Format d'annotation pour CADERIGE (V1.51)

(Diffusion restreinte au consortium Caderige)

Versions principales du document

V 1.0 : Réunion du 20 Février 2001 au LRI
V 1.1 : Réunion du 10 avril 2001 au LRI
V 1.25 : Réunion 17 octobre 2001 au LRI (+ discussion 20 novembre 2001 au LIMSI)
V 1.4 : Réunion du 4 juin au LRI
V 1.5 : Réunion du 1er juillet à l'INRA Paris

1 Description des balises d'annotations

Voici les balises proposées pour annoter les phrases présentant ou non des interactions entre les entités biologiques pour les organismes *procaryotes* et *eucaryotes*. Dans une première partie (cf. 1.1) on présente le format général d'un *document* annoté, dans une seconde (cf. 1.2), le format d'annotation d'une *phrase* décrivant une interaction. La seconde partie (cf. 2) présente les DTD utilisées. Les exemples sont décrits dans la section suivante (cf. 3).

Pour annoter ce document on utilise le formalisme suivant : Q/R/C (auteur) : texte … ; la première lettre indiquant la nature de l'annotation : Question, Réponse ou simple Commentaire.

1.1 Format général d'un document annoté

<!DOCTYPE Caderige_Annotation " Caderige_Annotation-v1.5.dtd">

Balise permettant de préciser la provenance du document, sa date de modification et ses auteurs

<ANNOTATED-DOCUMENT
id = "nom court permettant d'identifier le document"
reference = "référence permettant de retrouver le document□idéalement un URL"
description = "provenance du document et traitements effectués en amont"
date = "date de dernière modification du document annoté"
author = "nom du (ou des) annotateurs" >

La balise <ABSTRACT> englobe les phrases (voir le bloc <SENTENCE> décrit ci-après) qui appartiennent à un même résumé, ou plus généralement à un même bloc de texte.

<ABSTRACT
id = "identificateur du résumé ou bloc de texte"
reference = "référence permettant de retrouver le résumé □idéalement un URL" >

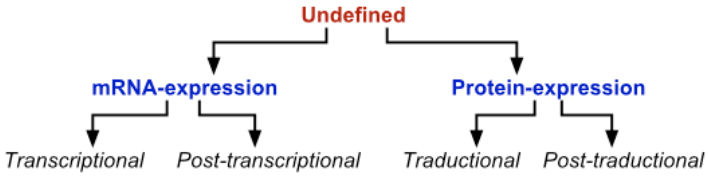
Dans la version actuelle, le document est constitué par une série de phrases délimitées par la balise <SENTENCE>. Pour chacune de ces phrases, on décrit la, (ou les), interactions, ou leur absence, à l'aide d'une des cinq balises suivantes□

- GENIC-INTERACTION,
- NON-GENIC-AGENT-INTERACTION,
- NON-GENIC-TARGET-INTERACTION,
- EXPERIMENT
- NO-INTERACTION.

Il est à noter que dans le cas des balises décrivant une interaction si cette phrase décrit simultanément N interactions, *cette balise doit être répétée N fois dans le bloc <SENTENCE>*, chacun des textes annotés décrivant l'une de ces interactions. Toutefois les trois premières sont exclusives avec les deux dernières EXPERIMENT et NO-INTERACTION.

<SENTENCE
id = "numéro de la phrase dans le résumé issu du préprocessing" >
title = {yes, no}> /* indique si la phrase est un titre

La balise GENIC-INTERACTION délimite une phrase décrivant une interaction dite «□génique□, c'est à dire dans laquelle *au moins l'un des agents et l'une* (Q□□Gilles) : ou bien toutes□ R(Sandrine)□toutes, voir avec Adeline !) *des cibles sont des protéines, des gènes ou des ARNm*. Chaque interaction doit-être identifié à l'aide d'un nom ou d'un numéro (la valeur par défaut peut être la chaîne vide lorsqu'il n'y a qu'une seule interaction décrite). L'attribut TYPE indique le «□*iveau de régulation*□ de l'interaction, c'est à dire à quel moment agit l'interaction. Les valeurs sont organisées au sein d'une taxonomie□



Q (Gilles)□remplacer partout le terme UNDERDETERMINED par UNDEFINED□
R (Sandrine)□voir avec Adeline

Les attributs ASSERTION et REGULATION permettent de préciser la *nature* de l'interaction. Ainsi, REGULATION décrit le sens de l'interaction s'il est connu, activation ou inhibition et ASSERTION indique la modalité, c'est-à-dire si la REGULATION est niée ou pas (voir ci-dessous la table indiquant les différentes combinaisons).

L'attribut UNCERTAINTY indique le *degré de certitude,* avec laquelle l'interaction est connue ou plus exactement avec laquelle elle est relatée dans le texte. L'attribut SELF-CONTAINED indique si la phrase est «□uto-suffisante□ pour comprendre la nature de l'interaction ou si elle fait implicitement référence à d'autres connaissances. Notons à ce propos que l'annotation doit s'effectuer le plus possible *en limitant* le nombre «□'inférences□ qui sont pas directement issues du contenu effectif de la phrase. Enfin, l'attribut CONFIDENCE Q (Gilles)□ je propose d'utiliser ce terme, classique pour les «□eviewer□ plutôt que TEXT-CLARITY□ R (Sandrine)□ voir avec Adeline indique le degré de confiance que l'annotateur accorde à son annotation (qui est liée au degré d'implicite présent dans le texte).

< GENIC-INTERACTION
id = "identificateur de l'interaction"
type = {mRNA-expression, transcriptional, post-transcriptional,
protein-expression, traductional, post-traductional, undefined}
assertion = {exist, non-exist} /* nature de la phrase
regulation = {activate, inhibit} /* nature de l'interaction
uncertainty = {certain, probable, doubtful } > /* degré de certitude
self-contained = {yes, no} /* phrase sans connaissance implicite
confidence = {good, medium, poor}

Texte annoté de la phrase (cf partie 2.2).

</ GENIC-INTERACTION >

La table ci-dessous indique le type d'interaction qui est décrit dans la phrase pour les différentes valeurs prises par les deux attributs. Notons que le second attribut REGULATION peut-être non-valué (chaîne vide) et que dans ce cas la signification dépend de la valeur du premier attribut TYPE.

		REGULATION		
		«activate»	*«inhibit»*	*«»*
ASSERTION	*«exist»*	activation	inhibition	interaction mais de nature indéterminée
	«non-exist»	pas d'activation	pas d'inhibition	pas d'interaction du tout

OU

La balise NON-GENIC-AGENT-INTERACTION délimite une phrase décrivant une interaction dans laquelle *l'ensemble des agents est constitué par des opérateurs «non-géniques»* c'est à dire qui ne sont ni des protéines, ni des gènes, ni des ARNm. L'interaction doit par contre contenir au moins une cible de type «génique».

Les attributs spécifiant cette balise sont les mêmes que ceux précédemment utilisés pour décrire la balise GENIC-INTERACTION.
Q (Gilles) même dans le cas du TYPE ?
R (Sandrine) oui même si la réponse sera souvent Undefined

< NON-GENIC-AGENT-INTERACTION
id = "identificateur de l'interaction"
type = {mRNA-expression, transcriptional, post-transcriptional,
protein-expression, traductional, post-traductional, undefined}
assertion = {exist, non-exist} /* nature de la phrase
regulation = {activate, inhibit} /* nature de l'interaction
uncertainty = {certain, probable, doubtful } > /* degré de certitude
self-contained = {yes, no} /* phrase sans connaissance implicite
confidence = {good, medium, poor}

Texte annoté de la phrase (cf partie 2.2).

</ NON-GENIC-AGENT-INTERACTION >

OU

La balise NON-GENIC-TARGET-INTERACTION délimite une phrase décrivant une interaction dans laquelle *l'ensemble des cibles de l'interaction est constitué par des cibles «non-géniques»* c'est à dire ni des protéines, ni des gènes, ni des ARNm. L'interaction doit par contre mettre en œuvre au moins un agent «génique». Comme précédemment les attributs restent les mêmes.
Q (Gilles) même dans le cas du TYPE ?
R (Sandrine) Non ici c'est inadapté !
Q (Gilles) OK, alors on supprime l'attribut ou on met un autre ensemble de valeurs possibles et dans ce cas lesquelles ?

< NON-GENIC-TARGET-INTERACTION
id = "identificateur de l'interaction"
type = {mRNA-expression, transcriptional, post-transcriptional, protein-expression, traductional, post-traductional, undefined}
assertion = {exist, non-exist} /* nature de la phrase
regulation = {activate, inhibit} /* nature de l'interaction
uncertainty = {certain, probable, doubtful } > /* degré de certitude
self-contained = {yes, no} /* phrase sans connaissance implicite
confidence = {good, medium, poor}

Texte annoté de la phrase (cf partie 2.2).

</ NON-GENIC-TARGET-INTERACTION >

OU (exclusivement)

La balise EXPERIMENT contient une phrase non annotée décrivant une condition expérimentale ne rapportant pas une interaction, mais laissant supposer que l'expérience a pour but de tester une interaction, par exemple l'action d'une hormone sur l'expression d'un gène. Q (Gilles) peut-on dire que EXPERIMENT se rapporte aussi à des résumés dont ni les agents ni les cibles sont géniques ? R (sandrine) plutot non, voir avec Philippe
Contrairement à la balise précédente, cette balise ne doit apparaître qu'une seule fois dans le bloc <SENTENCE>.

<EXPERIMENT>

Texte non annoté de la phrase.

</ EXPERIMENT >

OU (exclusivement)

Cette balise contient une phrase non pertinente du point de vue des interactions

< NO-INTERACTION

Texte non annoté de la phrase.

</NO-INTERACTION>

Pour chacun des résumés, l'annotateur peut ajouter un bloc de commentaires en texte libre.

<COMMENT>

Commentaires optionnels sur la phrase expliquant les annotations éventuelles, ou les doutes. Le contenu de ce texte ne sera pas utiisé par le mécanisme d'apprentissage

</COMMENT>

</SENTENCE>

< /ABSTRACT >

</ANNOTATED-DOCUMENT>

1.2 Format d'annotation d'une interaction

1.2.1 Principes généraux

Chaque phrase annotée ne concerne qu'une seule interaction. Si une phrase décrit simultanément plusieurs interactions, elle doit apparaître autant de fois dans le document. Pour chaque interaction, l'annotation vise à mettre en évidence dans la phrase *les groupes de mots* décrivant :

- les agents (A) : les entités qui sont à la base de l'interaction,
- les cibles (T) : les entités sur lesquelles agit cette interaction,
- l'interaction (I) : le type de contrôle qui est effectué par l'interaction,
- la fiabilité (C) : le degré de confiance que l'on a dans cette interaction décrite.

Afin que le schéma soit conforme au standard XML, les balises qui servent à annoter les différentes parties d'une phrase doivent être imbriquées de manière *hiérarchique stricte*. De plus, d'un point de vue sémantique, on a opté pour l'hypothèse simplificatrice selon laquelle les balises utilisées pour décrire les quatre types d'information (agent, cible, interaction, pertinence) ne peuvent pas être imbriquées. De manière encore plus drastique, cette imbrication n'est pas, en règle générale, autorisée pour un même type de balises. Voici un exemple de ces contraintes :

- Annotation interdite : <AF1> ... <IF> ... </IF> ... </AF1>
- Forme autorisée : <AF1> ... </AF1> <IF> ... </IF> <AF1> ... </AF1>

- Annotation interdite : <AF1> ... <AF2> ... </AF2> ... </AF1>
- Forme autorisée : <AF1> ... </AF1> <AF2> ... </AF2> <AF1> ... </AF1>

Ainsi, on effectue une partition stricte de la phrase, chaque partie étant étiquetée par un seul type de balise, voire une seule balise ; par contre, comme on le voit dans le dernier exemple, une même balise peut peut apparaître plusieurs fois si elle permet d'étiquetter des zones non contigües dans la phrase. Dans ce cas c'est la réunion des différents groupes de mots balisés qui constituent

l'information. L'avantage de cette vision est que l'on peut introduire des contraintes dans la DTD et ainsi bien mieux assurer le contrôle et la validation de l'annotation. Du point de vue de l'édition, l'annotation "visuelle" des texte (via l'utilisation de couleurs, polices ...) peut s'effectuer de manière plus simple puisqu'il n'y aurait plus de "chevauchements" entre des zones de type différent.

1.2.2 Description des Agents

Globalement, il y a deux familles de balises permettant de décrire les agents intervenant dans l'interaction. La première : <AFn> (Agent Fragment n) sert à désigner dans la phrase l'ensemble des fragments de textes qui décrivent et qualifient un agent donné (ce qu'on appelle la "partie large"). La seconde balise : <An> (Agent n) qui doit toujours être incluse dans la première sert à désigner l'endroit précis de la phrase où l'agent est effectivement identifié (ce qu'on appele la "partie centrale"). La partie "centrale" peut désigner des zones non contigues dans le texte. Ces deux familles de balises ont chacune un indice n compris entre 1 et 9 permettant d'identifier sans ambiguïté les différents agents qui apparaissent dans la phrase, sans qu'il y ait de sémantique particulière attachée à l'attribution de ces numéros. Lorsqu'une zone AF délimite le même fragment de texte qu'une zone A, les deux balises sont simultanément présentes pour des raisons de cohérence : <AFn> <An> ... </An> </AFn>. Par ailleurs, de manière logique, l'indice de la balise <An> doit être le même que celui de l'<AFn> englobant.

Enfin, on distingue deux types d'agents : « géniques » (notés GA) et « non géniques » (notés NGA). Les agents géniques « GAn » (et GAFn associés) ne peuvent apparaitre que dans les balises GENIC-INTERACTION et NON-GENIC-TARGET-INTERACTION. Les agents non géniques « NGAn » (et NGAFn associés) ne peuvent apparaître que les balises GENIC-INTERACTION et NON-GENIC-AGENT-INTERACTION.

1.2.2.1 Agents géniques <GAn>

Tous les agents géniques présents dans une phrase doivent être désignés à l'aide d'une balise <GAn> (incluse dans une balise « centrale » <GAFn>).

L'attribut TYPE indique la nature de l'agent et l'attribut ROLE précise son implication dans l'interaction. L'attribut DIRECT indique si cet agent agit sur la cible directement c'est-à-dire sans faire intervenir une autre entité génique, à quelque niveau moléculaire d'interaction que ce soit ; la valeur de cet attribut est donc « no » s'il y a d'autres gènes, protéines ou ARNm qui interviennent comme intermédiaires dans l'interaction.

```
<GAFn>

        Englobe le texte concernant l'agent et éventuellement une balise de type <GA>

        <GAn
                type = {gene, protein, arn, undefined}      /* nature de l'agent
                role = {required, modulate, undefined }     /* role dans l'interaction
                direct = {yes, no, undefined }              /* l'agent agit-il directement ?
        >
                                Texte identifiant précisément l'agent.
        </GAn>
</GAFn>
```

Notons que dans cette version de la DTD (Cf version 1.4) nous n'indiquons pas la (variation de la) concentration de l'agent qui déclenche de l'interaction (parmi□basal, increase, decrease).

1.2.2.2 Agents non géniques <NGAn>

Les agents non géniques <NGAn> sont décrits de selon les même règles d'annotation que les agents géniques. Le TYPE indique si l'agent est un «□produit□» susceptible d'être produit par l'organisme, ou s'il s'agit d'un agent externe. Dans cette version de la DTD on ne conserve pas l'influence de l'agent sur l'interaction (valeurs possibles parmi□neutral, strengthen, weaken).

<NGAFn>

Englobe le texte concernant l'agent et éventuellement une balise de type <NGA>

<NGAn
type = {endogenous, exogenous} /* type d'agent évoqué
>
Texte identifiant précisément l'agent.
</NGAn>
</NGAFn>

1.2.3 Description des Cibles

Les balises mises en œuvre pour annoter les fragments de phrases identifiant les «□entités cibles□» de l'interaction sont très semblables à celles utilisées dans le cas des agents (aux attributs prêts). Comme pour les agents, les étiquettes contiennent un numéro (sans sémantique particulière), cela permet de différencier les cibles lorsqu'une même interaction porte sur plusieurs cibles.

1.2.3.1 Cibles géniques <GTn>

<GTFn>

Englobe le texte concernant la cible et éventuellement une balise de type <GT>

<GTn type = {gene, protein, arn , undefined} >

Texte identifiant la cible.
</GTn>
</GTFn>

1.2.3.2 Cibles non géniques <NGTn>

<NGTFn>

Englobe le texte concernant la cible et éventuellement une balise de type <T>

<NGTn>
Texte identifiant la cible.

</NGTn>
</NGTFn>

1.2.4 Description des Interactions

L'annotation de l'interaction qui lie Agents et Cibles s'effectue par l'intermédiaire des balises <IF> et <I> qui permettent respectivement d'identifier les fragments "larges" et "centraux" de la phrase. Contrairement aux balises précédentes, il n'y a pas de numérotation de ces balises dans la mesure où par convention chaque description ne porte que sur une seule interaction.

<IF>

Englobe le texte concernant l'interaction et éventuellement une balise de type <I>

<I>

Texte identifiant précisément la nature de l'interaction.

</I>
</IF>

1.2.5 Description de la fiabilité (optionnel)

Ces dernières balises permettent d'annoter de manière optionnelle, les mots dans la phrase qui permettent d'associer un degré de certitude à l'interaction (au sens de la connaissance que l'on peut avoir de la réalité du phénomène décrit).

<CF>

Englobe le texte concernant l'expression d'un degré de confiance et une balise <C>

<C>

Texte identifiant le degré d'incertitude.

</C>
</IF>

2 Description de la DTD

Voici la DTD correspondant à la description effectuée dans la première partie.

```
<!-- DTD Caderige V1.51 -->

<!-- le code pourrait etre simplifie en definissant des ENTITIES
     voir http://ctdp.tripod.com/independent/web/dtd/index.html pour un
     tutorial sur l'écriture des DTD -->

<!ELEMENT annotated-document (abstract+) >

<!ATTLIST annotated-document     id           CDATA #REQUIRED>
<!ATTLIST annotated-document     reference    CDATA #REQUIRED>
<!ATTLIST annotated-document     description  CDATA #REQUIRED>
<!ATTLIST annotated-document     date         CDATA #REQUIRED>
<!ATTLIST annotated-document     author       CDATA #REQUIRED>

<!ELEMENT abstract (sentence+) >

<!ATTLIST abstract  id           CDATA #REQUIRED>
<!ATTLIST abstract  reference    CDATA #REQUIRED>

<!ELEMENT sentence ((genic-interaction+|
                                 non-genic-agent-interaction+|
                                 non-genic-target-interaction+|
                                 experiment|no-interaction),comment?) >

<!ATTLIST sentence  id           CDATA        #REQUIRED>
<!ATTLIST sentence  title        (yes|no)     "yes">


<!--   il faut avoir au moins un agent genique et une cible genique apres on peut avoir
une combinaison d'agents non geniques et cibles quelconques (? VOIR DISCUSSION ) -->

<!ELEMENT genic-interaction
                   (gaf1+|gtf1+|if+|(#PCDATA|gaf2|gaf3|gaf4|gtf2|gtf3|gtf4|cf|
                   ngaf1|ngaf2|ngaf3|ngaf4|ngtf1|ngtf2|ngtf3|ngtf4)*) >

<!--   il faut avoir au moins un agent non genique et une cible genique apres
on peut avoir une combinaison d'agents non geniques et cibles quelconques -->

<!ELEMENT non-genic-agent-interaction
                   (gtf1+|ngaf1+|if+|(#PCDATA|gtf2|gtf3|gtf4|cf|
                   ngaf2|ngaf3|ngaf4|ngtf1|ngtf2|ngtf3|ngtf4)*) >

<!--   il faut avoir au moins un agent genique et une cible non genique apres on peut
avoir une combinaison d'agents geniques ou non et cibles non geniques -->

<!ELEMENT non-genic-target-interaction
                   (gaf1+|ngtf1+|if+|(#PCDATA|gaf2|gaf3|gaf4|cf|
                   ngaf1|ngaf2|ngaf3|ngaf4|ngtf2|ngtf3|ngtf4)*) >

<!ELEMENT experiment       (#CDATA) >
<!ELEMENT no-interaction   (#CDATA) >
```

```
<!-- attributs de genic-interaction -->

<!ATTLIST genic-interaction      id     CDATA  #REQUIRED>
<!ATTLIST genic-interaction      type (mRNA-expression|transcriptional|
                                 post-transcriptional|protein-expression|
                                 traductional|post-traductional|undefined)
                                  "mRNA-expression" >
<!ATTLIST genic-interaction      assertion    (exist|non-exist)   "exist" >
<!ATTLIST genic-interaction      regulation   (activate|inhibit)  "activate" >
<!ATTLIST genic-interaction      uncertainty  (certain|probable|doubtful)  "certain">
<!ATTLIST genic-interaction      self-contained      (yes|no)         "yes" >
<!ATTLIST genic-interaction      confidence   (good|medium|poor)  "good" >

<!-- attributs de non-genic-agent-interaction -->

<!ATTLIST non-genic-agent-interaction   id     CDATA  #REQUIRED>
<!ATTLIST non-genic-agent-interaction   type
                                        (mRNA-expression|transcriptional|
                                        post-transcriptional|protein-expression|
                                        traductional|post-traductional|undefined)
                                         "mRNA-expression" >
<!ATTLIST non-genic-agent-interaction   assertion  (exist|non-exist)  "exist" >
<!ATTLIST non-genic-agent-interaction   regulation (activate|inhibit) "activate" >
<!ATTLIST non-genic-agent-interaction   uncertainty
                                        (certain|probable|doubtful)  "certain">
<!ATTLIST non-genic-agent-interaction   self-contained      (yes|no)     "yes" >
<!ATTLIST non-genic-agent-interaction   confidence   (good|medium|poor)  "good" >

<!-- attributs de non-genic-target-interaction -->

<!ATTLIST non-genic-target-interaction  id     CDATA  #REQUIRED>
<!ATTLIST non-genic-target-interaction  type   (to-be-determined)
                                               "to-be-determined">
<!ATTLIST non-genic-target-interaction  assertion  (exist|non-exist)  "exist" >
<!ATTLIST non-genic-target-interaction  regulation (activate|inhibit) "activate" >
<!ATTLIST non-genic-target-interaction  uncertainty
                                               (certain|probable|doubtful) "certain">
<!ATTLIST non-genic-target-interaction         self-contained      (yes|no)     "yes" >
<!ATTLIST non-genic-target-interaction  confidence   (good|medium|poor)  "good" >

<!-- définition des zone « larges », pour le moment on ne gère que les numéro 1-4 -->

<!ELEMENT gaf1 (#PCDATA|ga1)* >
<!ELEMENT gaf2 (#PCDATA|ga2)* >
<!ELEMENT gaf3 (#PCDATA|ga3)* >
<!ELEMENT gaf4 (#PCDATA|ga4)* >

<!ELEMENT ngaf1 (#PCDATA|nga1)* >
<!ELEMENT ngaf2 (#PCDATA|nga2)* >
<!ELEMENT ngaf3 (#PCDATA|nga3)* >
<!ELEMENT ngaf4 (#PCDATA|nga4)* >

<!ELEMENT gtf1 (#PCDATA|gt1)* >
<!ELEMENT gtf2 (#PCDATA|gt2)* >
<!ELEMENT gtf3 (#PCDATA|gt3)* >
<!ELEMENT gtf4 (#PCDATA|gt4)* >

<!ELEMENT ngtf1 (#PCDATA|ngt1)* >
<!ELEMENT ngtf2 (#PCDATA|ngt2)* >
```

```
<!ELEMENT ngtf3 (#PCDATA|ngt3)* >
<!ELEMENT ngtf4 (#PCDATA|ngt4)* >

<!ELEMENT if (#PCDATA|i)* >

<!ELEMENT cf (#PCDATA|c)* >

<!-- agents geniques -->

<!ELEMENT ga1       (#CDATA) >
<!ATTLIST ga1type   (gene|protein|arn|undefined)     "protein" >
<!ATTLIST ga1role   (required|modulate|undefined)    "modulate" >
<!ATTLIST ga1direct (yes|no|undefined)                     "yes" >

<!ELEMENT ga2       (#CDATA) >
<!ATTLIST ga2type   (gene|protein|arn|undefined)     "protein" >
<!ATTLIST ga2role   (required|modulate|undefined)    "modulate" >
<!ATTLIST ga2direct (yes|no|undefined)                     "yes" >

<!ELEMENT ga3       (#CDATA) >
<!ATTLIST ga3type   (gene|protein|arn|undefined)     "protein" >
<!ATTLIST ga3role   (required|modulate|undefined)    "modulate" >
<!ATTLIST ga3direct (yes|no|undefined)                     "yes" >

<!ELEMENT ga4       (#CDATA) >
<!ATTLIST ga4type   (gene|protein|arn|undefined)     "protein" >
<!ATTLIST ga4role   (required|modulate|undefined)    "modulate" >
<!ATTLIST ga4direct (yes|no|undefined)                     "yes" >

<!-- agents non geniques -->

<!ELEMENT nga1      (#CDATA) >
<!ATTLIST nga1      type   (endogenous|exogenous)          "endogenous">

<!ELEMENT nga2      (#CDATA) >
<!ATTLIST nga2      type   (endogenous|exogenous)          "endogenous">

<!ELEMENT nga3      (#CDATA) >
<!ATTLIST nga3      type   (endogenous|exogenous)          "endogenous">

<!ELEMENT nga4      (#CDATA) >
<!ATTLIST nga4      type   (endogenous|exogenous)          "endogenous">

<!-- cibles geniques -->

<!ELEMENT gt1       (#CDATA) >
<!ATTLIST gt1type   (gene|protein|arn|undefined)     "protein" >

<!ELEMENT gt2       (#CDATA) >
<!ATTLIST gt2type   (gene|protein|arn|undefined)     "protein" >

<!ELEMENT gt3       (#CDATA) >
<!ATTLIST gt3type   (gene|protein|arn|undefined)     "protein" >

<!ELEMENT gt4       (#CDATA) >
<!ATTLIST gt4type   (gene|protein|arn|undefined)     "protein" >

<!-- cibles non geniques -->
```

```
<!ELEMENT ngt1      (#CDATA) >
<!ELEMENT ngt2      (#CDATA) >
<!ELEMENT ngt3      (#CDATA) >
<!ELEMENT ngt4      (#CDATA) >

<!-- autres zones étroites et divers -->

<!ELEMENT i (#CDATA) >

<!ELEMENT c (#CDATA) >

<!ELEMENT comment (#CDATA) >
```

3 Exemples de document

Voici un exemple de document reprenant une partie des phrases utilisées par Claire Nédellec dans le document V1.0. Pour allèger l'étiquetage, on considèrera dans cette exemple que toutes les phrases proviennent d'un même résumé. Enfin, de manière à faciliter (?) la lecture le code de couleur suivant est utilisé dans les exemples, une impression couleur est recommandée :

	Agents	Cibles	Interaction	Fiabilité
Partie large	<AFn> .. </AFn>	<TFn> .. </TFn>	<IF> .. </IF>	<CF> .. </CF>
Parte centrale	<An> .. </An>	<Tn> .. </Tn>	<I> .. <I>	<C> .. </C>

<!DOCTYPE Caderige_Annotation " Caderige_Annotation-v1.5.dtd">

<ANNOTATED-DOCUMENT
id = "Exemple d'annotation DTD"
reference = "http://medline"
description = "Requête effectuée sur MedLine avec *bacillus subtilis transcription*"
date = "1/07/02"
author = "S. Lagarrigue, P. Bessière, A. Nazarenko, G. Bisson" >

<ABSTRACT id = "UI 1212" reference = "" >

<SENTENCE id = "4" title= no >

<EXPERIMENT>
The existence of the feedback loop was demonstrated by using antibodies prepared against SpoIIID to measure the level of spoIIID during sporulation of wild-type cells, mutants defective in sigma K production, and a mutant engineered to produce sigma K earlier than normal.
</EXPERIMENT>

<COMMENT>
L'expérience a pour but de tester une interaction entre SpoIIID et spoIIID dans une boucle de rétroaction impliquant sigma K. On peut s'attendre à ce que les phrases suivantes décrivent les résultats de l'expérience.
</COMMENT>

</SENTENCE>

<SENTENCE id = "15" title= no >

<NO-INTERACTION>
The sigma H factor, on the other hand, is dispensable for the switch in the position of the ftsZ assembly site.
</NO-INTERACTION>

</SENTENCE>

<SENTENCE id = "17" title= no >

<GENIC-INTERACTION
id = "1"
type = transcriptional
assertion = exist
regulation = activate
uncertainty = certain
self-contained = yes
confidence = good

<IF> <I> Induction of </I> </IF> the <TF1> Bacillus subtilis <T1 type=gene> kinA gene </T1>, which codes for a major kinase of the phosphorelay pathway </TF1>, <IF> required </IF> the <AF1> <A1 type=gene role=undefined direct=undefined> spo0H gene </A1>, coding for the sigma H protein </AF1>

</GENIC-INTERACTION >

</SENTENCE>

<SENTENCE id = "2" title=no >

<GENIC-INTERACTION
id = "1"
type = transcriptional
assertion = exist
regulation = activate
uncertainty = certain
self-contained = yes
text-clarity = good

<CF> <C> Previous studies showed </C> </CF> that <AF1> <A1 type=gene role=activate direct=undefined > spoIIID </A1> </AF1> <IF> <I> is needed to produce </I> </IF> <TF1> <T1 type=protein> sigma K</T1> </TF1>, but suggested that spoIIID represses sigma K directed transcription of genes encoding spore coat proteins.

</GENIC-INTERACTION >

<GENIC-INTERACTION
id = "2"
type = transcriptional
assertion = exist
regulation = inhibit
uncertainty = probable
self-contained = yes
text-clarity = good

Previous studies showed that spoIIID is needed to producesigma K, <CF> but <C> suggested </C> </CF> that <AF1> <A1 type="gene" role="modulate" direct= undefined> spoIIID </A1> </AF1> <IF> <I> represses </I> </IF> <TF1> sigma K directed transcription of <T1 type="gene"> genes encoding spore coat proteins </T1> </TF1>.

</GENIC-INTERACTION >

<GENIC-INTERACTION
id = "3"
type = transcriptional
assertion = exist
regulation = inhibit
uncertainty = probable
self-contained = yes
text-clarity = good

Previous studies showed that spoIIID is needed to produce sigma K, <CF> but <C> suggested </C> </CF> that spoIIID represses <AF1> <A1 type="protein" role="required" direct=undefined> sigma K</A1> </AF1> <IF> <I> directed transcription of </I> </IF> <TF1> <T1 type="gene"> genes encoding spore coat proteins </T1> </TF1>.

</GENIC-INTERACTION >

</SENTENCE>

<SENTENCE id = "1" >

<GENIC-INTERACTION
id = "1"
type = mRNA-expression
assertion = exist
regulation = inhibit
uncertainty = certain
self-contained = yes
text-clarity = good

Rather, <AF1> <A1 type=protein role=modulate direct=undefined> Sigma K </A1> </AF1> <CF> <C> appears to </C> </CF> <IF> <I> negatively regulate </I> the synthesis </IF> of <TF1> <T1 type=arn> spoIIID mRNA </T1> </TF1> by accelerating the disappearance of sigmaE RNA polymerase, which transcribes spoIIID.

</GENIC-INTERACTION >

<GENIC-INTERACTION
id = "2"
type = undefined

assertion = exist
regulation = inhibit
uncertainty = certain
self-contained = yes
confidence = good

Rather, <AF1> <A1 type=protein role=modulate direct=undefined> Sigma K </A1> </AF1> appears to negatively regulate the synthesis of spoIIID mRNA <IF> by <I>accelerating the disappearance </I> </IF> <TF1> <T1 type=protein> sigmaE </T1> RNA polymerase, which transcribes spoIIID </TF1>

</GENIC-INTERACTION >

<COMMENT>
R (claire) : Si on laisse tel que le TF1, on ne voit pas que Sigma K a un effet négatif indirect sur SpoIIID *parce qu'il concourt à faire disparaître* SigmaE RNA polymerase qui transcrit SpoIIID. On a ici le problème de la gestion des relations ternaires, on le traite en divisant les phrases par couples d'interaction, la disparition de SigmaE RNA polymerase est relatée ci-dessous.
</COMMENT>

<GENIC-INTERACTION
id = "3"
type = transcriptional
assertion = exist
regulation = activate
uncertainty = certain
self-contained = yes
confidence = good

Rather, Sigma K appears to negatively regulate the synthesis of spoIIID appears to negatively regulate the synthesis of spoIIID mRNA > by accelerating the disappearance of <AF1> <A1 type=protein role=required direct=yes> SigmaE RNA polymerase </A1> </AF1>, which <IF> <I>transcribes</I> </IF> <TF1> <T1 type=protein> spoIIID </T1> </TF1>

</GENIC-INTERACTION >

<COMMENT>
Le découpage en 2 interactions résout le problème soulevé dans V1.0. Il y a en fait trois interactions (et non 2) qui sont décrites dans cette phrase ...
</COMMENT>

</SENTENCE>

<SENTENCE id = "?" >

<GENIC-INTERACTION
id = "1"

type = transcriptional
assertion = exist
regulation = activate
uncertainty = certain
self-contained = yes
text-clarity = good

<IF> A <I> low level </I> of </IF> <AF1> <A1 type=protein role=modulate direct=yes> GerE </A1> </AF1>, <IF> <I> activated </I> transcription of </IF> <TF1> <T1 type=protein> CotD </T1> </TF1> by <AF2> <A2 type=protein role=required> GerE RNA polymerase </A2> </AF2>, but <CF> <C> in vitro </C> </CF>

</GENIC-INTERACTION >

<COMMENT>
On peut mettre la valeur Required pour le ROLE de l'agent 2 car il est possible d'utiliser une règle systèmatique associant la notion de RNA polymerase avec Required.
</COMMENT>

</SENTENCE>

<SENTENCE id = "?" >

<GENIC-INTERACTION
id = "1"
type = transcriptional
assertion = exist
regulation = inhibit
uncertainty = probable
self-contained = yes
text-clarity = good

These results <CF> suggest that </CF> <AF1> <A1 type=protein role=modulate direct=yes> yfhP </A1> </AF1>, <CF> <C> may act </C> </CF> as a <IF> <I> negative regulator </I> for the transcription of </IF> <TF1> <T1 type=gene> yfhQ </T1> </TF1>, <TF2> <T2 type=gene> yfhR </T2> </TF2>, <TF3> <T3 type=gene> sspE </T3> </TF3> and <TF4> <T4 type=gene> yfhP </T4> </TF4>.

</GENIC-INTERACTION >

</SENTENCE>

</ABSTRACT>

</ANNOTATED-DOCUMENT>