États des lieux du Web sémantique

Ivan Herman, W3C

19èmes Journées Francophones d'Ingénierie des Connaissances

19 juin, 2008, Nancy, France

Nous avons des technologies de base à notre disposition

- Spécifications stables depuis 2004 : RDF, OWL
- Spécification pour interroger les données depuis 2008 : SPARQL
- Technologies pour créer et accéder aux données en RDF : GRDDL, RDFa, POWDER, …
- Certains vocabulaires de base sont devenus omniprésents (FOAF, Dublin Core,…)

Beaucoup d'outils (*liste incomplète*)

- Des catégories d'outils :
  - banque de triplets, moteurs d'inférence, convertisseurs,
  - moteurs de recherche, middleware, CMS,
  - navigateurs pour le Web sémantique, outils de développement, Wikis sémantique, …
- Quelques noms :
  - Jena, AllegroGraph, Mulgara, Sesame, flickurl,
  - TopBraid Suite, Virtuoso, Falcon, Drupal 7, Redland, Pellet, Disco, Oracle 11g, RacerPro, IODT,
  - Ontobroker, OWLIM, Tallis Platform, RDF Gateway, RDFLib, Open Anzo, DartGrid, Zitgist, Ontotext,
  - Protégé, Thetus publisher, SemanticWorks, SWI-Prolog, RDFStore, …

Beaucoup d'outils (cont.)

- Capacité de stockage, vitesse, etc, évoluent de jour en jour
- Certains des outils sont dans le domaine public, certains non ; certains sont complets, d'autres non ; *c'est la situation des outils en général*, *rien de spécial !*
- *N'importe qui peut développer une application pour le Web sémantique*

Une communauté active

- Beaucoup de tutoriels, guides, livres, conférences, articles
  - comme pour les outils, il y en a qui sont bien, d'autres non, tout comme pour d'autres domaines…
- Une communauté active de développeurs
  - blogs, IRC, listes de diffusion, wikis : plus que ce qu'une personne peut absorber seul…
- Certains avancent le chiffre de 10^7 pour le nombre de documents sur le Web sémantique

Quelques exemples de communautés

- Communautés importantes qui commencent à utiliser cette technologie : bibliothèques numériques, défense, e-Gouvernement, secteurs d'énergie, de finances, santé, …
- Le Web sémantique apparaît aussi dans le monde du « Web 2.0/Web 3.0 »
  - échange de données sociales (« *social networks* »)
  - applications personnelles
  - multimédia (vidéo, audio, photo)
  - etc.

Mais, c'est quoi le Web sémantique?

- Il y a de plus en plus de types d'applications se référant au Web sémantique :
  - intégration de données en utilisant RDF, SKOS, OWL,…
  - ingénierie des connaissances à base d'ontologies (complexes ou non)
  - une meilleure gestion des données, d'archivage, de catalogues, de bibliothèques numériques,
  - gestion et coordination des services Web
  - agents intelligents
  - amélioration des moteurs de recherche (utilisant des vocabulaires spécifiques d'un domaine…)
  - et, bien sûr, un mélange de tout cela…
- Qu'est-ce qui lie entre eux toutes ces applications ?

Sommes-nous à ce stade?

Sommes-nous à ce stade? (cont.)

- Peut-être, mais le fait d'être un éléphant n'est pas nécessairement un problème ! Ça montre que :
  - le Web sémantique a atteint une maturité
  - il y a beaucoup d'intérêt, de développement, d'expérimentation
  - les divers domaines d'application choisissent ce dont ils ont besoin…
    - certains ont besoin d'une gestion sophistiquée de connaissances, ils utilisent donc des ontologies complexes …
    - d'autres se concentrent sur des vocabulaires sémantiquements plus simple mais avec un volume important de données
- Et tant mieux, il y a de la place pour tout le monde !

C’est quoi le Web sémantique?

- Il est, néanmoins, bon d’insister sur certains principes…
- Le Web sémantique est :
  - un moyen standard de spécifier des données et leurs *relations*
  - un moyen d’étendre les principes du Web des documents aux données ; créer un *Web des données*
  - découvrir de nouvelles relations par le biais du Web (tout comme pour les documents)

C’est quoi le Web sémantique? (cont.)

- C’est le *Web* sémantique, et pas seulement de la sémantique
  - les données, les ontologies, les vocabulaires, etc, doivent être partagés, réutilisés, et à l’échelle du Web
  - il est possible d’utiliser l’infrastructure du Web pour désigner aussi des « choses » plus abstraites…
    - p.e. : **`http://www.ivan-herman/me`** *me* désigne (pas ma page d’accueil, pas mon fichier FOAF, mais moi!)
  - … et ajouter des relations qui peuvent inclure ces « choses » !

Quelque mots sur nos technologies récentes

Interroger RDF : SPARQL

- Pouvoir interroger un graphe RDF est indispensable
  - pourrions-nous imaginer des bases de données relationnelles sans SQL?
- SPARQL est :
  - un langage d'interrogation à base de motifs de graphes
  - un protocole pour utiliser SPARQL avec, par exemple, HTTP
  - un format XML pour les résultats de l'interrogation

SPARQL (cont.)

- De nombreuses implémentations sont déjà disponibles
- Il existe également des points d'accès (« *endpoints* ») SPARQL sur le Web :
  - envoyer une requête et une référence aux données par HTTP GET, recevoir le résultat en XML ou JSON
    - certains d'entre eux peuvent être installés facilement sur n'importe quelle machine
  - les bases de données offrent souvent ces points d'accès à leurs données locales
  - *les applications n'ont pas nécessairement besoin de programmation directe de RDF, il suffit d'utiliser un processeur SPARQL*

Un mot d'avertissement sur SPARQL…

- Certaines fonctionnalités manquent :
  - pas de contrôle et/ou de description sur les systèmes d'inférence sur le graphe (RDFS? OWL-DL? OWL-Lite?…)
  - pas de modification du contenu (« *update* »)
  - c'est compliqué d'interroger des collections ou des conteneurs
  - pas de fonctions arithmétiques pour des sommes, moyennes, min, max,…
  - pas de contrôle sur l'agrégation des interrogations
- Ce sera pour la prochaine version…

Il y a de plus en plus de données publiques

- IgentaConnect : metadonnées bibliographiques, plus de 200 millions de triplets
- Représentation de WordNet en RDFS/OWL : fichier accessible en RDF/XML (150MB)
- « Département/canton/commune » : publication de données géographique de l'INSEE
- Geonames Ontology and Data : 6 millions (et plus) de données géographiques
- RDF Book Mashup : information sur des livres d'Amazon, par exemple

Des ponts vers les bases de données

- Une énorme quantité de données est stockée dans des bases de données
- Des « ponts » sont définis :
  - des couches entre RDF et les données relationnelles
  - les tableaux sont convertis, au niveau conceptuel, en RDF :
    - conversion physique complète ; ou bien
    - extraction des relations RDF, mais les données restent ; ou bien
    - génération des interrogation en SQL « on the fly »
    - etc.

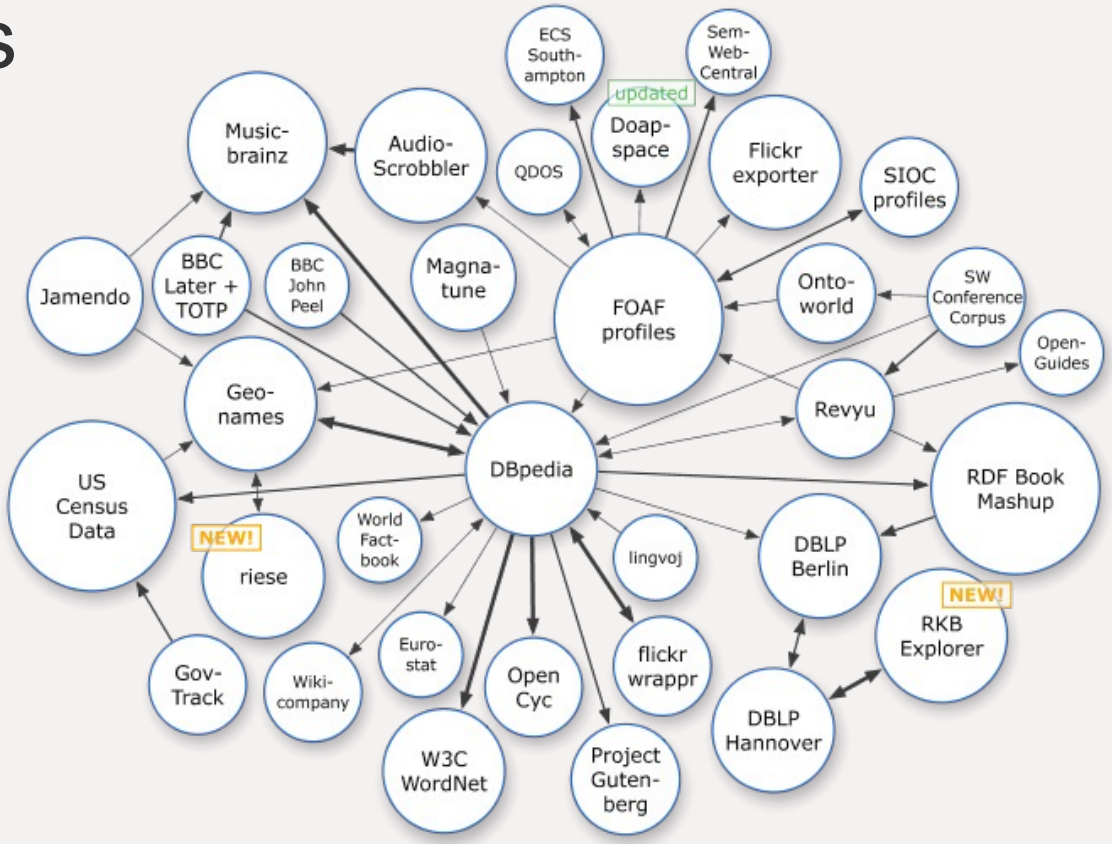
Des ponts vers les bases de données

- Le travail pour une étude sur les techniques des ponts vient de commencer au W3C
- SPARQL devient l'outil de choix pour interroger les données (par l'intermédiaire des points d'accès)

Projet « Linking Open Data »

- Objectif : « exposer » des données publiques
- Ajouter des liens RDF entre les données venant de bases de données différentes
- Mettre en place des points d'accès SPARQL pour interroger les données

  - des milliards de triplets
  - des millions de « liens »



Exemple : DBpedia

- DBpedia est un effort communautaire pour :
  - extraire des informations structurées (*infobox*) de Wikipedia
  - fournir des points d'accès en SPARQL
  - créer des liens en RDF vers d'autres données sur le Web

Informations structurées de Wikipedia

```
http://en.wikipedia.org/wiki/Nancy
```

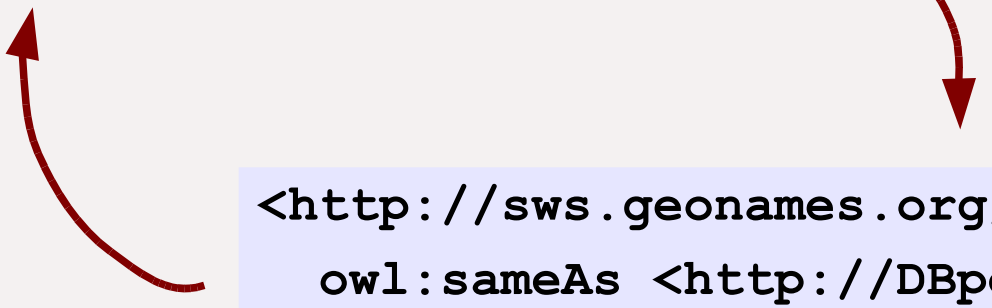
```
<http://dbpedia.org/resource/Nancy>
  rdfs:label "Nancy"@fr, "Нанси"@ru ;
  dbpedia:nomcommune "Nancy"@fr ;
  dbpedia:altMaxi  "353"^^dbunits:Meter ;
  dbpedia:altMini  "188"^^dbunits:Meter ;
  dpbedia:maire
       dbpedia:André_Rossinot ;
  foaf:homepage
       <http://www.mairie-nancy.fr> ;
 ...
```

Coordinates	48°41'25"N, 6°11'00"E
Administration	
Country	France
Region	Lorraine
Department	Meurthe-et-Moselle (*préfecture*)
Arrondissement	Nancy
Canton	Chief town of 4 cantons
Intercommunality	Communauté urbaine du Grand Nancy
Mayor	André Rossinot (2001-2008)
Statistics	
Elevation	188 m–353 m (avg. 212 m)
Land area[1]	15.01 km²
Population[2] (2005)	105,400
- Density	6,902/km² (1999)
Miscellaneous	
INSEE/Postal code	54395/ 54000

Liens automatiques entres données

```
<http://dbpedia.org/resource/Nancy>
  owl:sameAs <http://sws.geonames.org/2990999>;
 ...
```

DBpedia



Geonames

```
<http://sws.geonames.org/2990999>
  owl:sameAs <http://DBpedia.org/resource/Nancy>
  wgs84_pos:lat "48.69333267211914"^^xsd:float ;
  wgs84_pos:long "6.184444427490234"^^xsd:float ;
 ...
```

Les processeurs peuvent passer d'une description à l'autre automatiquement…

Projet « Linking Open Data » (cont.)

- Un projet communautaire majeur
  - n'importe qui peut y participer, abonnez-vous à la liste :
    - http://lists.w3.org/Archives/public/public-lod/
  - ou consultez le site :
    - http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData
  - si vous connaissez une autre base de données publique, contacter le projet !
- Des applications utilisant cet ensemble de données commencent à apparaître

Extraire des (meta)données

- Une approche en développement
  - analyser le texte avec des techniques de traitement automatique de la langue (TAL)
  - créer des (meta)données en RDF
  - certains systèmes le font d'une manière invisible (p.e. Twine)
  - d'autres offrent des services publics qui peuvent être ajoutés à d'autres systèmes (par exemple, Open Calais de Reuters)

Extraire les données structurés

- Des outils, des services, etc, apparaissent :
  - obtenir les données associées à des images, par exemple :
    - service pour obtenir RDF des images sur flickr
    - service pour obtenir RDF d'XMP
  - des scripts pour convertir des feuilles de calcul en RDF
  - etc
- Beaucoup de ces outils sont encore des « hacks » individuels, mais ils montrent une tendance générale
- D'autres outils vont apparaître
  - il y a une page wiki qui contient des références à ceux qui existent déjà

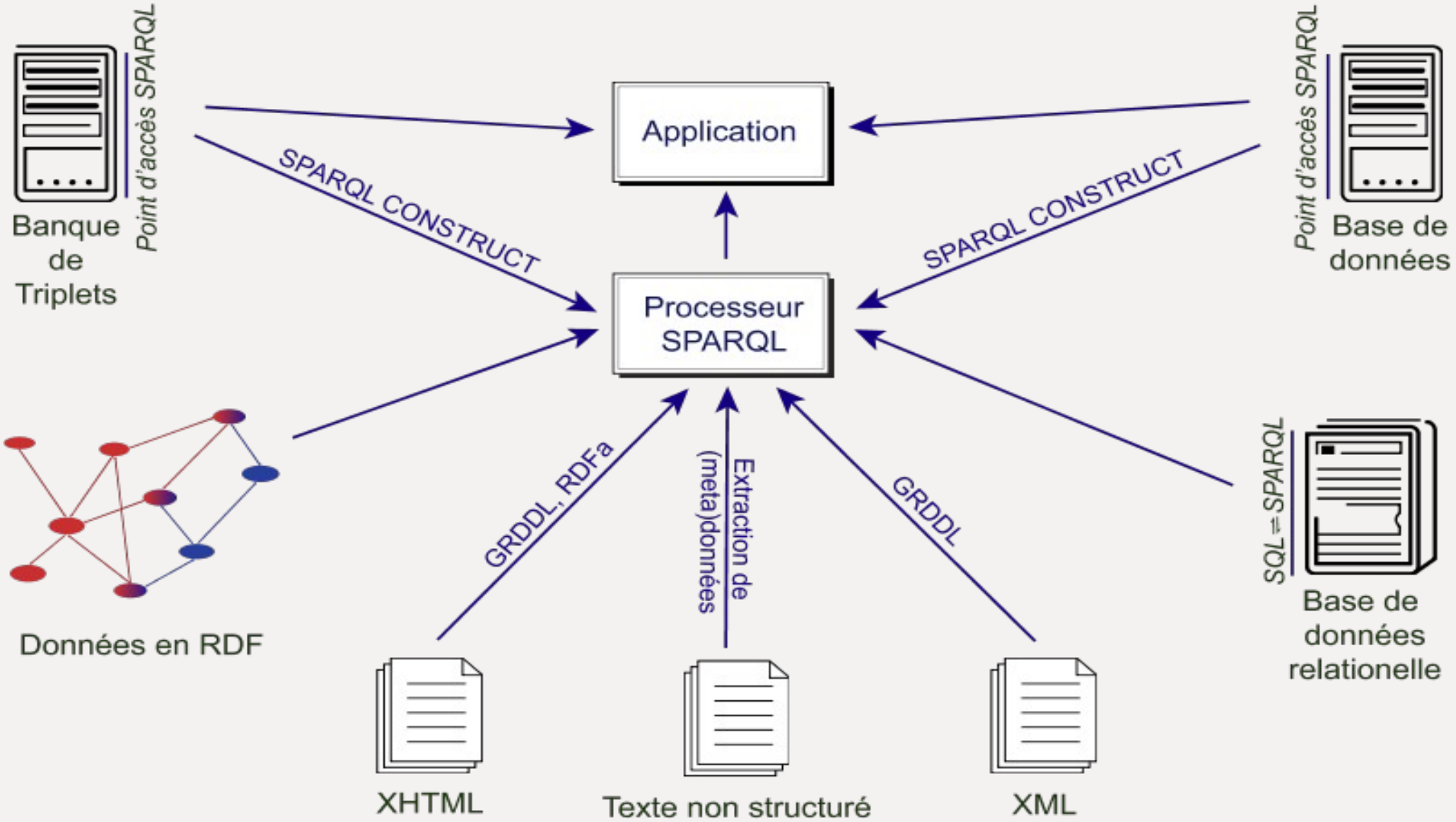
Extraire les données structurés en RDF : GRDDL

- GRDDL est un moyen d'accéder à des données en XML/XHTML et de les transformer en RDF :
  - GRDDL définit des attributs pour accéder à un script qui transforme les données en RDF
- Une façon de créer de nouveaux ponts vers RDF
  - un lien possible aux microformats
  - se rattacher aux applications XML de toutes sortes

Ajouter des structures RDF à XHTML : RDFa

- RDFa définit un ensemble d'attributs pour ajouter des données structurées à (X)HTML
  - des processeurs spécialisés produisent du RDF à partir de XHTML+RDFa
- Il devient facile d'utiliser n'importe quel vocabulaire RDF en XHTML
  - RDFa utilise des espaces de noms pour départager les terminologies et pour pouvoir les mélanger

SPARQL comme point d'unification !



Naturellement, il y a encore à faire…

- Problèmes techniques non résolus
- Évolution des technologies actuelles (OWL, SPARQL, même RDF)
- Problèmes de communication, idées fausses sur les technologies
- Nécessité pour plus d'exemples d'applications, de déploiement

Quelques problèmes techniques

- Sécurité, confiance, provenance
  - combiner des techniques de chiffrement avec le modèle RDF, signer une partie d'un graphe, etc
  - modèles de confiance
  - protocoles de provenance
- Fusion et alignement d'ontologies ; développement distribué de vocabulaires, traitement de versions…
- Incertitude : relations probabilistes ou floues, raisonnement appropriés, …
- Identification par le biais des URIs (nature exacte de leur utilisation, catalogues généraux,…)
- etc

Travaux du W3C

- Travaux en cours :
  - intégration des règles et du Web sémantique
  - nouvelle version d'OWL
  - systèmes pour décrire des taxonomies (SKOS)
  - utilisation du Web sémantique par certaines communautés (eGouvernement, santé, …)
  - …
- Futurs travaux éventuels :
  - nouvelle version de SPARQL
  - sécurité, provenance, données privées
  - nouvelle version de RDF (peut-être…)
  - utilisation du Web sémantique pour le multimédia sur le Web
  - …

Règles : ce que nous aimerions exprimer

- Un exemple :
  - “si deux personnes ont le même nom et la même adresse e-mail, ou le même nom et la même page d’accueil, alors ils sont identiques”
- C’est à dire (dans une syntaxe ad-hoc):

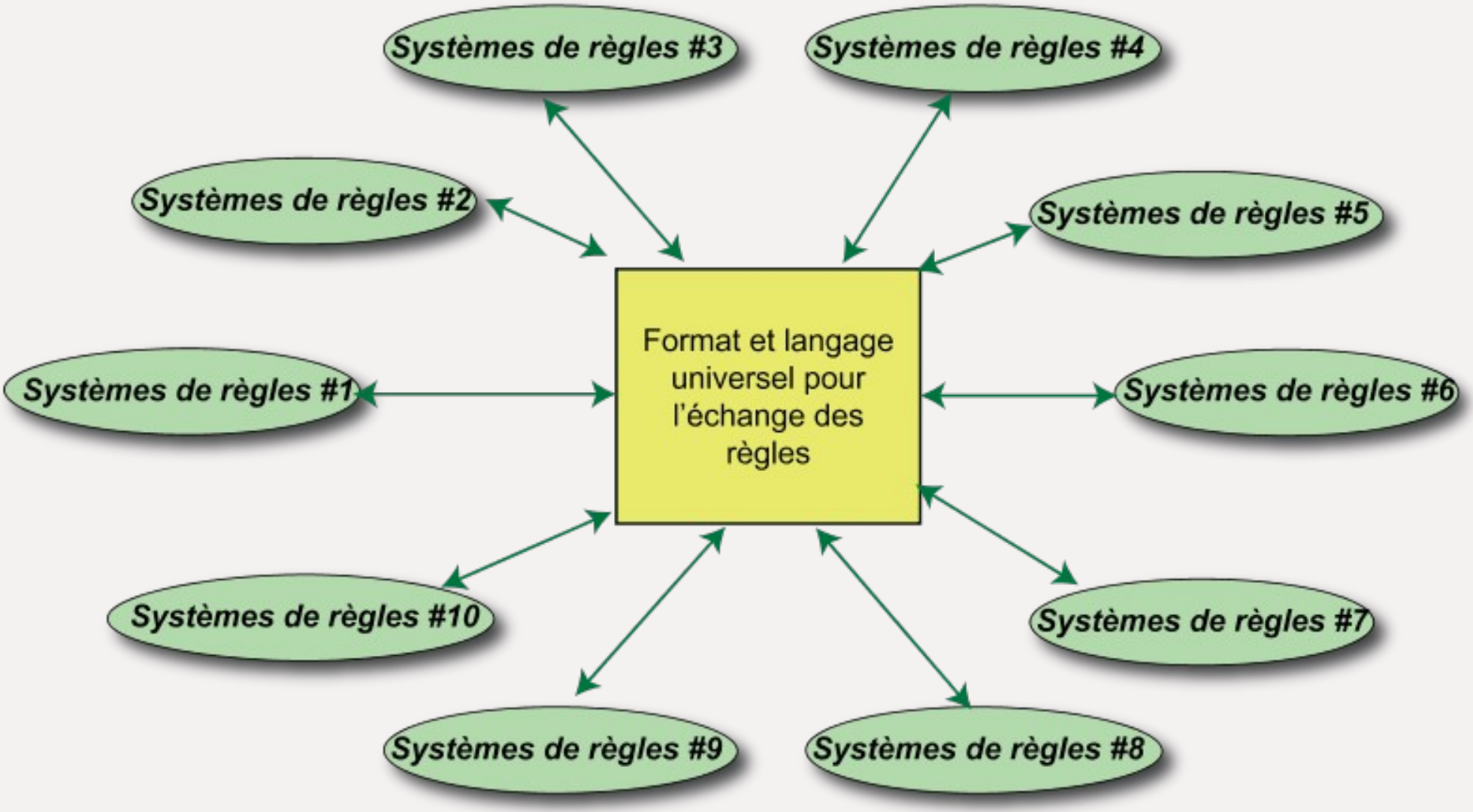
```
If { ?x rdf:type foaf:Person.
     ?y rdf:type foaf:Person.
     ?x foaf:name ?n.
     ?x foaf:homepage ?h.
     ?y foaf:name ?n.
     ?y foaf:homepage ?h. }
then { ?x owl:sameAs ?y }

If { ?x rdf:type foaf:Person.
     ?y rdf:type foaf:Person.
     ?x foaf:name ?n.
     ?x foaf:mailbox ?h.
     ?y foaf:name ?n.
     ?y foaf:mailbox ?m. }
then { ?x owl:sameAs ?y }
```

Une nouveauté : *échanger* des règles

- Les applications peuvent vouloir *échanger* leurs règles :
  - représentation neutre des règles d'un produit de votre entreprise afin que d'autres puissent vous trouver sur le Web
  - échanger les filtres spam d'un système à l'autre
- D'où le nom du groupe : *Rule Interchange Format*
  - un langage qui :
    - exprime les règles pour être utilisé avec, entre autre, RDF
    - puisse être utiliser comme format d'échange

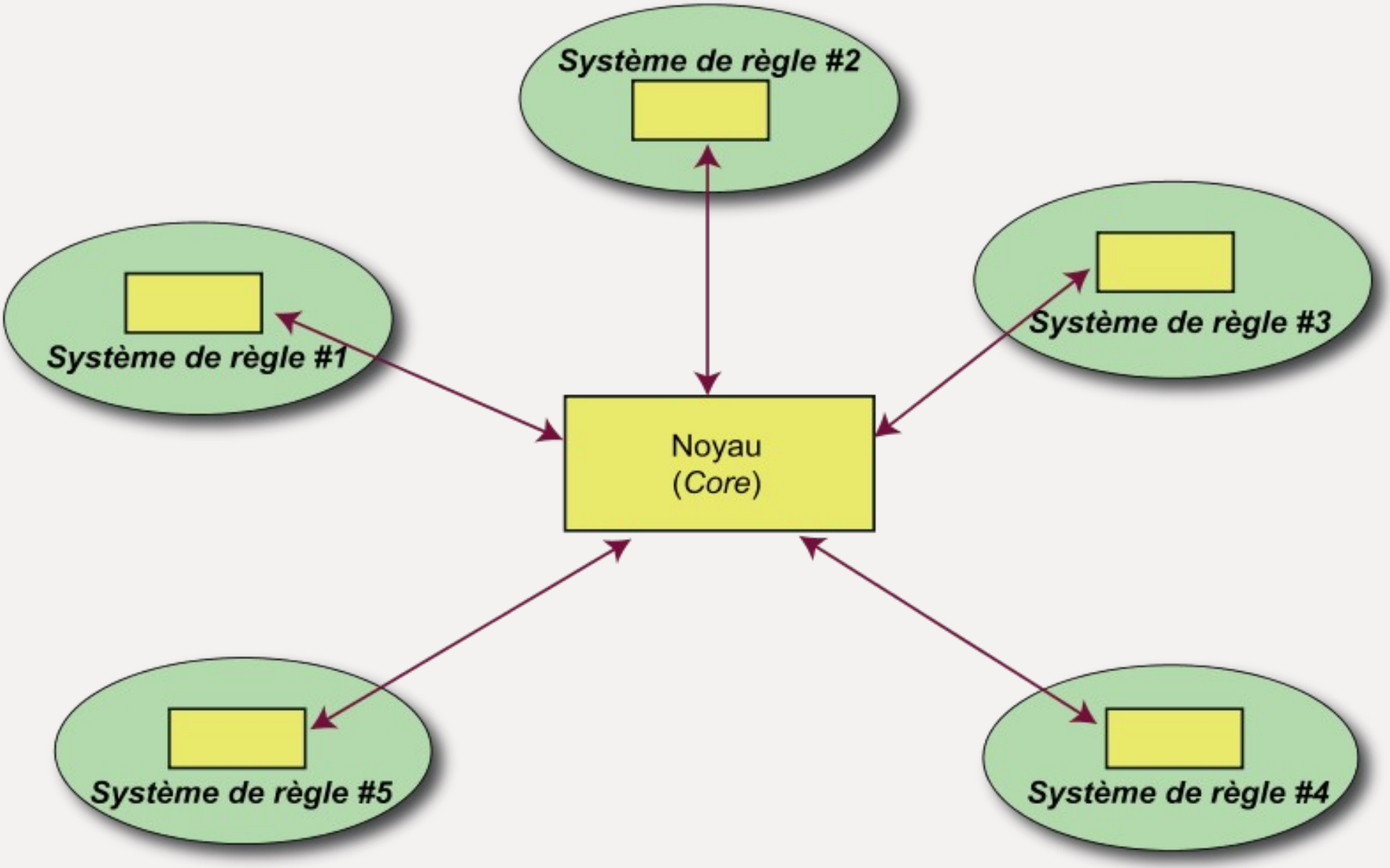
Dans un monde idéal…



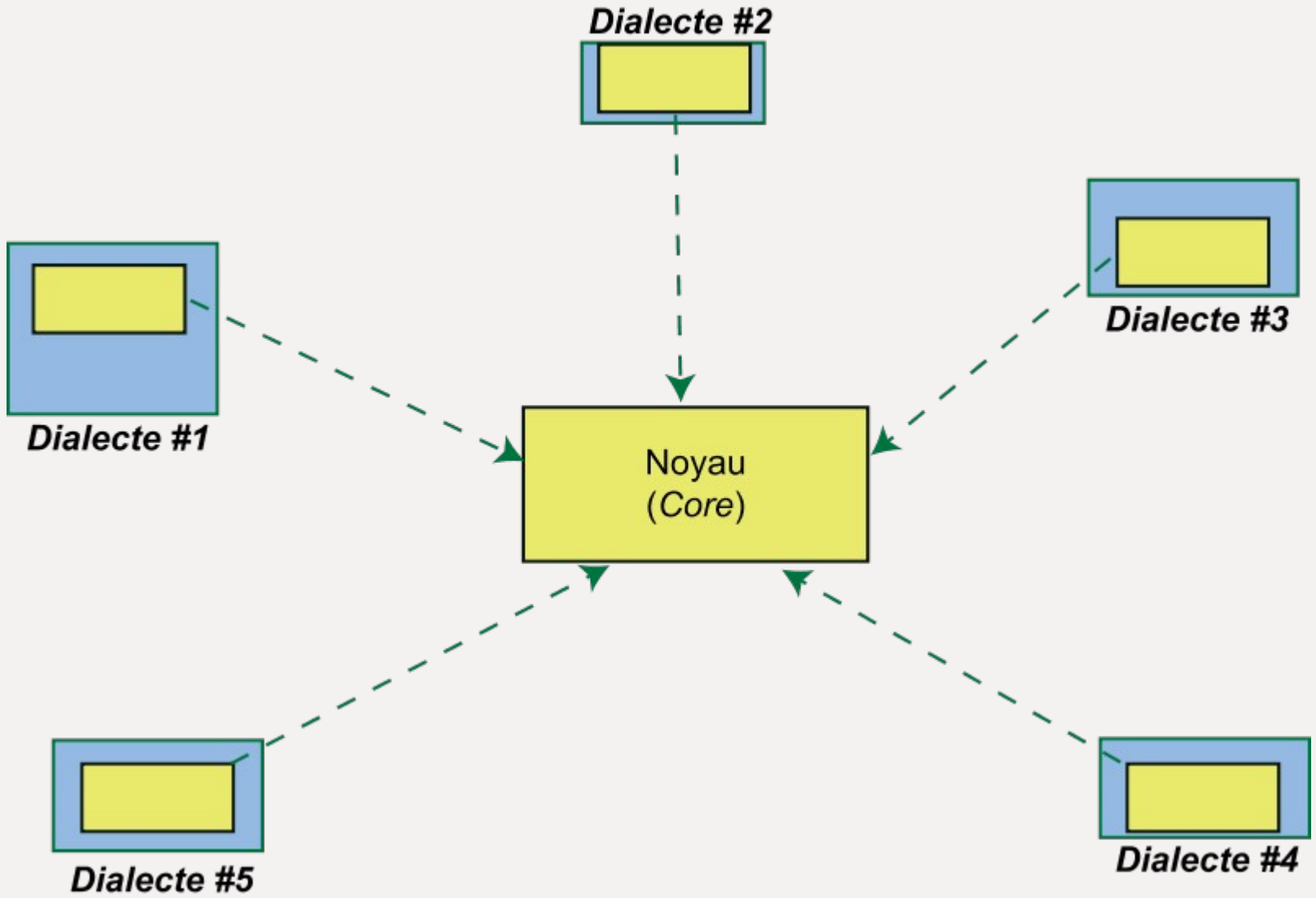
Dans le monde réel…

- Les systèmes de règles peuvent être très différents
  - sémantiques différentes (aux niveaux des modèles, des systèmes de preuves, …)
  - différences avec les règles de production (références à des procédures, transitions d'états, etc)
- Ce genre de langage universel n'est pas faisable

Les noyaux RIF : échange partiel

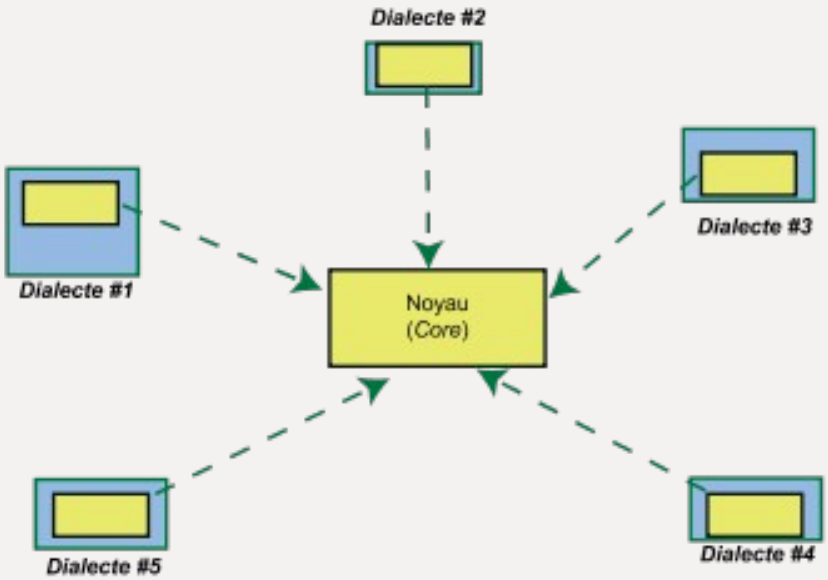


Les « dialectes » RIF

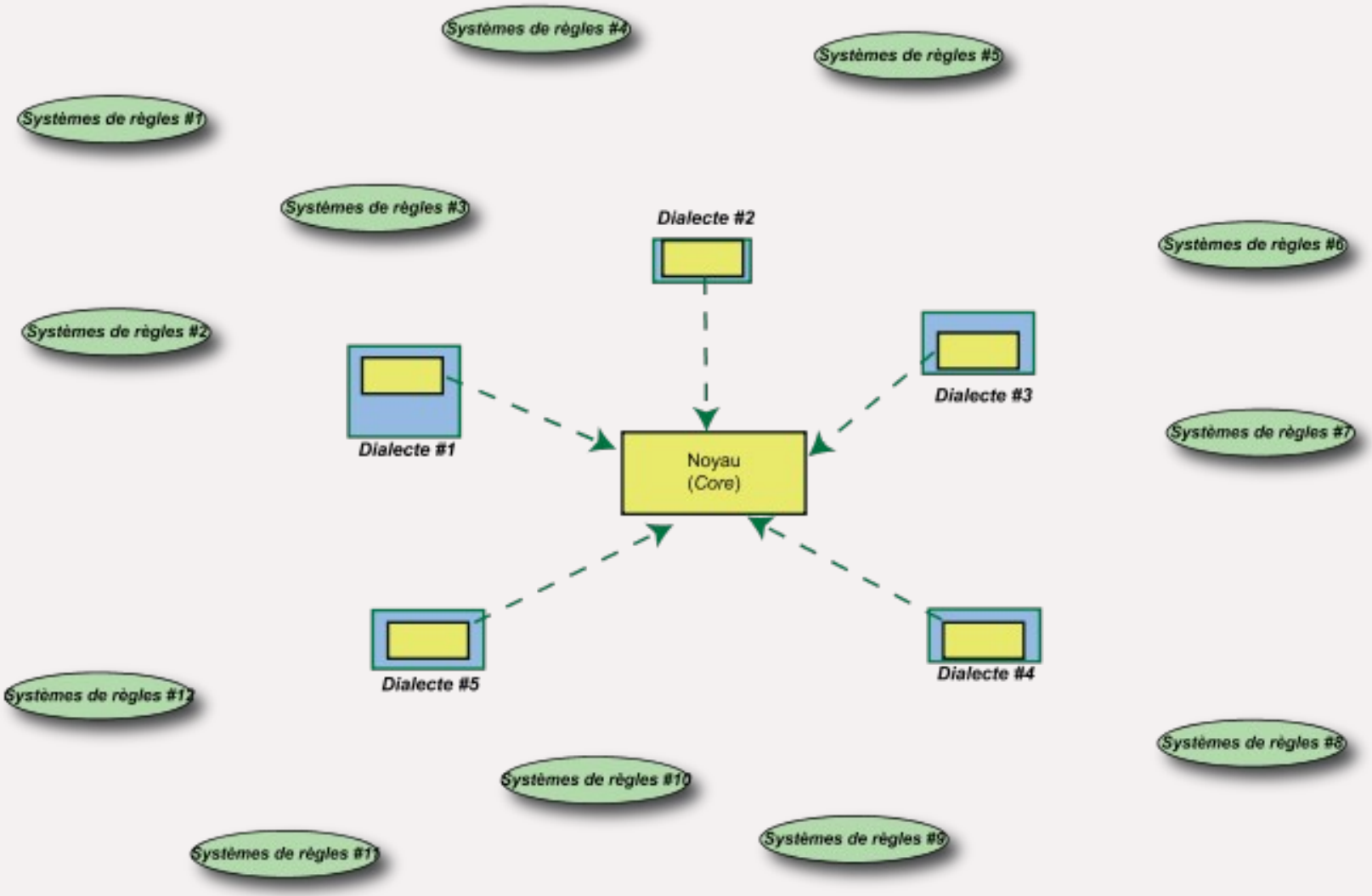


- Dialectes possibles : F-logic, règles de productions, règles floues …

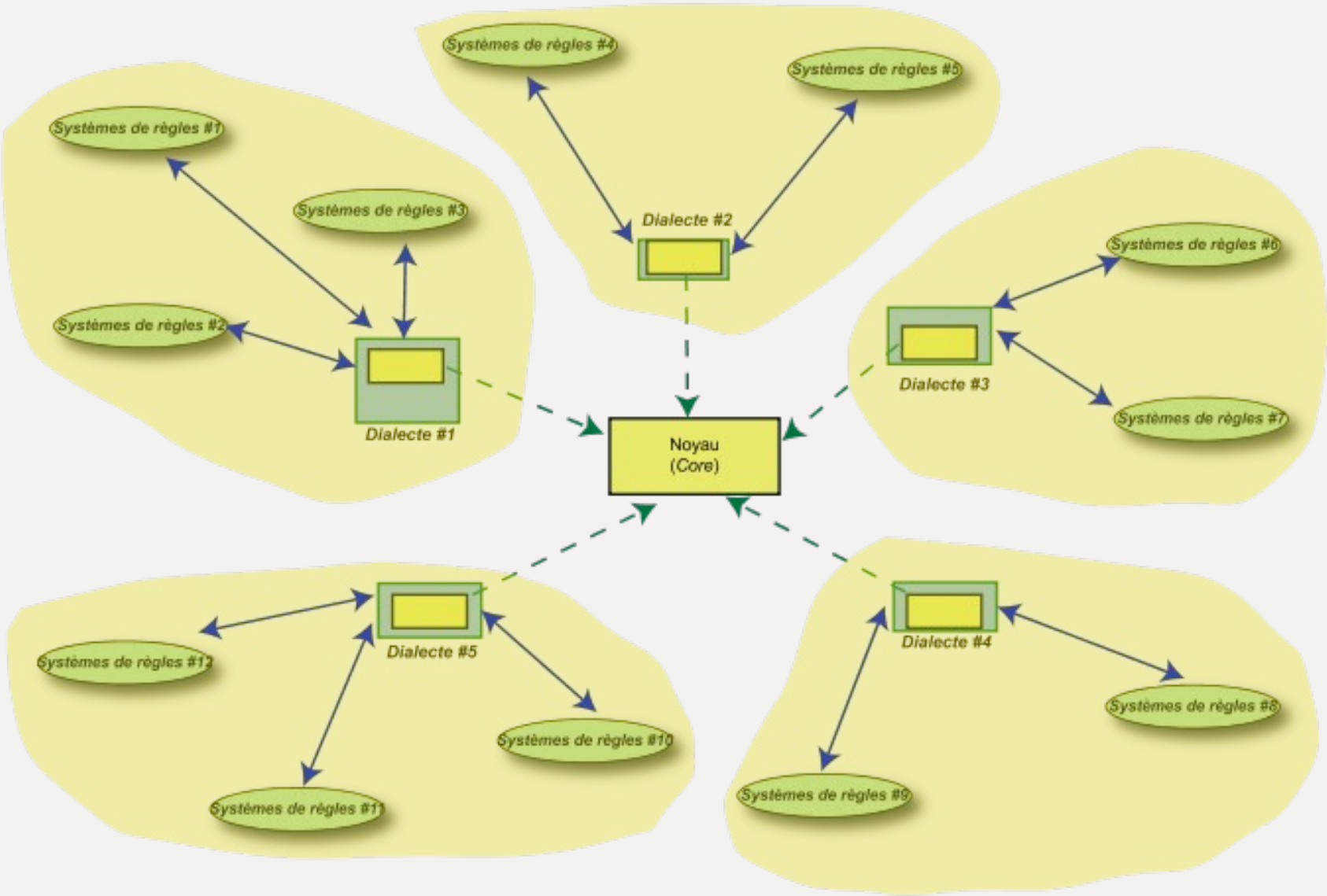
Rôle des dialectes



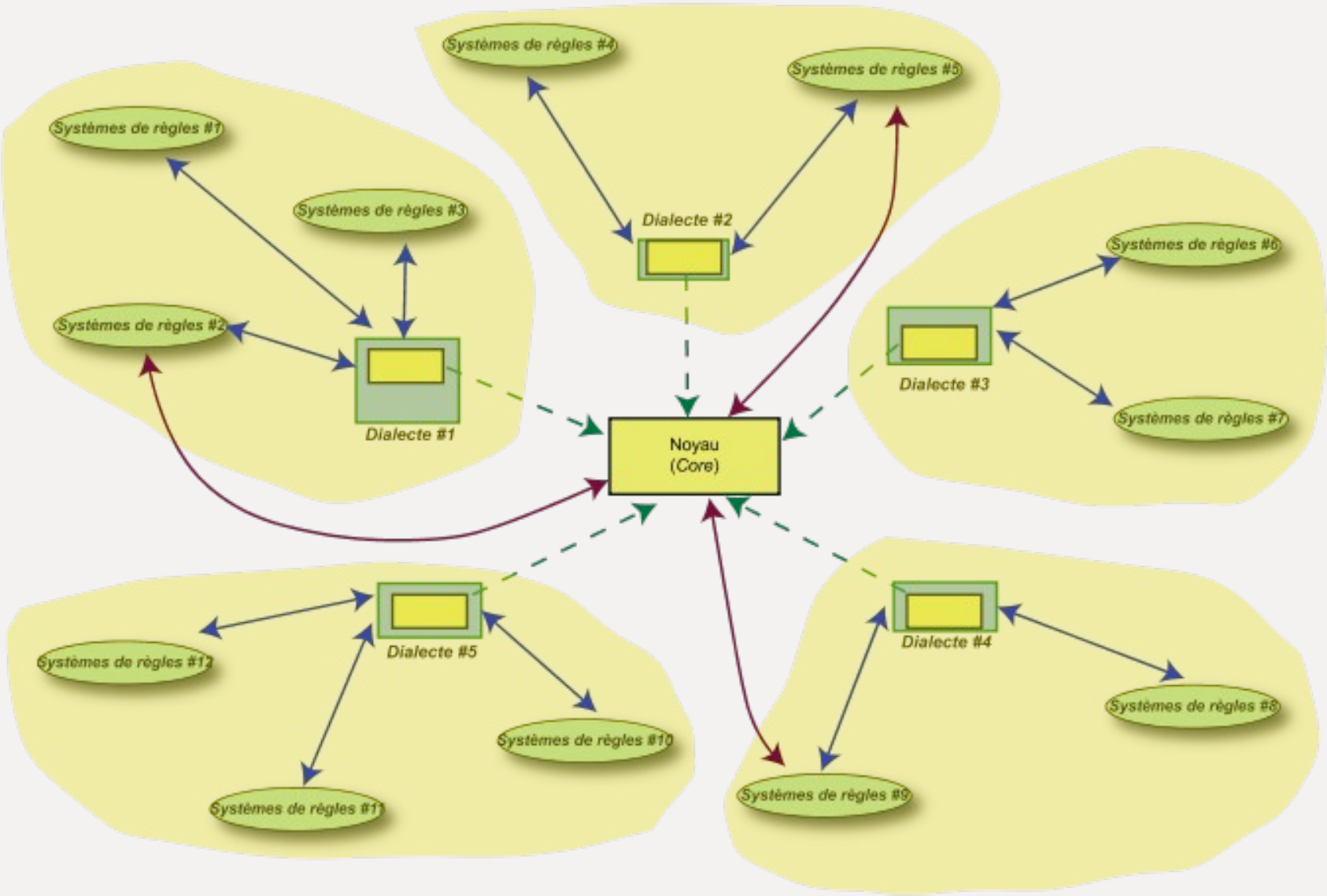
Rôle des dialectes



Rôle des dialectes



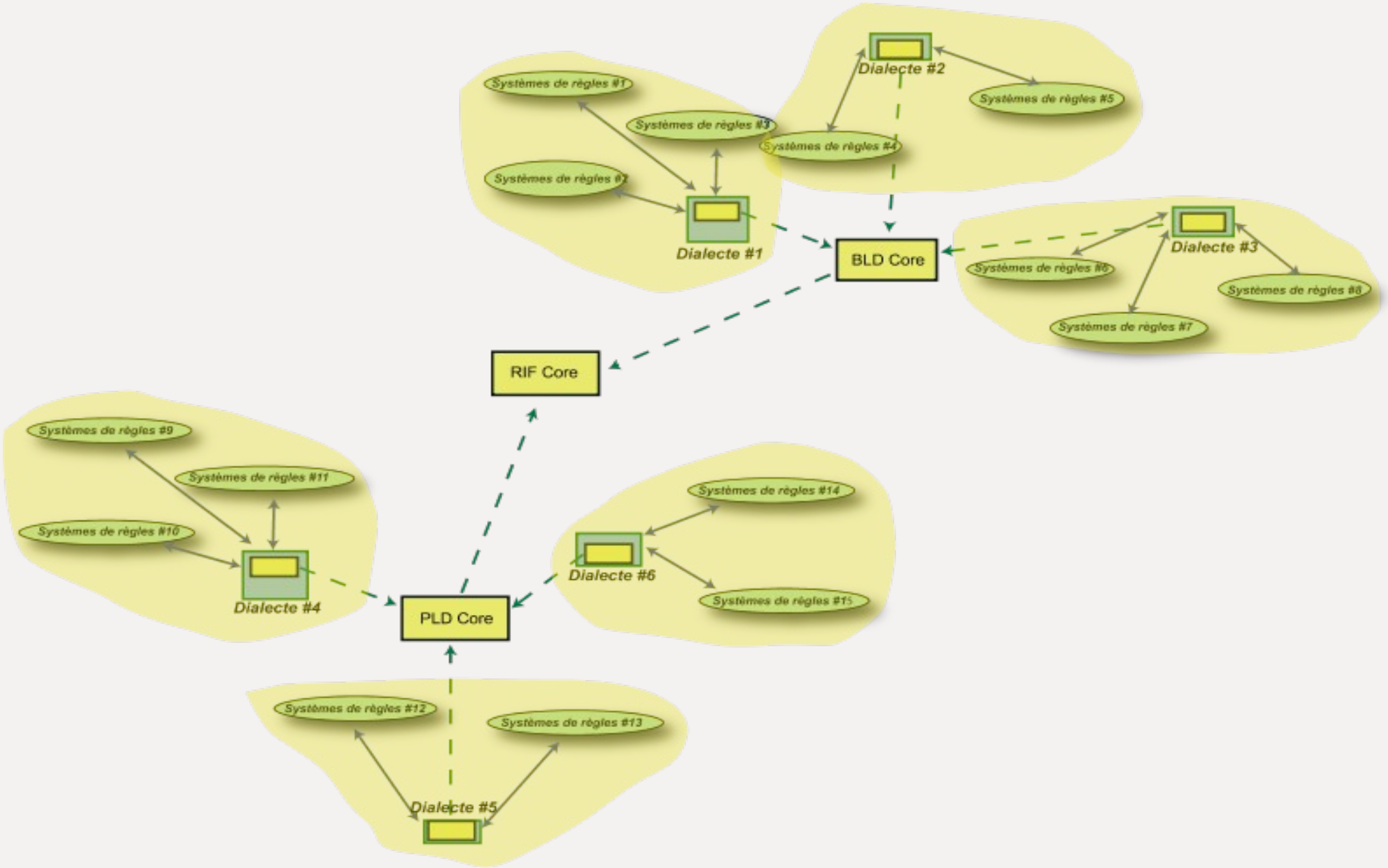
Rôle des dialectes



Néanmoins…

- Même cette approche ne fonctionne pas à 100%
- La différence entre les règles de productions et règles logique « classique » est trop importante
- Il est nécessaire d'avoir une hiérarchie de noyaux :
  - *« Basic Logic Dialect »* et « *Production Rule Dialect »* comme noyaux pour des *familles* de langages
  - un *« RIF Core »* commun pour les lier entre eux

Hiérarchie de noyaux



Schématiquement…

- Le « Core » : sous-ensemble partagée des langages principaux
  - « positive Horn » sans fonctions, avec quelque types de données simples
- Le « BLD (Basic Logic Dialect) » est du genre :
  - « **si condition est vrai alors ceci est vrai** »
  - les conditions peuvent inclure des fonctions, des hiérarchies de classes et de prédicats, égalité
- Le « PLD (Production Logic Dialect) » est du genre :
  - « **si condition est vrai alors fait quelque chose** »

Où en est RIF ?

- Il y a un « draft » pour le BLD
  - une syntaxe XML, et un mécanisme d'extension pour les dialectes
  - le langage peut être utilisé
    - avec ou sans RDF et/ou OWL
    - comme un langage d'échange
- Le plan est d'avoir BLD comme recommandation en 2009
- Les travaux sur le PLD ont également commencé
- Le « Core » viendra comme une abstraction commune de BLD et de PLD

Le nouveau groupe OWL

- Un nouveau groupe de travail à été formé pour une révision d'OWL
- L'objectif du groupe :
  1. ajouter quelques extensions utiles et réalisables
  2. définir de nouveaux « profils » d'OWL

« OWL 2 » : nouvelles fonctionnalités

- « Qualified cardinality restrictions » : par exemple « l'instance d'une classe doit avoir deux chats noirs »
- « Property Chains » : par exemple « si **x** est le frère de **y** et **y** est le père de **z**, alors **x** est l'oncle de **z** »
- « Punning » : le même symbole peut signifier une classes *et* une instance (avec restrictions) même en OWL DL
- Des construction de types de données, au lieu d'utiliser XML Schema

« OWL 2 » : des profils plus simples

- Pour un certain nombre d'applications, RDFS n'est pas suffisant, mais OWL Lite est déjà trop complexe
- Il y a une demande pour une version « light » d'OWL : quelques possibilités supplémentaires ajoutées à RDFS, facilement réalisables
- Plusieurs profils sont considérés par le groupe (EL++, DL-Lite, OWL-R)

Un des profils de OWL 2: OWL-R

- Extension relativement simple de RDFS
- Peut être défini par un ensemble de règles classiques
- Couvre un pourcentage important des applications plus « simples » (comme ceux à base des « Linking Open Data »)

`sameAs`, `unionOf`, `intersectionOf`, `oneOf`, …
`equivalentProperty`, `equivalentClass`,…
`functionalProperty`, `inverseOf`, …
`(max)Cardinality` (1 or 0),…
no existentials in consequents

Applications du Web sémantique

- *Le Web sémantique n'est pas seulement un sujet de recherche universitaire*
- Il y a un nombre croissant d'applications :
  - Sun, Vodafone, Bankinter, Oracle, Radar Networks, Digg, Yahoo!, Microsoft, Unilever, …
- Consultez la liste au W3C : **`http://www.w3.org/2001/sw/sweo/public/UseCases`**

Merci pour votre attention!

- Ces slides sont accessible sur le Web :

http://www.w3.org/2008/Talks/0618-Nancy-IH/