Procédure de diffusion des publications de l'ATALA sur les archives ouvertes

Yannick Parmentier, Sylvain Pogodalla, Rachel Bawden, Matthieu Labeau, Iris Eshkol-Taravella

Rapport technique Septembre 2023

Auteurs : · e · s :

- Yannick Parmentier, Université de Lorraine / LORIA
- Sylvain Pogodalla, Inria Nancy Grand Est
- Rachel Bawden, Inria Paris
- Matthieu Labeau, TELECOM Paris
- Iris Eshkol-Taravella, Université Paris Nanterre / Modyco

PROCÉDURE DE DIFFUSION DES PUBLICATIONS DE L'ATALA SUR LES ARCHIVES OUVERTES

Association pour le Traitement Automatique des LAngues (ATALA)

Table des matières

1 Historique des publications de l'ATALA

Les activités de l'ATALA, association savante de type loi 1901 créée en 1959 œuvrant au développement du Traitement Automatique des Langues (TAL), incluent l'édition de la revue TAL depuis 1960, l'organisation de la conférence annuelle TALN, de sa session étudiante RECITAL (Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues) et d'ateliers associés depuis 1997, et l'organisation de journées d'étude sur des thématiques particulières.

1.1 La revue TAL

La revue TAL, est une revue à comité de lecture soutenue par l'Institut des Sciences Humaines et Sociales (InSHS) du CNRS, qui accepte des soumissions en français ou anglais. Elle a été publiée par HERMÈS Science / Lavoisier et diffusée en format papier jusqu'en 2006. Depuis cette date, elle est diffusée en format électronique sur le site de l'ATALA [1], et depuis peu également sur l'ACL Anthology (archive ouverte administrée par l'Association for Computational Linguistics) [2]. Des travaux sont en cours pour la mise en place d'une diffusion systématique des articles de la revue sur l'archive ouverte HAL.

1.2 La conférence TALN / RECITAL

Les conférences TALN et RECITAL sont organisées conjointement chaque année depuis 1994 (1999 pour RECITAL). Depuis 1997, l'organisation de TALN est désignée suite à un appel à candidatures. Le comité permanent de l'ATALA (CPerm) est en charge du recueil des candidatures, et participe au suivi de l'organisation des différentes éditions [3].

Jusqu'en 2012, les actes de TALN / RECITAL étaient produits par l'équipe organisatrice de la conférence, et diffusés généralement sur support papier, puis électronique (CD, clé USB). En 2012 (Grenoble), les actes de TALN / RECITAL ont pour la première fois été hébergés sur l'ACL Anthology (Bird et al. 2008). À la même époque ont été mises en ligne les "TALN archives", une archive ouverte des publications des conférences TALN / RECITAL compilée de manière semi-automatique (Boudin 2013) [4]. Les actes des éditions 2013 (Batz-sur-Mer), 2014 (Marseille) et 2015 (Caen) ont par la suite également été mis en ligne sur les TALN archives et sur l'ACL Anthology.

Lors de l'édition 2017 (Orléans), un script d'automatisation de la compilation des actes PDF à partir des données entrées par les auteur·e·s sur l'application de gestion de soumissions `easychair` [5] a été développé. Suite à cette expérience, des travaux ont alors été entrepris sous l'impulsion d'Emmanuel Morin et du CPerm,

1. https://www.atala.org/revuetal
2. https://aclanthology.org/venues/tal/
3. https://www.atala.org/-Conference-TALN-RECITAL
4. http://talnarchives.atala.org/index.html
5. https://easychair.org

pour développer un outil pérenne permettant non seulement de compiler de manière automatique les actes au format PDF, mais également dans les formats attendus par les archives ouvertes TALN archives et ACL Anthology.

Une première version de cet outil (appelé `taln2acl` [6]) a été utilisée pour compiler et diffuser les actes de TALN / RECITAL 2020 (Nancy), et également pour rétro-ingérer dans l'ACL Anthology les actes des éditions antérieures à 2012, et postérieures à 2015 [7]. Le support de l'archive ouverte HAL (Baruch 2007) et de la bibliographie DBLP (Ley 2002) y ont été ajoutés dans la foulée [8].

Sous l'impulsion de Sylvain Pogodalla, responsable des actes de l'édition 2020, une collection HAL nommée TALN-RECITAL [9], ainsi qu'une sous-collection HAL nommée JEP-TALN-RECITAL2020 [10] ont été créées. L'outil `taln2acl` a par la suite été utilisé pour compiler les actes des éditions 2021 (Lille) et 2022 (Avignon), et téléverser ces derniers dans l'ACL Anthology, les TALN archives et HAL (en mettant à jour les collections et sous-collections [11] correspondantes).

En 2023 (Paris), l'équipe organisatrice de TALN / RECITAL a décidé d'expérimenter une nouvelle application de gestion des soumissions, à savoir la plateforme `sciencesconf` [12]. À cette occasion, l'outil `taln2acl` a été étendu (et renommé `taln2x` [13]) afin de prendre en charge le format d'export de cette plateforme [14].

Depuis 2023, le Conseil d'Administration de l'ATALA comporte une personne chargée de mission "actes de la conférence TALN". Cette personne a pour but d'accompagner les équipes organisatrices de TALN / RECITAL dans la compilation et la diffusion des actes de la conférence. Par ailleurs, lors de l'appel à candidature à l'organisation de TALN / RECITAL, le CPerm demande à présent aux équipes organisatrices candidates d'identifier une personne responsable des actes. Le but de ces mesures est de fluidifier la procédure de compilation et de diffusion des actes, et d'éviter qu'elle ne repose sur un nombre très limité de personnes.

1.3 Les journées d'études de l'ATALA

L'ATALA organise également des journées d'études [15] donnant parfois lieu à la publication d'actes. Ces derniers pourraient à terme également être inclus dans la procédure de diffusion des publications de l'association. Ce point doit encore être discuté avec les personnes responsables des journées d'études au sein du Conseil d'Administration de l'association.

6. https://gitlab.com/parmenti/taln2acl
7. https://aclanthology.org/venues/jeptalnrecital/
8. Le support du format DBLP n'a en pratique jamais été utilisé car DBLP moissonne l'ACL Anthology.
9. https://hal.science/TALN-RECITAL
10. https://hal.science/JEP-TALN-RECITAL2020
11. https://hal.science/TALN-RECITAL2021 https://hal.science/TALN-RECITAL2022
12. https://www.sciencesconf.org/
13. https://talnarchives.gitlabpages.inria.fr/taln2x
14. Ce format est documenté à l'adresse https://doc.sciencesconf.org/gestion-editoriale/.
15. https://www.atala.org/journees

2 Description du processus de publication des actes de TALN / RECITAL

2.1 Sélection des articles présentés à la conférence

Comme de nombreux autres événements scientifiques (colloques, ateliers, journées d'études, journées scientifiques, etc.), les équipes d'organisation des conférences TALN / RECITAL sélectionnent les travaux qui vont être présentés lors de la conférence (et publiés dans les actes) au moyen d'un appel à soumission (Call for Papers [16]) et d'une application de gestion des soumissions [17]. Cet appel contient notamment les modalités de soumission (thèmes visés, contraintes de forme, dates limites, adresse de la plateforme de recueil des soumissions, etc.). L'application de gestion des soumissions offre plusieurs fonctionnalités très utiles (en plus du stockage des articles soumis) :

— communication avec les membres du comité de relecture ;

— affectation des relectures semi-automatisée (au moyen d'un système de vœux de relecture) ;

— collecte des avis de relecture ;

— diffusion des notifications et avis aux auteur·e·s ;

— versionnage des articles ;

— etc.

Comme mentionné précédemment, les systèmes de gestion de soumissions utilisés à ce jour par les équipes organisatrices de TALN / RECITAL sont `easychair` et `sciencesconf`. Il existe cependant d'autres systèmes de gestion de soumissions utilisés par la communauté du TAL (on peut mentionner en particulier `softconf` [18] et `openreview` [19]). Le choix de l'outil de gestion des soumissions est laissé à l'équipe organisatrice de la conférence et dépend généralement de contraintes financières et / ou pratiques.

Les étapes du processus de sélection sont décrites en Figure 1 ci-dessous.

16. Voir par exemple https://coria-taln-2023.sciencesconf.org/resource/page/id/4.
17. Voir par exemple https://easychair.org
18. https://softconf.com/
19. https://openreview.net/about

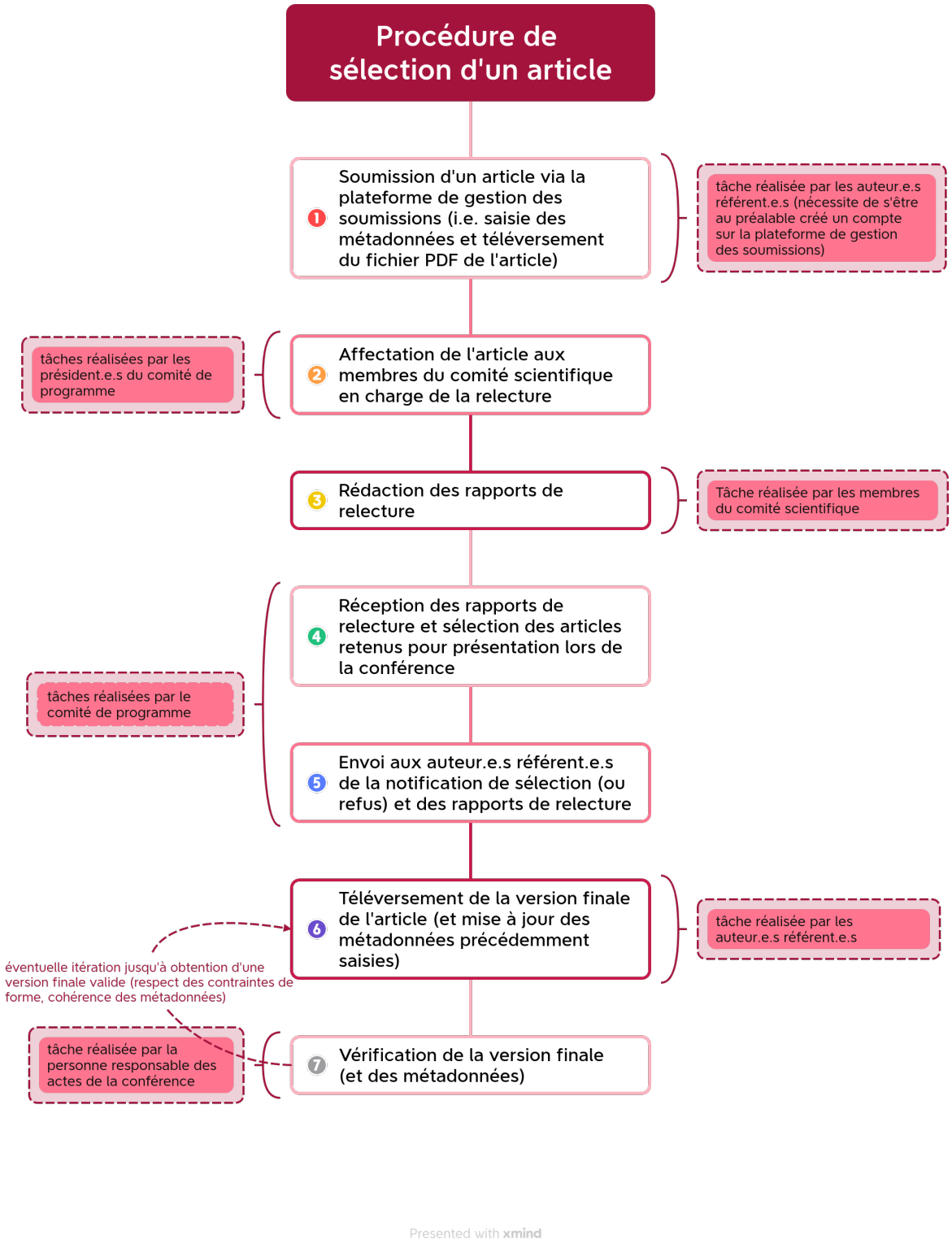

Figure 1 – Processus de sélection des articles présentés à la conférence.

2.2 Compilation des actes de la conférence

Les actes de la conférence TALN / RECITAL (et des ateliers associés) sont compilés de manière automatique à partir d'informations fournies par l'équipe organisatrice (e.g. préface, liste des comités, sessions, sponsors), ou par les auteur·e·s (métadonnées et fichier PDF de chaque article [20]) via la plateforme de gestion des soumissions.

> Chaque session (e.g. articles longs, articles courts, prises de position, démonstrations) donne lieu à un volume dédié [21] dans les actes de la conférence.

Concrètement, pour la personne responsable des actes, il s'agit de préparer :

— un fichier de configuration des actes au format YAML [22] contenant les métadonnées de la conférence [23],

— un dossier par volume respectant certaines contraintes de forme (e.g. avoir un sous-dossier nommé `pdf` contenant les différents articles de la session et un fichier `articles.csv` contenant les métadonnées des articles en question telles que produites par le système de gestion des soumissions) [24],

puis d'appeler l'outil `taln2x` (outil accessible en ligne de commande) pour lancer la compilation.

La version PDF des actes utilise un modèle (template) LaTeX fourni par l'outil `taln2x` et modifiable [25]. Cette version PDF (qu'elle soit compilée automatiquement ou manuellement) est pré-requise avant de pouvoir procéder au téléversement des actes dans les différentes archives ouvertes. À l'issue de la compilation, on obtient pour chaque volume :

— les actes complets (i.e. un fichier PDF) avec table des matières et hyperliens ;

— un extrait des actes hors articles (i.e. un fichier PDF allant de la page de garde jusqu'à la table des matières incluse, correspondant à ce que l'on nomme généralement le "frontmatter") ;

— un dossier contenant chacun des articles du volume avec entête et pied de page mis à jour (i.e. incluant les informations de la conférence, la licence de diffusion CC-BY, et la pagination).

Une fois les actes au format PDF compilés, il est possible de procéder à une nouvelle compilation dans un autre format attendu par une archive ouverte cible (les archives et formats de sortie supportés sont décrits en Annexe page 15). Pour cela, il suffit d'appeler à nouveau l'outil `taln2x` en changeant les options de compilation [26].

20. Dans certains cas, la personne responsable des actes peut demander aux auteur·e·s les sources LaTeX de leur article.
21. Les volumes peuvent par ailleurs être découpés en parties (e.g. pour inclure des sessions thématiques).
22. https://yaml.org/spec/1.2.2/
23. Un fichier exemple est fourni par l'outil `taln2x`.
24. Pour plus d'informations, voir https://talnarchives.gitlabpages.inria.fr/taln2x/formats/.
25. Ce modèle est inspiré par celui utilisé pour les actes des conférences de l'ACL (https://github.com/acl-org/easy2acl/tree/master/book-proceedings).
26. Les options de compilation sont décrites à https://talnarchives.gitlabpages.inria.fr/taln2x/options/.

2.3 Interventions humaines

La compilation des actes nécessite certaines interventions humaines :

— pour vérifier que les métadonnées saisies par les auteur·e·s sont correctes et cohérentes par rapport à la version finale de l'article [27] ;

— pour vérifier que les actes compilés (en PDF ou dans les autres formats attendus) sont bien formés et cohérents, et sinon modifier les options de configuration et / ou métadonnées erronées et relancer la compilation ;

— pour téléverser les actes dans les différentes archives ouvertes [28].

Historiquement, ces interventions reposaient sur un très petit nombre de personnes (en l'occurrence la personne en charge de la maintenance de l'outil `taln2x` et la personne en charge des actes au sein du comité d'organisation de la conférence).

Afin de pérenniser la diffusion des actes de TALN / RECITAL sur les archives ouvertes, l'ATALA a défini en 2023 des personnes ressources (responsable des actes au sein du comité d'organisation de la conférence, responsable des actes des conférences TALN au sein du CA de l'ATALA). Par ailleurs, l'installation de l'outil de compilation des actes `taln2x` a été grandement simplifiée [29], et sa documentation largement étendue [30]. À terme, ces mesures devraient rendre le procédé de diffusion des actes le plus fluide possible, avec un minimum d'interventions humaines.

Actuellement, l'étape la plus chronophage concerne la vérification (i) des métadonnées saisies dans le système de gestion des soumissions (qui ne sont parfois pas cohérentes par rapport à la version finale de l'article), et (ii) des fichiers PDF des versions finales (qui contiennent parfois des erreurs de formattage). Une meilleure communication sur ces points auprès de la communauté (via la liste de diffusion de l'ATALA et dans les appels à soumissions) devrait aider à obtenir des saisies de meilleure qualité.

27. Il arrive parfois que les métadonnées ne soient pas à jour, par exemple si le titre de l'article a changé entre la version soumise et la version finale, ou encore si les auteur·e·s n'ont pas mis à jour leur profile sur la plateforme de gestion des soumissions.

28. Le téléversement passe généralement par la mise à disposition de fichiers compressés via un outil de type ftp, filesender ou autre, sauf dans le cas de HAL où le téléversement se fait via un outil dédié (webservice ou application `X2hal` accessible à l'adresse `https://x2hal.inria.fr`).

29. Cet outil s'installe à présent simplement via `pip`, le gestionnaire de paquets python, qui prend en charge l'installation des dépendences de l'outil.

30. Voir `https://talnarchives.gitlabpages.inria.fr/taln2x`.

3 Vue d'ensemble

3.1 Déroulement des opérations

La Figure 2 ci-dessous donne une vue d'ensemble du procédé de compilation et diffusion des actes de la conférence TALN / RECITAL sur les archives ouvertes (et des personnes impliquées).

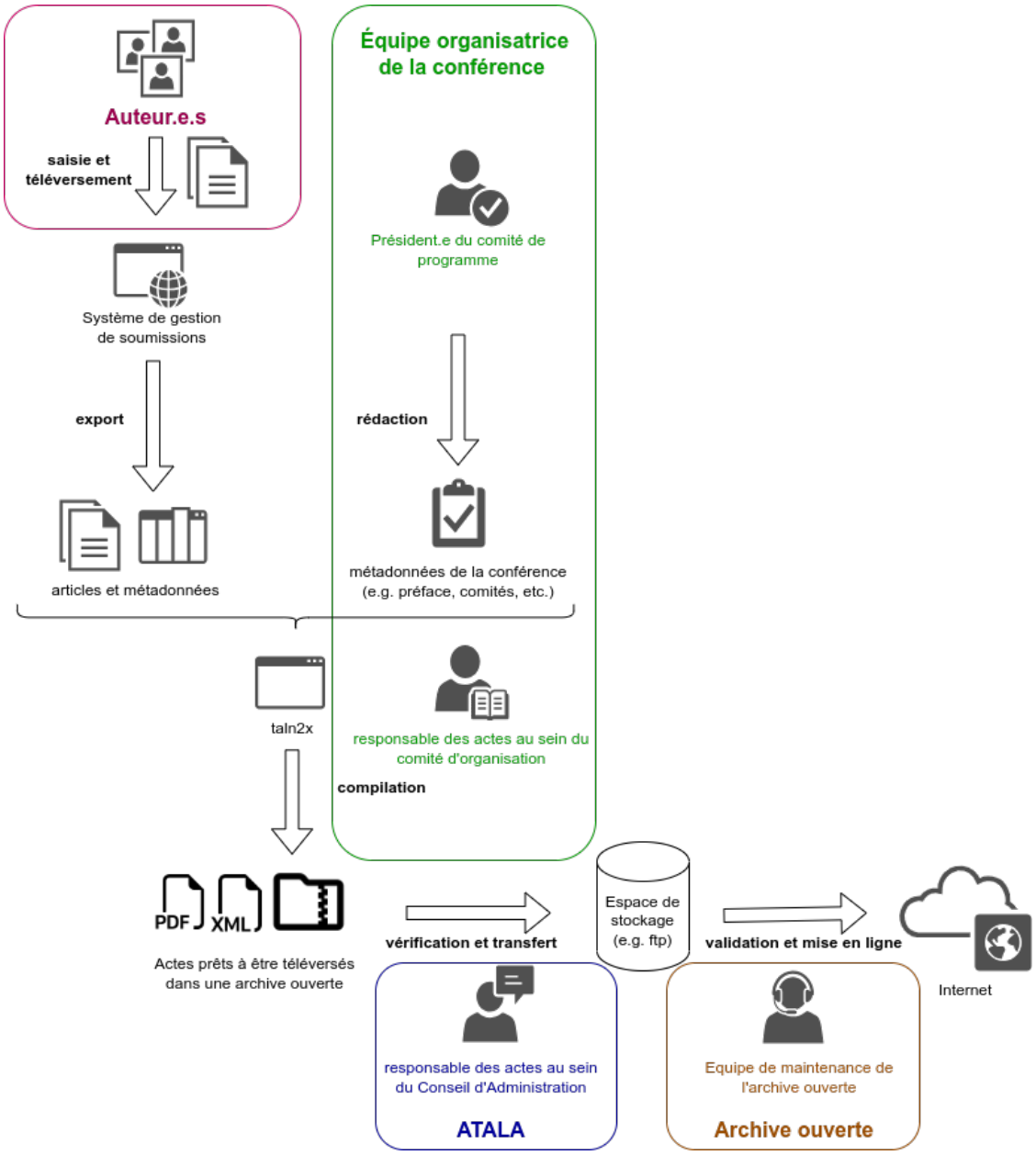

Figure 2 – Procédé de compilation et diffusion des actes.

3.2 Points de vigilance

Les expériences passées (depuis 2020) de compilation et diffusion automatique des actes de TALN / RECITAL ont permis de mettre en lumière certains points de vigilance :

— la vérification des métadonnées et des fichiers PDF des articles saisis par les auteur·e·s ne doit pas être négligée, car il reste souvent des erreurs (ordre des auteur·e·s, version finale du titre, etc. pour ce qui concerne les métadonnées, numéros de page non retirés, utilisation de paquets LaTeX particuliers impactant la fonte, etc. pour ce qui concerne les fichiers PDF), il est important de définir le calendrier de la conférence en conséquence (i.e. laisser du temps entre la date limite de dépôt de la version finale par les auteur·e·s et la date limite pour la mise en ligne des actes) [31] ;

— l'ingestion des actes dans une archive ouverte est un processus qui prend du temps (il est souvent assujetti a minima à une vérification manuelle de forme), là aussi il convient de définir le calendrier de la conférence en conséquence, notamment si l'on souhaite que les actes soient en ligne au moment de la conférence [32] ;

— la diffusion des articles retenus sur les archives ouvertes n'est possible que si les droits en ce sens ont été obtenus par l'ATALA, soit par la signature d'une autorisation (formulaire dit de cession des droits), soit par l'utilisation d'une licence ouverte (e.g. licence creative commons) annoncée lors de l'appel à soumission [33] ;

— dans le cas spécifique de l'archive ouverte HAL, la création d'une collection permet de mettre en place une indexation particulière et d'offrir un accès direct aux publications de l'ATALA sur HAL, il faut en faire la demande de manière anticipée par la personne responsable des actes de la conférence [34].

31. Une liste non exhaustive des erreurs de formattage rencontrées en 2020 est donnée en Annexe page 17.

32. Dans le cas de HAL, il est utile de prendre contact avec le support pour prévenir qu'il va y avoir un dépôt automatique de plusieurs dizaines d'articles.

33. Cette dernière option est actuellement utilisée par l'ATALA pour les conférences TALN / RECITAL.

34. Voir https://doc.archives-ouvertes.fr/gerer-une-collection/. À noter, le tamponnage des articles de la collection se fait de préférence lors du dépôt par `taln2x` (il peut cependant se faire a posteriori au moyen de mots-clés).

4 Axes de travail

Le processus de compilation des actes de TALN / RECITAL est à présent éprouvé (il est utilisé depuis plusieurs années pour diffuser les actes de la conférence sur des archives ouvertes). Les actions entreprises récemment par l'ATALA (e.g. nomination d'une personne responsable des actes au sein du CA) devraient par ailleurs permettre de rendre ce processus encore plus fluide, dans la mesure où les tâches sont à présent partagées entre plusieurs personnes ressources clairement identifiées.

Les axes de travail envisagés concernent plus particulièrement l'extension de l'outil de compilation des actes pour supporter d'autres systèmes de gestion de soumissions, et le développement de bonnes pratiques pour le dépôt des actes sur l'archive ouverte HAL.

Sur ce dernier point, il a été constaté dernièrement des interrogations de la part des auteur·e·s sur leur implication dans le processus de mise en ligne de leurs articles sur HAL. En effet, un usage répandu de HAL consiste en le dépôt d'articles par les auteur·e·s (ces dernier·e·s gardent ainsi la propriété sur les données saisies sur HAL et la possibilité de modifier celles-ci directement). Au niveau de l'ATALA, il a été décidé de procéder à un téléversement automatique des actes de TALN / RECITAL sur HAL afin (i) d'offrir une plus grande exhaustivité des articles déposés, (ii) une plus grande homogénéité entre les différents articles déposés (e.g. utilisant tous le même titre pour les actes), et (iii) d'avoir des références plus complètes (e.g. incluant les numéros de pages). Ce téléversement automatique s'accompagne de nouveaux besoins et soulève quelques questions :

— comment permettre facilement aux auteur·e·s d'obtenir la propriété de leurs articles déposés par l'ATALA sur HAL [35] ?

— comment garantir la correction des métadonnées des articles déposés sur HAL (notamment les affiliations des auteur·e·s sachant que certain·e·s auteur·e·s n'ont pas forcément d'affiliation existante sur HAL, e.g. des personnes issues du monde industriel) ?

— comment gérer les doublons en cas de soumissions par l'ATALA et également par les auteur·e·s (e.g. en cas de mise en ligne d'une version de type pre-print) [36] ?

— comment étendre la hiérarchie des collections pour créer une collection par volume (i.e. une collection TALN, une collection RECITAL, une collection par atelier) ?

— comment gérer les collections et sous-collections HAL de la conférence et de ses ateliers (e.g. pour y intégrer les articles déjà déposés) ?

— comment étendre la couverture des actes de TALN / RECITAL dans HAL (i.e. comment y intégrer des articles des éditions passées, en tenant compte des doublons) ?

35. À noter, HAL s'est doté récemment d'une fonctionnalité permettant d'activer automatiquement lors du dépôt d'un article, un partage de propriété à tou·te·s ses auteur·e·s disposant d'un identifiant HAL. Par ailleurs, chaque auteur·e d'un article déposé par un tiers (automatiquement ou non) sur HAL peut demander le partage de propriété de l'article depuis l'interface web de HAL.

36. Cette question est liée à la procédure de fusion d'articles dans HAL, et notamment à la mise à jour des métadonnées (e.g. comment agréger le nombre de vues des articles fusionnés ?).

— comment permettre la mise à jour des fichiers PDF d'articles ayant été déposés dans HAL par un tiers (e.g. pour remplacer une version préliminaire par la version publiée) ?

— comment faciliter la modération des fichiers PDF déposés (e.g. gérer les PDF non conformes aux règles de HAL [37]) ?

37. Voir https://doc.archives-ouvertes.fr/reussir-mon-depot-guide-pratique/.

5 Références

Baruch, Pierre (oct. 2007). “Open Access Developments in France : the HAL Open Archives System”. In : Learned Publishing 20. see also ; voir aussi : P. Baruch, La diffusion libre du savoir Accès libre et Archives ouvertes, http://archivesic.ccsd.cnrs.fr/sic_00169330/fr/, p. 267-282. doi : 10.1087/095315107X239636. url : https://hal.science/hal-00176428.

Bird, Steven et al. (mai 2008). “The ACL Anthology Reference Corpus : A Reference Dataset for Bibliographic Research in Computational Linguistics”. In : Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC’08). Marrakech, Morocco : European Language Resources Association (ELRA). url : http://www.lrec-conf.org/proceedings/lrec2008/pdf/445_paper.pdf.

Boudin, Florian (juin 2013). “TALN Archives : une archive numérique francophone des articles de recherche en Traitement Automatique de la Langue”. In : Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles. Les Sables d’Olonne, France : Association pour le Traitement Automatique des Langues, p. 507-514. url : http://talnarchives.atala.org/TALN/TALN-2013/taln-2013-court-001.

Ley, Michael (2002). “The DBLP Computer Science Bibliography : Evolution, Research Issues, Perspectives”. In : String Processing and Information Retrieval. Sous la dir. d’Alberto H. F. Laender et Arlindo L. Oliveira. Berlin, Heidelberg : Springer Berlin Heidelberg, p. 1-10. isbn : 978-3-540-45735-0.

6 Annexe

6.1 Formats des archives ouvertes supportées par `taln2x`

À ce jour, l'outil utilisé pour la compilation des actes en vue d'une intégration à une archive ouverte supporte les formats d'archives ouvertes suivants : TALN archives, HAL, ACL Anthology, DBLP. Ces formats sont détaillés ci-dessous.

6.1.1 Actes au format ACL Anthology

Les actes au format ACL Anthology correspondent, pour chaque volume, à un dossier contenant : [38]

— un fichier texte nommé `meta` contenant les métadonnées du volume (éditeurs, titre, etc.) ;

— un dossier nommé `cdrom` contenant à sa racine un fichier PDF correspondant au frontmatter de la conférence, un fichier bibtex pour les actes complets, un dossier nommé `pdf` contenant chacun des articles au format PDF, et un dossier nommé `bib` contenant la référence bibtex de chacun des articles.

Pour intégrer les actes à l'ACL Anthology, une fois les différents volumes compilés, il faut les regrouper dans une archive (e.g. ZIP) qui sera mise en ligne et dont le lien sera envoyé à la personne en charge de l'Anthology.

6.1.2 Actes au format TALN-archives

Les actes au format TALN-archives correspondent, pour chaque volume, à un dossier contenant :

— un fichier XML avec les métadonnées du volume et celles de chacun des articles ; [39]

— un fichier bibtex contenant ces mêmes métadonnées ;

— un dossier nommé `actes` contenant les PDF de chacun des articles ;

— un dossier nommé `bib` contenant les références bibtex de chacun des articles.

Pour intégrer les actes aux TALN archives, il faut compresser le dossier de chaque volume, puis faire parvenir ces archives à la personne en charge de la maintenance des TALN archives.

À noter : afin de pouvoir rétro-ingérer les actes des éditions passées de TALN dans l'ACL Anthology, le format TALN-archives a été utilisé (i.e. `taln2x` peut prendre en entrée un dossier de type TALN-archive et en extraire les informations sur les actes pour les compiler vers un autre format).

6.1.3 Actes au format HAL

La compilation des actes au format HAL génèrent des données dans deux formats supportés par HAL :

38. Pour plus d'informations, voir https://acl-org.github.io/ACLPUB/anthology.html
39. Voir https://github.com/boudinfl/taln-archives

1. le format utilisé par le webservice de HAL (désigné sous les termes d'import SWORD), correspondant à un fichier ZIP par article contenant le PDF de l'article et ses métadonnées au format XML ; [40]
2. le format bibtex utilisé l'application X2HAL. [41]

Suivant le format utilisé, le téléversement des actes dans HAL peut se faire soit :

— de manière non-interactive, en utilisant le format SWORD et les requêtes HTTP d'import contenues dans des scripts bash générés automatiquement par `taln2x` ;

— de manière interactive, en utilisant l'application `X2hal` [42] et en y téléversant le fichier bibtex généré par `taln2x`.

Dans le premier cas (mode non-interactif), il est possible d'activer diverses options du webservice HAL, comme par exemple la détection automatique des affiliations ou encore l'utilisation du dépôt HAL de test (environnement de pré-production).

Dans le second cas (mode interactif), la personne qui téléverse les actes bénéficie d'une interface graphique conviviale qui lui permet notamment de détecter les doublons et de vérifier les métadonnées. À noter, pour pouvoir déposer les fichiers PDF des articles avec X2hal, il faut que ces derniers soient accessible en ligne par ailleurs (par exemple sur les TALN-archives) et que leur adresse (URL) soit indiquée dans le fichier bibtex.

6.1.4 Actes au format DBLP

`taln2x` permet enfin de compiler les actes au format DBLP. [43] Concrètement, il s'agit d'un fichier XML contenant les métadonnées de la conférence et des articles de ses actes. Il est ensuite possible de soumettre ce fichier aux personnes en charge de DBLP pour que ces dernières intégrent les actes à la base de données.

40. Voir https://api.archives-ouvertes.fr/docs/sword/
41. Voir https://doc.archives-ouvertes.fr/x2hal/
42. https://X2hal.inria.fr
43. Voir https://dblp.org/faq/1474621.html

6.2 Erreurs rencontrées lors de la vérification des données soumises

Lors de l'édition 2020 de TALN / RECITAL, les erreurs de formattage suivantes avaient été rencontrées dans les articles déposés par les auteur·e·s :

- changement des dimensions de l'article (e.g. document au format letter plutôt qu'A4) ; [44]
- changement des polices (fonte, taille) ;
- erreurs dans les auteur·e·s (version finale toujours anonyme, ordre des auteur·e·s, etc.) ;
- accents manquant à "Résumé" et "Mots-clés" ;
- titre de l'article et titre des sections utilisant des majuscules superflues (en français, les règles sont les même que pour le texte courant) ;
- non-respect du nombre de pages ;
- présence de saut de lignes / paragraphes impromptus ;
- hyperliens non-cliquables ;
- bibliographie mal formattée :
  - champs doi manquant ou erroné (url complète au lieu de l'identifiant, champs non cliquable) ;
  - pour les publications de type CoRR, identifiant arXiv manquant ;
  - pour les articles de journaux, numéro de volume manquant ;
  - présence de références bibliographiques erronées (champs manquants), ou inutiles (références provenant des fichiers exemples).

44. Les dimensions peuvent être vérifiées au moyen d'un outil tel que `pdfinfo`.