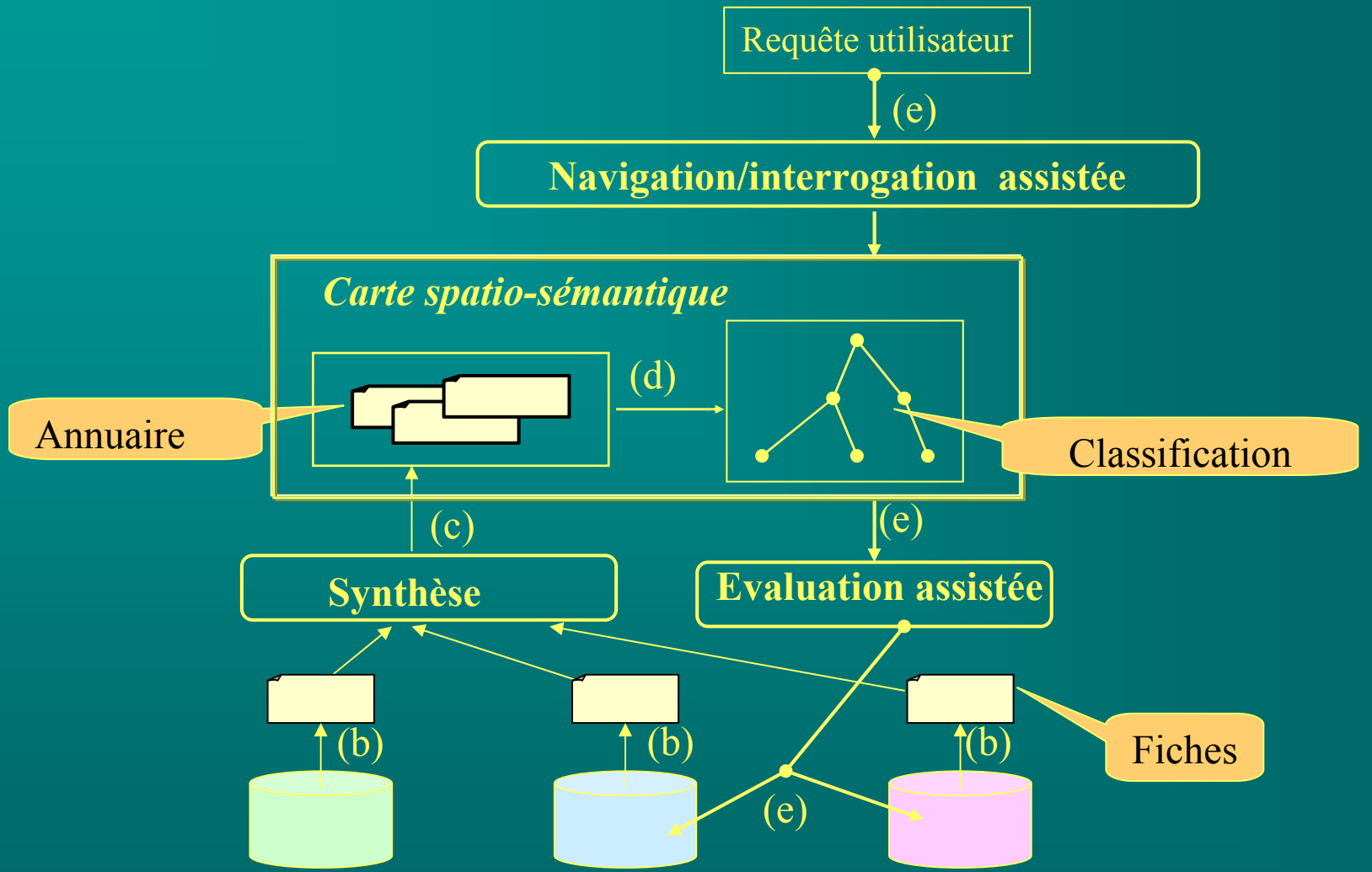
- Dans version automatique, une seule source interrogée
  => choix manuel de la meilleure source

- 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