
**Gestion des ressources langagières —
Cadre d'annotation morphosyntaxique
(MAF)**

*Language resource management — Morpho-syntactic annotation
framework (MAF)*

iTeh Standards
(<https://standards.iteh.ai>)
Document Preview

ISO 24611:2012

<https://standards.iteh.ai/catalog/standards/iso/442a57b7-65de-4d5f-9c8f-34c401137730/iso-24611-2012>



iTeh Standards
(<https://standards.iteh.ai>)
Document Preview

ISO 24611:2012

<https://standards.iteh.ai/catalog/standards/iso/442a57b7-65de-4d5f-9c8f-34c401137730/iso-24611-2012>



DOCUMENT PROTÉGÉ PAR COPYRIGHT

© ISO 2012, Publié en Suisse

Droits de reproduction réservés. Sauf indication contraire, aucune partie de cette publication ne peut être reproduite ni utilisée sous quelque forme que ce soit et par aucun procédé, électronique ou mécanique, y compris la photocopie, l'affichage sur l'internet ou sur un Intranet, sans autorisation écrite préalable. Les demandes d'autorisation peuvent être adressées à l'ISO à l'adresse ci-après ou au comité membre de l'ISO dans le pays du demandeur.

ISO copyright office
Ch. de Blandonnet 8 • CP 401
CH-1214 Vernier, Geneva, Switzerland
Tel. + 41 22 749 01 11
Fax + 41 22 749 09 47
copyright@iso.org
www.iso.org

Sommaire

Page

Avant-propos.....	v
Introduction.....	vi
1 Domaine d'application.....	1
2 Références normatives.....	1
3 Termes et définitions	1
4 Le métamodèle MAF	4
4.1 Vue d'ensemble	4
4.2 Métamodèle MAF	5
5 Segmentation	6
5.1 Aspect général	6
5.2 Description formelle: <token>.....	7
5.3 Notation enchâssée	8
5.4 Représentation alternative pour les documents conformes à la TEI	8
5.5 Notation déportée	9
5.6 Attributs informatifs	10
5.7 Compléter la notation enchâssée.....	10
5.7.1 Joindre des segments dans le mode enchâssé.....	11
5.7.2 Segments chevauchants	11
6 Les mots-formes en tant qu'unités linguistiques	12
6.1 Description formelle: <wordForm>.....	13
6.2 Attachement de segment	13
6.2.1 Un segment, un mot-forme.....	13
6.2.2 Plusieurs segments contigus, un mot-forme	13
6.2.3 Plusieurs segments discontigus, un mot forme	13
6.2.4 Absence de segment, un mot-forme.....	14
6.2.5 Un segment, plusieurs mots-formes	14
6.3 Référencer les entrées lexicales.....	15
6.4 Mots-formes composés.....	16
6.5 Identification des mots-formes au sein d'un document conforme à la TEI	16
7 Contenu morphosyntaxique	19
7.1 Aspect général	19
7.2 Utiliser les structures de traits	19
7.3 Balises morphosyntaxiques compactes.....	20
7.4 Les bibliothèques FSR	20
7.5 Conception des ensembles de balises	21
7.6 Description formelle: <tagset>	23
8 Gestion des ambiguïtés.....	23
8.1 Ambiguïtés du contenu des mots-formes	23
8.2 Ambiguïtés lexicales.....	24
8.3 Ambiguïtés structurelles	24
8.3.1 Ambiguïtés structurelles avec des mots-formes	24
8.3.2 Ambiguïtés structurelles avec les segments.....	25
8.4 Variantes structurées simplement.....	25

8.4.1	Représentation linéaire non ambiguë.....	25
8.4.2	Représentation mixte linéaire et en treillis	26
8.5	Expanser les variantes simplifiées	27
8.5.1	Séparer les segments et les mots-formes	27
8.5.2	Envelopper dans les treillis locaux.....	27
8.5.3	Fusion de treillis locaux.....	28
8.5.4	Suppression de <wfAlt>	30
8.6	Description formelle: <wfAlt> and <fsm>	30
Annexe A (informative) Exemple encodé selon la sérialisation MAF.....		31
Annexe B (normative) Spécification MAF.....		34
B.1	Eléments	34
B.1.1	<dcs/>.....	34
B.1.2	<fsm>	35
B.1.3	<maf>.....	35
B.1.4	<tagset>	36
B.1.5	<token>	36
B.1.6	<transition>	37
B.1.7	<wfAlt>.....	37
B.1.8	<wordForm>	38
B.2	Classes de modèles	39
B.3	Classes d'attributs	39
B.3.1	att.token.information	39
B.3.2	att.token.join.....	40
B.3.3	att.token.span.....	40
B.3.4	att.wordForm.content	40
B.3.5	att.wordForm.tokens	41
B.4	Macros	41
B.4.1	data.certainty.....	41
B.4.2	data.code	41
B.4.3	data.count	42
B.4.4	data.duration.w3c.....	42
B.4.5	data.enumerated	42
B.4.6	data.key.....	43
B.4.7	data.language.....	43
B.4.8	data.name.....	44
B.4.9	data.numeric.....	45
B.4.10	data.pointer.....	45
B.4.11	data.probability	46
B.4.12	data.temporal.w3c	46
B.4.13	data.truthValue	46
B.4.14	data.word	47
B.4.15	data.xTruthValue	47
Annexe C (normative) Catégories de données morphosyntaxiques.....		48
Bibliographie		62

Avant-propos

L'ISO (Organisation internationale de normalisation) est une fédération mondiale d'organismes nationaux de normalisation (comités membres de l'ISO). L'élaboration des Normes internationales est en général confiée aux comités techniques de l'ISO. Chaque comité membre intéressé par une étude a le droit de faire partie du comité technique créé à cet effet. Les organisations internationales, gouvernementales et non gouvernementales, en liaison avec l'ISO participent également aux travaux. L'ISO collabore étroitement avec la Commission électrotechnique internationale (IEC) en ce qui concerne la normalisation électrotechnique.

Les procédures utilisées pour élaborer le présent document et celles destinées à sa mise à jour sont décrites dans les Directives ISO/IEC, Partie 1. Il convient, en particulier de prendre note des différents critères d'approbation requis pour les différents types de documents ISO. Le présent document a été rédigé conformément aux règles de rédaction données dans les Directives ISO/IEC, Partie 2 (voir www.iso.org/directives).

L'attention est appelée sur le fait que certains des éléments du présent document peuvent faire l'objet de droits de propriété intellectuelle ou de droits analogues. L'ISO ne saurait être tenue pour responsable de ne pas avoir identifié de tels droits de propriété et averti de leur existence. Les détails concernant les références aux droits de propriété intellectuelle ou autres droits analogues identifiés lors de l'élaboration du document sont indiqués dans l'Introduction et/ou dans la liste des déclarations de brevets reçues par l'ISO (voir www.iso.org/brevets).

Les appellations commerciales éventuellement mentionnées dans le présent document sont données pour information, par souci de commodité, à l'intention des utilisateurs et ne sauraient constituer un engagement.

Pour une explication de la signification des termes et expressions spécifiques de l'ISO liés à l'évaluation de la conformité, ou pour toute information au sujet de l'adhésion de l'ISO aux principes de l'Organisation mondiale du commerce (OMC) concernant les obstacles techniques au commerce (OTC), voir le lien suivant: www.iso.org/iso/fr/avant-propos.html.

Le comité chargé de l'élaboration du présent document est l'ISO/TC 37, *Terminologie et autres ressources langagières et ressources de contenu*, sous-comité SC4, *Gestion de ressources linguistiques*.

Introduction

L'ISO/TC 37/SC 4 se concentre sur la définition des modèles et des formats utilisés pour représenter les ressources linguistiques annotées. A cette fin, il généralise la stratégie de modélisation initialisée par son comité frère le SC 3 pour la représentation des données terminologiques [Romary, 2001], selon laquelle les modèles de données linguistiques sont considérés comme la combinaison d'un patron de données génériques (un métamodèle), qui est ensuite perfectionné au moyen d'une sélection de catégories de données qui fournissent les descripteurs correspondant à ce niveau spécifique d'annotation. Ces modèles sont définis indépendamment des formats spécifiques et permettent à l'implémenteur de disposer de l'outil conceptuel nécessaire pour concevoir et comparer les formats en fonction de leurs niveaux d'interopérabilité.

Pour représenter tout type d'annotation, il est important de mettre à disposition une sémantique claire et fiable pour les divers descripteurs utilisés, soit sous la forme de traits valués formels, soit directement comme objets d'une représentation exprimée par exemple en XML. Pour que cette sémantique puisse être partagée entre différents schémas d'annotation et d'applications d'encodage, il convient de l'implémenter comme un registre centralisé de concepts: aussi, nous considérerons ces concepts comme des catégories de données. En tant que telles, il convient que ces catégories de données remplissent les conditions suivantes:

- d'un point de vue technique, elles doivent fournir des références uniques et stables (implémentées sous la forme d'identifiants pérennes au sens de l'ISO 24619) de telle manière que le concepteur d'un schéma spécifique d'encodage puisse les référencer dans ses spécifications. Ainsi, deux annotations seront considérées comme équivalentes quand elles feront référence à la même catégorie de données (en tant que trait et valeur).
- d'un point de vue descriptif, il convient que chaque référence sémantiquement unique soit associée à une documentation précise combinant une explication en prose de la signification du descripteur avec l'expression des contraintes spécifiques qui portent sur la catégorie.

Ces dernières années, l'ISO a développé un cadre général pour représenter et maintenir un tel registre de catégories de données couvrant tous les domaines des ressources linguistiques. Cette initiative, spécifiée par l'ISO 12620, a abouti à l'implémentation d'un environnement mis en ligne afin d'une part de fournir l'accès à toutes les catégories de données qui ont été normalisées dans le contexte des activités liées aux diverses ressources linguistiques au sein de l'ISO, et d'autre part spécifiquement au titre de la maintenance du registre de catégories de données. Le système propose aussi un accès aux diverses catégories de données que les praticiens des technologies du langage ont définies dans le cadre de leur propre travail et qu'ils ont partagé ensuite avec la communauté.

Le registre de catégories de données, accessible via l'implémentation ISOCat (www.isocat.org) est juste un espace d'objets sémantiques n'offrant qu'un ensemble limité de contraintes ontologiques. L'objectif est de faciliter la maintenance d'un environnement au sein duquel de nouvelles catégories sont facilement insérées et réutilisées sans qu'il soit nécessaire de procéder à une vérification approfondie de la cohérence par rapport au reste du registre. En effet, les contraintes de base sont intrinsèques au modèle de catégorie de données tel que défini par l'ISO 12620:

- de simples relations génériques-spécifiques quand elles sont utiles à une identification exacte des descripteurs d'interopérabilité entre catégories de données. Par exemple, le fait que /properNoun/ soit une sous-catégorie de /noun/ permet de comparer des annotations morphosyntaxiques fondées sur différents niveaux de granularité;

- la description des domaines conceptuels au sens de l'ISO 11179 pour identifier, quand elle est connue ou identifiable la valeur possible de la dite catégorie de donnée complexe. Par exemple, elle peut être utilisée pour enregistrer que la valeur possible de `/grammaticalGender/` (limitée à un petit groupe de langues [Romary 2011]), peut être un sous-ensemble de `{/ masculine/, /feminine/ et /neutral/}`;
- des contraintes linguistiques spécifiques, soit sous la forme de notes d'application ou comme des restrictions explicites portant sur les domaines conceptuels des catégories de données. Par exemple, il est possible d'exprimer explicitement que `/grammaticalGender/` en français ne peut prendre que les deux valeurs: `{/masculine/ et /feminine/}`.

La présente Norme internationale fournit un cadre complet pour la représentation des annotations morphosyntaxiques (aussi dénommées annotations en partie du discours). Ce niveau d'annotation correspond à un premier niveau d'abstraction par rapport aux données linguistiques (textuelles ou parlées), dont la structure et la complexité peuvent varier considérablement en fonction de la langue à annoter, de même que selon les caractéristiques de l'outil d'annotation ou du schéma d'annotation utilisé.

Pour résoudre les problématiques complexes de l'ambiguïté et du déterminisme en annotation morphosyntaxique, la présente Norme internationale introduit un méta-modèle qui établit une distinction nette entre les deux niveaux que sont les segments (représentant le découpage de surface de la source) et les mots-formes (identifiant les abstractions lexicales associées aux groupes de segments). Ces deux niveaux partagent les caractéristiques suivantes: d'une part, ils peuvent être représentés comme de simples séquences et des graphes locaux tels que segmentations multiples et éléments ambigus, et d'autre part, toute combinaison N à M peut relier les segments et les mots-formes.

En tant que segments linguistiques (quelquefois dénommés 'tokens' ou 'markables' dans la littérature technique anglaise [par exemple, Carletta et al. 1997]), ces *segments* peuvent être enchâssés dans le document source comme une balise en ligne, ou peuvent y faire référence par l'intermédiaire d'annotations déportées ('stand-off annotation' en anglais).

En tant qu'abstractions linguistiques, les mots-formes peuvent être qualifiés par divers traits linguistiques caractérisant les propriétés morphosyntaxiques qui sont instanciées dans la réalisation de l'entrée lexicale dans le texte annoté. Ces propriétés peuvent prendre diverses formes: de la simple indication d'un lemme à une référence explicite à une entrée lexicale dans un dictionnaire. Dans la plupart des applications existantes de l'annotation morphosyntaxique, les propriétés linguistiques sont exprimées au moyen de balises; ces codes font référence aux structures de traits basiques (voir les exemples dans Monachini and Calzolari, 1994). Ces codes peuvent aussi fournir de l'information morphologique, incluant la partie du discours (par exemple, nom, adjectif ou verbe), et des traits comme le nombre, le genre, la personne, le mode et le temps du verbe.

En phase avec la stratégie générale de modélisation de l'ISO/TC 37, la présente Norme internationale/le cadre MAF fournit les moyens de mise en relation des balises morphosyntaxiques exprimées en tant que structures de traits (conformes à l'ISO 24610) avec les catégories de données d'ISOCat. Une annexe normative de la présente Norme internationale explicite un jeu de base de catégories de données qui peuvent être utilisées comme référence pour la plupart des tâches d'annotation morphosyntaxiques dans un contexte multilingue. Néanmoins, si des utilisateurs de la présente Norme internationale estiment que ces catégories sont inappropriées du point de vue de la couverture, du domaine d'application ou de la sémantique, ils sont invités à utiliser ISOCat pour définir leurs propres catégories en conformité avec les principes de l'ISO/TC 37.

Associé au méta-modèle, le cadre MAF fournit aussi une syntaxe XML par défaut qui peut être utilisée pour sérialiser les modèles d'annotation conformes. Etant donné que de nombreux projets existants sont basés sur les lignes directrices émanant du consortium TEI (Text Encoding Initiative, www.tei-c.org) — particulièrement dans les humanités numériques, où un encodage correct des sources textuelles est essentiel — la présente Norme internationale fournira aussi des informations sur la façon

concilier le modèle MAF et les encodages conformes à la TEI. En effet, les lignes directrices de la TEI offrent d'ores et déjà une grande variété de constructions et de mécanismes pour prendre en charge les nombreux défis posés par les corpus oraux et leurs annotations (Romary and Witt, 2012).

Enfin, il convient de noter que la présente Norme internationale constitue la base conceptuelle permettant d'élaborer la série de normes ISO 24614 relative à la segmentation des unités lexicales. La totalité des règles et principes généraux définis dans l'ISO 24614-1 de même que les contraintes exprimées dans des parties complémentaires traitant de langues spécifiques, doivent être appréhendés dans le respect de la dichotomie segment / mot-forme.

iTeh Standards
(<https://standards.iteh.ai>)
Document Preview

[ISO 24611:2012](https://standards.iteh.ai/catalog/standards/iso/442a57b7-65de-4d5f-9c8f-34c401137730/iso-24611-2012)

<https://standards.iteh.ai/catalog/standards/iso/442a57b7-65de-4d5f-9c8f-34c401137730/iso-24611-2012>

Gestion des ressources langagières — Cadre d'annotation morphosyntaxique (MAF)

1 Domaine d'application

La présente Norme internationale fournit un cadre pour la représentation des annotations des mots-formes dans les textes; ces annotations concernent les segments, leurs relations avec les unités lexicales, et leurs propriétés morphosyntaxiques.

Elle présente un métamodèle pour l'annotation morphosyntaxique qui référence les catégories de données dans le registre des catégories de données ISOCat (DCR tel que défini dans l'ISO 12620). Elle décrit aussi une sérialisation XML pour l'annotation morphosyntaxique, avec les équivalences des lignes directrices de la TEI (Text Encoding Initiative).

2 Références normatives

Les documents référencés sont indispensables à l'application de ce document. Pour les références datées, seule l'édition citée s'applique. Pour les références non datées, la dernière édition du document référencé s'applique (incluant ses éventuels amendements).

ISO 24610-1, *Gestion des ressources linguistiques — Structures de traits — Partie 1: Représentation de structures de traits*

[ISO 24611:2012](https://standards.iteh.ai/catalog/standards/iso/442a57b7-65de-4d5f-9c8f-34c401137730/iso-24611-2012)

<https://standards.iteh.ai/catalog/standards/iso/442a57b7-65de-4d5f-9c8f-34c401137730/iso-24611-2012>

3 Termes et définitions

Pour les besoins du présent document, les termes et définitions donnés dans l'ISO 24610-1 ainsi que les suivants s'appliquent:

3.1

GOA

DAG

graphe orienté acyclique

graphe contenant des arcs orientés et sans cycle

Note 1 à l'article: les graphes orientés acycliques sont des sous-ensembles des automates finis (3.4).

3.3

structure de trait

ensemble des spécifications de trait, utilisé dans le cadre d'annotation morphosyntaxique (MAF) pour exprimer le contenu morphosyntaxique

Note 1 à l'article: les structures de trait sont spécifiées dans l'ISO 24610-1.

3.4

AEF

FSA

automate fini

graphes comprenant plusieurs états avec un état initial et un état final, et un ensemble fini de transitions pour passer d'un état à l'autre

Note 1 à l'article: Voir aussi GOA (3.1).

3.5

graphème

unité minimale dans une langue écrite

EXEMPLE Lettre, pictogramme, idéogramme, numérique, ponctuation.

3.6

flexion

modification ou balise d'un lexème qui reflète ses propriétés morphosyntaxiques

3.7

forme fléchie

forme qu'un mot peut prendre dans une phrase ou une proposition

Note 1 à l'article: Une forme fléchie d'un mot est associée avec une combinaison de traits morphologiques comme le nombre grammatical ou le cas.

3.8

lemme

forme lemmatisée

forme conventionnelle choisie pour représenter un lexème

Note 1 à l'article: Dans les langues européennes, le lemme est habituellement le *singulier* s'il y a une variation en nombre, le *masculin* s'il y a une variation en genre, et l'*infinitif* pour tous les verbes. Dans certaines langues, certains noms sont défectifs au singulier, auquel cas on choisit le pluriel. Pour les verbes en arabe, le lemme est habituellement la troisième personne du singulier à l'aspect accompli

3.9

lexème

morphème généralement associé à un ensemble de mots-formes partageant un sens en commun

3.10

entrée lexicale

conteneur pour gérer un ensemble de mots-formes et éventuellement un ou plusieurs sens pour décrire un lexème

3.11

lexique

ressource comprenant une collection d'entrées lexicales pour une langue

3.12

morphème

plus petite unité linguistique porteuse de sens dans un discours et qui ne peut être divisée en de plus petites unités porteuses de sens

Note 1 à l'article: Un morphème est soit grammatical (grammème), soit lexical (lexème).

3.13**trait morphologique****trait morphosyntaxique**

trait induit à partir de la forme fléchie d'un mot

Note 1 à l'article: Le registre de catégories de données ISOCat fournit une liste complète de valeurs pour les langues européennes.

EXEMPLE "genre grammatical".

3.14**morphologie**

description de la structure et de la formation des mots-formes

3.15**balise morphosyntaxique****balise**

structure de trait utilisée systématiquement pour qualifier un mot-forme

3.16**ensemble de balises****tagset**

jeu d'étiquettes de segments

ensemble complet de balises, utilisé pour la description morphosyntaxique d'une langue

Note 1 à l'article: Pour décrire un ensemble de balises, il y a lieu d'utiliser le registre de catégories de données ISOCat.

3.17**partie du discours****catégorie grammaticale**

catégorie affectée à un mot en vertu de ses propriétés grammaticales et sémantiques

EXEMPLE nom, verbe.

Note 1 à l'article: Le registre de catégories de données ISOCat fournit une liste complète de valeurs pour des parties du discours.

3.18**phonème**

plus petite unité du système audio-phonatoire d'une langue

3.19**écriture**

ensemble de caractères graphiques utilisés dans la forme écrite d'une ou plusieurs langues

3.20**relation syntagmatique**

relation par laquelle sont associées des unités linguistiques de discours

3.21**segment**

séquence non vide contiguë de graphèmes ou de phonèmes dans un document

Note 1 à l'article: Pour des raisons éditoriales, certains schémas d'annotations peuvent étendre la notion de segment à une séquence vide. Voir la section sur l'attachement de segment (6.2).

3.22

segmentation

processus identifiant les segments

3.23

transcription

forme résultant de l'application d'une méthode cohérente d'écriture d'une production orale

3.24

translittération

forme résultant de la conversion d'une écriture dans une autre, habituellement via une correspondance une à une entre caractères

3.25

mot-forme

unité morphosyntaxique

unité linguistique contiguë ou non contiguë identifiée comme correspondant à une entrée lexicale dans une langue

Note 1 à l'article: Les mots-formes peuvent avoir ou ne pas avoir de réalisation acoustique ou graphique, ou peuvent correspondre à un ou plusieurs segments.

3.26

treillis de mots

ensemble des décompositions possibles d'un texte oral ou écrit, en mots-formes

Note 1 à l'article: Un treillis de mots possède les propriétés algébriques d'un graphe orienté acyclique avec un nœud initial et un nœud final.

Note 2 à l'article: Voir aussi GOA (3.1) et AEF (3.4).

4 Le métamodèle MAF

4.1 Vue d'ensemble

Les annotations morphosyntaxiques fournissent une couche importante des informations linguistiques d'un document. La présente Norme internationale est fondée sur un métamodèle qui opère une claire distinction entre les deux niveaux de segments (représentant la segmentation de surface de la source) et les mots-formes (identifiant des abstractions lexicales associées aux groupes de segments). Ces deux niveaux partagent les spécificités suivantes: d'une part, ils peuvent être représentés comme de simples séquences et comme des graphes locaux (par exemple, de multiples segmentations et des composés ambigus), et d'autre part, toute combinaison N à M peut relier les mots-formes et les segments. Cette Norme internationale délimite des séquences minimales et maximales dans les documents (qu'elles soient écrites ou orales) qui peuvent être identifiées comme mots-formes et cherche à catégoriser les critères linguistiques et distributionnels qui peuvent être utilisés pour marquer ces mots-formes au sein de contextes syntagmatiques plus étendus. Les unités minimales ne peuvent pas être décomposées plus avant en utilisant des critères similaires, mais elles peuvent néanmoins être subdivisées en fonction de propriétés morphologiques ou phonologiques. Les mots-formes peuvent être agrégés pour constituer des unités plus grandes (telles que mots composés ou des unités multi-mots) qui agissent en

tant qu'unités élémentaires pour d'autres niveaux de l'analyse linguistique, notamment la syntaxe. En particulier, les mots-formes correspondent au niveau non terminal défini dans l'ISO 24615.

4.2 Métamodèle MAF

La Figure 1 présente une vue simplifiée du métamodèle proposé pour les annotations morphosyntaxiques, alors que la Figure 2 présente une vue plus formelle fondée sur UML (Unified Modeling Language).

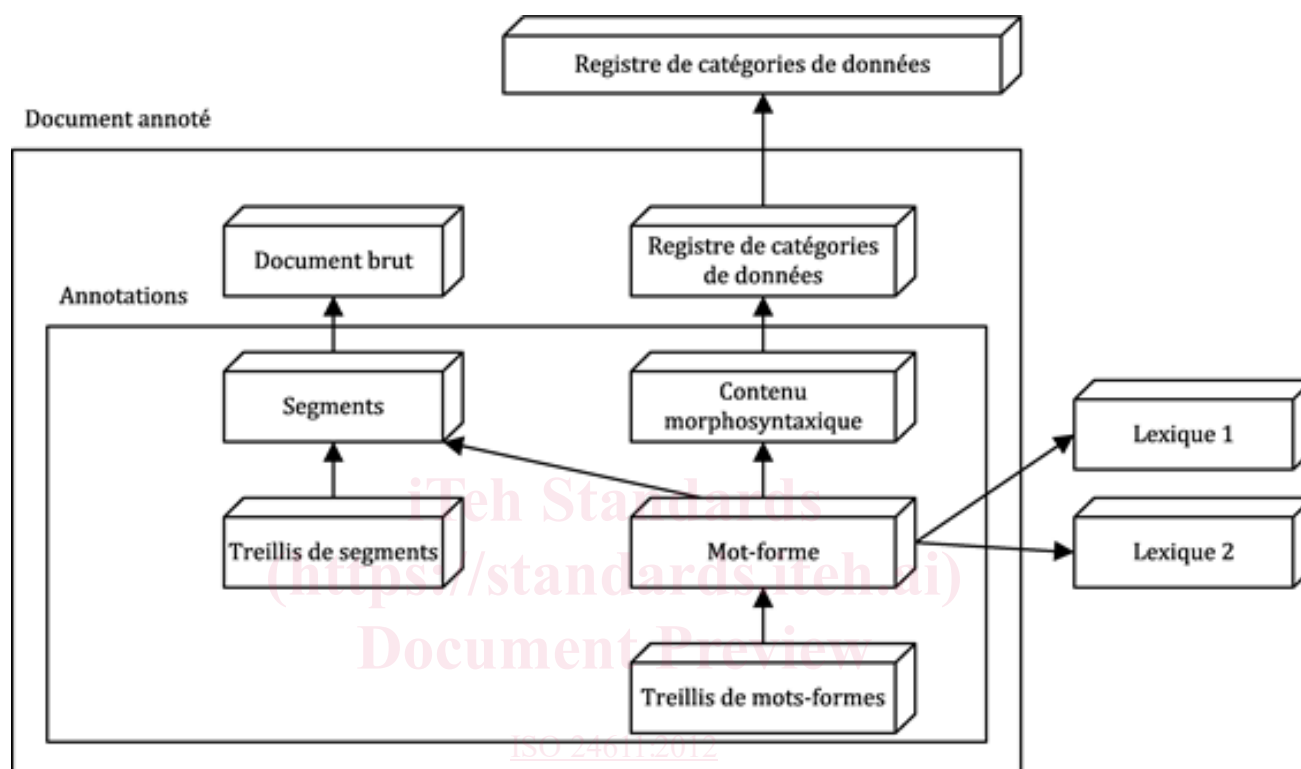


Figure 1 — Vue simplifiée du métamodèle MAF

Un document annoté comporte un document original et un ensemble d'annotations. Les annotations sont associées à des mots-formes correspondant à zéro ou plusieurs segments dans le document original. Un mot-forme peut aussi être associé à une entrée lexicale fournissant de l'information à propos de son lemme sous-jacent et sa forme fléchie. L'annotation morphosyntaxique associée au mot-forme est représentée par une balise dont la signification peut être exprimée sous la forme d'une structure de trait. L'ensemble des balises utilisées par un schéma d'annotation particulier est appelé jeu de balises, et correspond à ce qui est défini dans l'ISO 24610-2 pour les représentations de structures de traits (FSR) en tant que bibliothèque de structures de traits. Il convient que chaque catégorie discrète contenue dans le jeu de balises soit descriptible dans les termes du registre des catégories de données décrit dans l'ISO 12620 et implémenté dans ISOCat. Du fait que l'annotation peut être appliquée à la fois aux segments et aux mots-formes, une ambiguïté structurelle peut apparaître. Ainsi, l'annotation est typiquement conceptualisée comme un ou plusieurs flux, chacun représentant un treillis de mots, ou plus formellement, comme un graphe orienté acyclique (DAG).

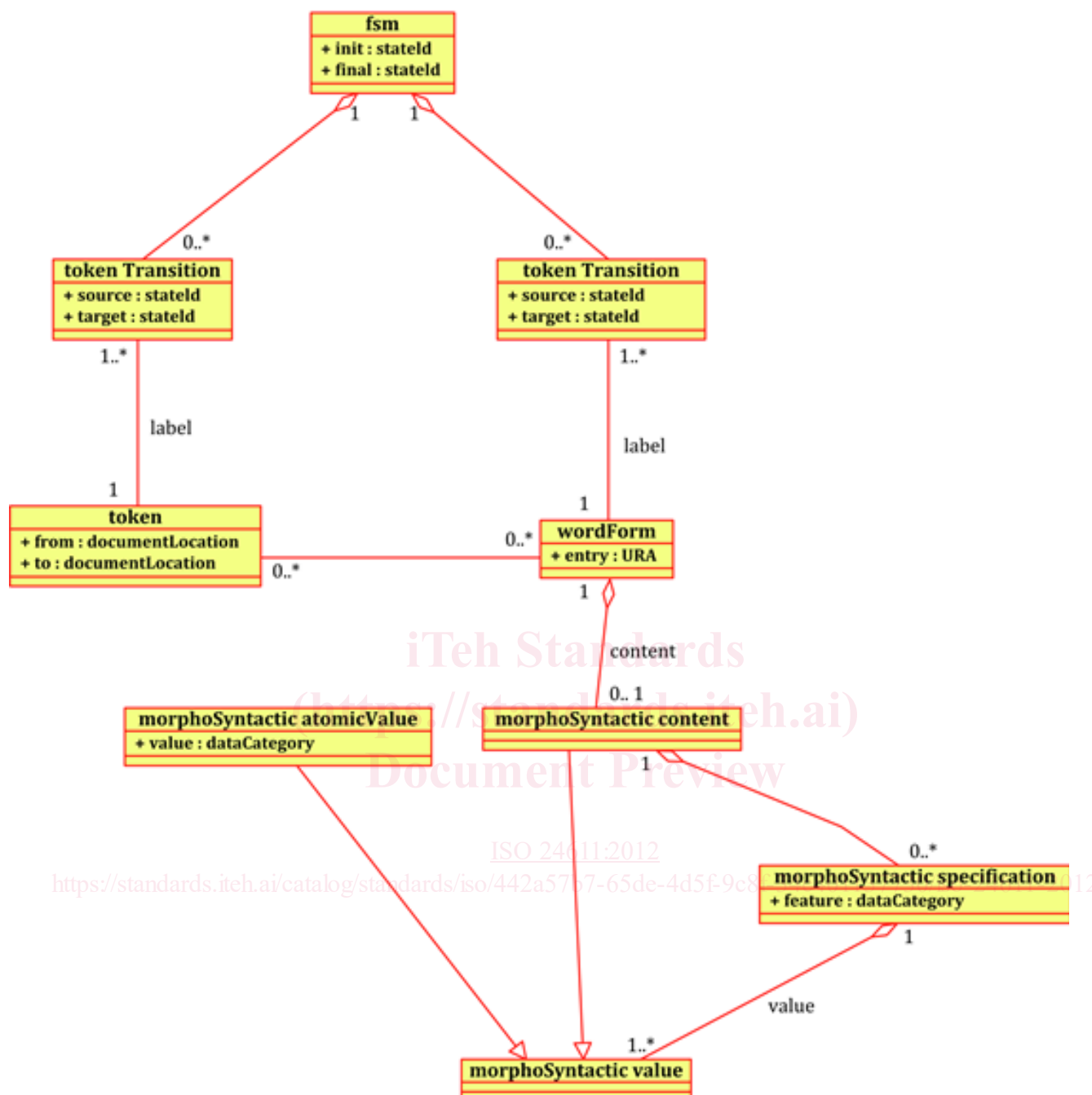


Figure 2 — Vue UML du métamodèle MAF

5 Segmentation

5.1 Aspect général

Les annotations morphosyntaxiques sont portées par les segments, appelés “tokens” en anglais, qui sont présents dans le flux du document, mais ceci n’implique pas que la segmentation résultante corresponde à une séquence de segments adjacents qui partitionne le document original. Il est particulièrement important de distinguer les mots-formes de leurs réalisations. Quelques sections d’un document peuvent ne porter aucune annotation (ex. marques typographiques, directives intermédiaires