Format d'annotation pour CADERIGE

Version 1.61 du 6/07/04

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 2002 au LRI
V 1.5 : Réunion du 1er juillet 2002 à l'INRA Paris
V 1.51 : Réunion d'octobre 2002 au LRI
V1.52 : Réunion 7 mai 2003 à l'INRA Jouy
V1.53 : Modifications mineures sur les ensembles de valeurs (06/03)
V1.60 : Ajout des balises de coréférences par l'INRA (04/04)
V1.61 : Discussion à l'INRIA (10/05/04)

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 Commentaire. Les parties du document qui sont modifiées depuis la dernière version sont identifiées par un filet bleu à droite.

1.1 Format général d'un document annoté

<!DOCTYPE Caderige_Annotation " Caderige_Annotation-v1.61.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" >

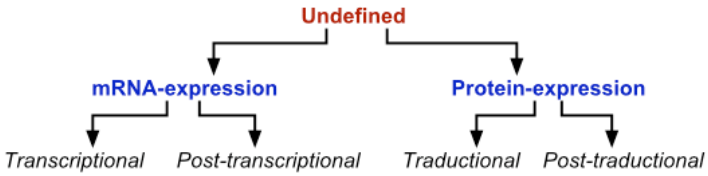
Dans la version actuelle, le document est constitué par une série de phrases délimitées par la balise <SENTENCE>. Pour chacune de ces phrases, on décrit ensuite la, (ou les), interactions, ou leur absence, à l'aide d'une des cinq balises suivantes :

- GENIC-INTERACTION
- NON-GENIC-AGENT-INTERACTION
- NON-GENIC-TARGET-INTERACTION
- EXPERIMENT
- CONTAINS-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. En partique on duplique donc la phrase autant de fois qu'il est nécessaire. Notons que les trois premières sont exclusives avec les deux dernières EXPERIMENT et CONTAINS-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'ensemble des cibles impliquées 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 « *niveau de régulation* » de l'interaction, c'est à dire à quel moment agit cette interaction. Les valeurs possibles sont organisées au sein de la taxonomie suivante :



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 « auto-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 « d'inférences » qui sont pas directement issues du contenu effectif de la phrase. Enfin, l'attribut CONFIDENCE indique le degré de confiance que l'annotateur accorde à son annotation (qui est liée au degré d'implicite présent dans le texte). Lorqu'un attribut accepte la valeur « Undefined » elle correspond à la valeur par défaut :

< 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 = { undefined, 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 ASSERTION et REGULATION.

		REGULATION		
		« activate »	*« inhibit »*	*« undefined »*
ASSERTION	*« exist »*	activation	inhibition	interaction 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.

< 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 = {undefined, 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/Sandrine) : doit-on supprimer le TYPE dans ce cas ?

< NON-GENIC-TARGET-INTERACTION
id = "identificateur de l'interaction"
type = {mRNA-expression, transcriptional, post-transcriptional, protein-expression, traductional, post-traductional, undefined}
assertion = {undefined, 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} Q (Gilles) : la valeur undefined a telle un sens dans ce cas (Cf DTD de l'INRA), à mon avis non …

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. 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.

< CONTAINS-NO-INTERACTION
Texte non annoté de la phrase.
</CONTAINS-NO-INTERACTION >

</SENTENCE>

< /ABSTRACT >

</ANNOTATED-DOCUMENT>

Dans l'ensemble des balises décrites dans cette partie (<ANNOTATED-DOCUMENT>, <ABSTRACT >, <SENTENCE> et les cinq balises caractérisant les phrases), l'annotateur peut ajouter librement des blocs de commentaires en texte libre délimité par la balise COMMENT :

<COMMENT> *Commentaires optionnels.*
</COMMENT>

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 l'interaction :

Informations nécessaires

- les agents (A) : les entités qui sont à la base de l'interaction,
- les cibles (T) : les entités sur lesquelles opère cette interaction,
- l'interaction (I) : le contrôle qui est effectué par l'interaction,

Informations optionnelles

- la fiabilité (C) : le degré de confiance que l'on a dans l'interaction décrite,
- la temporalité (TP) : les informations temporelles concernant l'interaction,
- la localisation (LC) : la localisation de l'interaction, des agents ou des cibles,
- l'expérimentation (EX) : la description des conditions expérimentales,
- les anaphores (AN) : les fragments correspondant à une anaphore.

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 huit types d'information (agent, cible, interaction, pertinence, etc) ne peuvent pas être imbriquées entre elles. 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 lorsqu'il est nécessaire d'étiquetter des zones non contigües dans la phrase qui correspondent à un seul et même type d'information. Dans ce cas c'est la réunion des différents groupes de mots balisés qui constitue 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 textes (via l'utilisation de couleurs, polices ...) peut s'effectuer en outre de manière plus simple puisqu'il n'y aurait plus de "chevauchements" entre des zones de types différents.

1.2.2 Description des Agents

Globalement, il y a trois familles de balises permettant de décrire les agents qui interviennent dans les différents types d'interaction.

- Les balises de type <AFn> (Agent Fragment n) qui servent 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" de la description.

- Les balises de type <An> (Agent n) qui doivent toujours être incluse dans la première et servent à désigner l'endroit précis de la phrase où l'agent est effectivement identifié, généralement par son nom, ce qu'on appele la "partie centrale".
- Les balises <CorefAn> (co-références de l'Agent n) qui doivent, elles aussi, être toujours incluses dans la première et qui permettent de mettre en évidence des co-références éventuelle à l'agent, indiquant par exemple dans le texte des désignations synonymiques de l'agent. Cette balise est optionnelle et ne peut pas être présente s'il n'y a pas de balise de type <An> .

Ces trois familles de balises ont chacune un indice n compris entre 1 et 9 (en pratique 4 dans les DTD actuelles) permettant d'identifier sans ambiguïté les différents agents impliqués (c-à-d qui sont nécessaires) dans l'interaction décrite dans la phrase. Notons qu'il n'y pas de sémantique particulière attachée à l'attribution de ces numéros. L'indice est placé dans la balise et non pas comme un attribut (ce qui pourrait-être plus logique) afin de permettre à l'éditeur d'annotation d'appliquer un style graphique différent à chaque agent et ainsi permettre de les différencier.

Lorsqu'une zone AF délimite le même fragment de texte qu'une zone A ou CorefA, les deux balises sont simultanément présentes par 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.

En pratique les balises A, AF et CorefA n'existent pas en tant que telles dans la DTD mais elles sont « spécialisées » selon deux types d'agents : « géniques » (notés GA) et « non géniques » (notés NGA). Les agents géniques « GAn » (et GAFn associés) ne peuvent apparaître que dans les balises GENIC-INTERACTION et NON-GENIC-TARGET-INTERACTION. Les agents non géniques « NGAn » (et NGAFn associés) ne peuvent apparaître que dans les balises GENIC-INTERACTION à la condition qu'il y ait au moins un autre agent génique qui participe à l'interaction et NON-GENIC-AGENT-INTERACTION si tous les agents sont non-géniques.

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>. Les coréférences à ces agents sont désignées à l'aide d'une balise <corefGAn>. Les deux balises sont incluse dans une balise « centrale » <GAFn>.

Dans <GA> la balise TYPE indique la nature de l'agent et l'attribut ROLE précise son implication dans l'interaction. L'attribut DIRECT indique qu'il y a effectivement un « contact moléculaire entre agent et cible » ne faisant pas intervenir une autre « entité génique », à quelque niveau moléculaire d'interaction que ce soit. Par exemple, c'est le cas lorsqu'une protéine se fixe sur le promoteur d'un gène pour le réguler, ou lorsqu'une protéine se fixe à une autre pour la cliver ou la phosphoryler, ce qui est moyen classique de l'activer. 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. Dans le cas de la balise <corefGA> l'attribut TYPE indique si la coréférence est un synonyme ou autre chose et l'attribut ID permet de différencier les différentes co-références si elles existent.

```
<GAFn>
        Englobe le texte concernant l'agent et éventuellement des balises de type <GA>
        et <corefGA>

        <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>

<corefGAn
type = {synonym,other} /* nature de la référence
id = "Identificateur de la coréférence"
>

Texte décrivant une coréférence à l'agent.

</corefGAn >
</GAFn>

Par exemple le fragment de phrase suivante [...] the phosphorylated protein, SpoOA, a major transcription factor is [...] sera annoté de la manière suivante au niveau des agents :

[...] <corefAG1 id=1 …> the phosphorylated protein </corefAG1>
<AG1 …> SpoOA </AG1>
<corefAG1 id=2 …> a major transcription factor </corefAG1> is [...]

Notons que dans cette version de la DTD (Cf version 1.4) nous n'indiquons plus la (variation de la) concentration de l'agent qui déclenche de l'interaction (parmi : basal, increase, decrease) dans la mesure ou cette valeur est trop difficile à préciser dans l'absolu.

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
C (Sandrine) : j'enlèverais cet attribut car il fait trop référence à des connaissances implicites. J'ai pu le remarquer avec les 2 étudiants qui ont annoter 1000 phrases. C'est déjà le cas pour distinguer parfois agent génique / d'agent non génique en particulier quand l'agent non génique est une substance décrite par une abréviation qui peut alors facilement se confondre avec un nom de gène
>

Texte identifiant précisément l'agent.

</NGAn>

<corefNGAn
type = {synonym,other} /* nature de la référence
id = "Identificateur de la coréférence"
>

Texte décrivant une coréférence à l'agent.

</corefNGAn >

</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 (et leurs coréférences éventuelles) sont très semblables à celles utilisées dans le cas des agents (aux attributs prêts). Comme pour les agents, on distingue entre les entités de nature génique (GT) et non-génique (NGT) et les balises 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>

<corefNGTn
type = {synonym,other} /* nature de la référence
id = "Identificateur de la coréférence"
>
Texte décrivant une coréférence à la cible.
</corefNGTn >

</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>

<corefNGTn
type = {synonym,other} /* nature de la référence
id = "Identificateur de la coréférence"
>
Texte décrivant une coréférence à la cible.
</corefNGTn >

</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 doit porter que sur une seule interaction. Notons que la description de la nature de l'interaction est effectué non pas à ce niveau mais au celui de la phrase à l'aide des attributs présents dans les balises <GENIC-INTERACTION>, <NON-GENIC-AGENT-INTERACTION>, et <NON-GENIC-TARGET-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 balises permettent d'annoter de manière optionnelle, les mots dans la phrase qui permettent d'associer un degré de certitude à l'interaction qui est décrite (au sens de la connaissance que la phrase apporte sur 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>
</CF>

1.2.6 Description des informations temporelles (optionnel)

Ces balises d'annoter les parties de la phrase qui apporte une information sur le moment ou l'interaction se produit. Cette notion devra être ultérieurement précisée car les échelles de temps peuvent être très variables. Par exemple, une interaction peut se produire en même temps qu'une autre interaction ou durant l'ensemble de la phase de développement d'un organisme …

<TPF>
Englobe le texte concernant l'expression d'un degré de confiance et une balise <C>
<TP>
Texte identifiant le degré d'incertitude.
</TP>
</TPF>

1.2.7 Description des informations sur la localisation (optionnel)

Ces balises d'annoter les parties de la phrases qui apporte une information sur la localisation de l'interaction et des entités qui y participent. Cette notion devra être ultérieurement précisée car les localisations décrites peuvent être très variables : organite, cellule, organe, …

<LCF>
Englobe le texte concernant l'expression d'un degré de confiance et une balise <C>
<LC>
Texte identifiant le degré d'incertitude.
</LC>
</LCF>

1.2.8 Description des informations sur les conditions expérimentale (optionnel)

Ces balises d'annoter les parties de la phrases qui apporte une information sur les conditions expérimentales qui ont permis de mettre en évidence l'interaction. C (Gilles) : peut-on donner ici 1-2 exemples de conditions expérimentales

<EXF>
Englobe le texte concernant l'expression d'un degré de confiance et une balise <C>
<EX>
Texte identifiant le degré d'incertitude.
</EX>
</EXF>

1.2.9 Description des informations sur les anaphores (optionnel)

Ces balises d'annoter les parties de la phrases qui correspondent à une anaphore. C (Gilles) : peut-on donner ici 1 exemple d'anaphore

<ANF>
Englobe le texte concernant l'expression d'un degré de confiance et une balise <C>
<AN>
Texte identifiant le degré d'incertitude.
</AN>
</ANF>

1.3 Description de la DTD

Voici la DTD correspondant à la description effectuée dans la première partie. Des modifications ont été introduites par l'INRA de manière à ce que lors de la validation on accepte la présence de partie de textes non annoté dans les phrases (ce qui est logique puisque les fragments ne se rapportent pas toujours à des informations pertinentes ou identifiables par les balises). Quelques points sont à vérifier/discuter :

- Il semble que les DTD (version <1.60) ne soient pas conformes aux définitions données dans ce document en ce qui concerne les types d'interactions. Par exemple, GENIC-INTERACTION ne doit pas contenir de cibles non-géniques.
- La DTD actuelle est plus permissive que les précédentes en ce qui concerne la présence des agents et des cibles il faudrait voir si l'on ne peut pas ajouter d'avantage de contraintes.
- Doit-on continuer à mettre des valeurs par défaut ? Dans la DTD de l'INRA tout est noté avec des REQUIRED ce qui force l'utilisateur à entrer/choisir explicitement une valeur.

```
<?xml version="1.0" encoding="UTF-8"?>
<!-- DTD Caderige V1.61 -->
<!-- du 07/07/04 -->

<!-- Le code pourrait etre simplifie en definissant des ENTITIES.  -->

<!-- On pourra voir le site du W3c http://www.w3schools.com/dtd/dtd_intro.asp ou
 http://ctdp.tripod.com/independent/web/dtd/index.html pour avoir un tutorial
 concernant l'ecriture des DTD -->

<!ELEMENT annotated-document (#PCDATA|abstract|comment)* >

<!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 (#PCDATA|sentence|comment)* >

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

<!-- les differents types d'interaction (ou non interaction), il est maintenant possible de tout
melanger et de mettre des commentaires ou l'on desire -->

<!ELEMENT sentence (#PCDATA
                    |genic-interaction
                    |non-genic-agent-interaction
                    |non-genic-target-interaction
                    |experiment
                    |contains-no-interaction
                    |comment )* >

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

<!-- Pour genic-interaction il faut que l'un au moins des agents (numerote 1 par convention) et
l'ensemble des cibles soient genique (proteines, des genes ou des ARNm) -->

<!ELEMENT genic-interaction  (#PCDATA
                              |gaf1  |gaf2 |gaf3 |gaf4
                                     |ngaf2|ngaf3|ngaf4
                              |gtf1  |gtf2 |gtf3 |gtf4
                              |if    |cf   |tpf  |lcf  |exf  |anf )* >
```

```
<!-- Alternative plus contrainte, mais est-elle correcte ? Si oui l'appliquer aux autres cas. Ici,
l'idée est de dire qu'à partir du moment ou l'on commence à baliser la phrase il doit y avoir au
moins un agent et une cible geniques et la descriptions de la zone d'interaction ...

<!ELEMENT genic-interaction  (#PCDATA | (gaf1+ | gtf1+ |if+
                                        | (#PCDATA
                                               |gaf2 |gaf3 |gaf4
                                               |ngaf2|ngaf3|ngaf4
                                               |gtf2 |gtf3 |gtf4
                                               |cf   |tpf  |lcf  |exf  |anf )*)) >
-->

<!-- Pour non-genic-agent-interaction il faut que l'ensemble des agents soient non genique et que
l'une au moins des cibles (numerote 1 par convention) soit génique -->

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

<!-- Pour non-genic-target-interaction il faut que l'ensemble des cibles soient non genique et que
l'un au moins des agents (numerote 1 par convention) soit génique -->

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

<!-- les deux cas ou il n'y a pas d'interaction -->

<!ELEMENT experiment   (#PCDATA) >
<!ELEMENT contains-no-interaction (#PCDATA) >

<!-- 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)           "undefined" >
<!ATTLIST genic-interaction   assertion             (exist|non-exist)     "exist" >
<!ATTLIST genic-interaction   regulation            (activate|inhibit|undefined) "undefined" >
<!ATTLIST genic-interaction   uncertainty           (certain|probable|doubtful)  "certain">
<!ATTLIST genic-interaction   self-contained        (yes|no)           "yes" >
<!ATTLIST genic-interaction   confidence            (good|medium|poor)    "good" > <!-- Undefined ?
-->

<!-- 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)           "undefined" >
<!ATTLIST non-genic-agent-interaction       assertion             (exist|non-exist)  "exist" >
<!ATTLIST non-genic-agent-interaction       regulation            (activate|inhibit|undefined)
"undefined" >
```

```
<!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>
<!-- Probleme du type -->
<!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|undefined)  "undefined"
>
<!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" >

<!-- definition des zone  larges , pour le moment on ne gere que les numeros 1-4 -->

<!ELEMENT gaf1 (#PCDATA|ga1|corefga1)* >
<!ELEMENT gaf2 (#PCDATA|ga2|corefga2)* >
<!ELEMENT gaf3 (#PCDATA|ga3|corefga3)* >
<!ELEMENT gaf4 (#PCDATA|ga4|corefga4)* >

<!ELEMENT ngaf1 (#PCDATA|nga1|corefnga1)* >
<!ELEMENT ngaf2 (#PCDATA|nga2|corefnga2)* >
<!ELEMENT ngaf3 (#PCDATA|nga3|corefnga3)* >
<!ELEMENT ngaf4 (#PCDATA|nga4|corefnga4)* >

<!ELEMENT gtf1 (#PCDATA|gt1|corefgt1)* >
<!ELEMENT gtf2 (#PCDATA|gt2|corefgt2)* >
<!ELEMENT gtf3 (#PCDATA|gt3|corefgt3)* >
<!ELEMENT gtf4 (#PCDATA|gt4|corefgt4)* >

<!ELEMENT ngtf1 (#PCDATA|ngt1|corefngt1)* >
<!ELEMENT ngtf2 (#PCDATA|ngt2|corefngt2)* >
<!ELEMENT ngtf3 (#PCDATA|ngt3|corefngt3)* >
<!ELEMENT ngtf4 (#PCDATA|ngt4|corefngt4)* >


<!-- autres zone  larges -->

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

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

<!ELEMENT tpf (#PCDATA|tp)* >        <!-- temporalite -->
<!ELEMENT lcf (#PCDATA|lc)* >        <!-- localisation -->
<!ELEMENT exf (#PCDATA|ex)* >        <!-- experimentation -->
<!ELEMENT anf (#PCDATA|an)* >        <!-- anaphore -->

<!-- agents geniques -->

<!ELEMENT ga1  (#PCDATA) >
<!ATTLIST ga1  type    (gene|protein|arn|undefined)  "undefined" >
<!ATTLIST ga1  role    (required|modulate|undefined) "undefined" >
<!ATTLIST ga1  direct (yes|no|undefined)             "undefined" >

<!ELEMENT ga2  (#PCDATA) >
<!ATTLIST ga2  type    (gene|protein|arn|undefined)  "undefined" >
<!ATTLIST ga2  role    (required|modulate|undefined) "undefined" >
<!ATTLIST ga2  direct (yes|no|undefined)             "undefined" >

<!ELEMENT ga3  (#PCDATA) >
<!ATTLIST ga3  type    (gene|protein|arn|undefined)  "undefined" >
<!ATTLIST ga3  role    (required|modulate|undefined) "undefined" >
<!ATTLIST ga3  direct (yes|no|undefined)             "undefined" >
```

```
<!ELEMENT ga4  (#PCDATA) >
<!ATTLIST ga4  type    (gene|protein|arn|undefined)  "undefined" >
<!ATTLIST ga4  role    (required|modulate|undefined) "undefined" >
<!ATTLIST ga4  direct (yes|no|undefined)                          "undefined" >

<!-- agents non geniques -->

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

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

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

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

<!-- cibles geniques -->

<!ELEMENT gt1  (#PCDATA) >
<!ATTLIST gt1  type    (gene|protein|arn|undefined)  "undefined" >

<!ELEMENT gt2  (#PCDATA) >
<!ATTLIST gt2  type    (gene|protein|arn|undefined)  "undefined" >

<!ELEMENT gt3  (#PCDATA) >
<!ATTLIST gt3  type    (gene|protein|arn|undefined)  "undefined" >

<!ELEMENT gt4  (#PCDATA) >
<!ATTLIST gt4  type    (gene|protein|arn|undefined)  "undefined" >

<!-- cibles non geniques -->

<!ELEMENT ngt1         (#PCDATA) >
<!ELEMENT ngt2         (#PCDATA) >
<!ELEMENT ngt3         (#PCDATA) >
<!ELEMENT ngt4         (#PCDATA) >

<!-- gestions des zones etroites -->

<!ELEMENT i (#PCDATA) >

<!ELEMENT c (#PCDATA) >

<!ELEMENT tp (#PCDATA) >
<!ELEMENT lc (#PCDATA) >
<!ELEMENT ex (#PCDATA) >
<!ELEMENT an (#PCDATA) >

<!ELEMENT comment (#PCDATA) >

<!-- gestion des coreferences -->


<!ELEMENT corefnga1    (#PCDATA) >
<!ATTLIST corefnga1    type    (synonym|other) #REQUIRED>
<!ATTLIST corefnga1    id      CDATA           #REQUIRED>

<!ELEMENT corefnga2    (#PCDATA) >
<!ATTLIST corefnga2    type    (synonym|other) #REQUIRED>
<!ATTLIST corefnga2    id      CDATA           #REQUIRED>

<!ELEMENT corefnga3    (#PCDATA) >
<!ATTLIST corefnga3    type    (synonym|other) #REQUIRED>
<!ATTLIST corefnga3    id      CDATA           #REQUIRED>
```

```
<!ELEMENT corefnga4   (#PCDATA) >
<!ATTLIST corefnga4   type    (synonym|other) #REQUIRED>
<!ATTLIST corefnga4   id      CDATA           #REQUIRED>



<!ELEMENT corefga1    (#PCDATA) >
<!ATTLIST corefga1    type    (synonym|other) #REQUIRED>
<!ATTLIST corefga1    id      CDATA           #REQUIRED>

<!ELEMENT corefga2    (#PCDATA) >
<!ATTLIST corefga2    type    (synonym|other) #REQUIRED>
<!ATTLIST corefga2    id      CDATA           #REQUIRED>

<!ELEMENT corefga3    (#PCDATA) >
<!ATTLIST corefga3    type    (synonym|other) #REQUIRED>
<!ATTLIST corefga3    id      CDATA           #REQUIRED>

<!ELEMENT corefga4    (#PCDATA) >
<!ATTLIST corefga4    type    (synonym|other) #REQUIRED>
<!ATTLIST corefga4    id      CDATA           #REQUIRED>


<!ELEMENT corefgt1    (#PCDATA) >
<!ATTLIST corefgt1    type    (synonym|other) #REQUIRED>
<!ATTLIST corefgt1    id      CDATA           #REQUIRED>

<!ELEMENT corefgt2    (#PCDATA) >
<!ATTLIST corefgt2    type    (synonym|other) #REQUIRED>
<!ATTLIST corefgt2    id      CDATA           #REQUIRED>

<!ELEMENT corefgt3    (#PCDATA) >
<!ATTLIST corefgt3    type    (synonym|other) #REQUIRED>
<!ATTLIST corefgt3    id      CDATA           #REQUIRED>

<!ELEMENT corefgt4    (#PCDATA) >
<!ATTLIST corefgt4    type    (synonym|other) #REQUIRED>
<!ATTLIST corefgt4    id      CDATA           #REQUIRED>

<!ELEMENT corefngt1   (#PCDATA) >
<!ATTLIST corefngt1   type    (synonym|other) #REQUIRED>
<!ATTLIST corefngt1   id      CDATA           #REQUIRED>

<!ELEMENT corefngt2   (#PCDATA) >
<!ATTLIST corefngt2   type    (synonym|other) #REQUIRED>
<!ATTLIST corefngt2   id      CDATA           #REQUIRED>

<!ELEMENT corefngt3   (#PCDATA) >
<!ATTLIST corefngt3   type    (synonym|other) #REQUIRED>
<!ATTLIST corefngt3   id      CDATA           #REQUIRED>

<!ELEMENT corefngt4   (#PCDATA) >
<!ATTLIST corefngt4   type    (synonym|other) #REQUIRED>
<!ATTLIST corefngt4   id      CDATA           #REQUIRED>
```

2 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>