Normes et patrimoine numérique

Laurent Romary

Normes et patrimoine numérique

Laurent Romary, équipe ALMAnaCH, Inria

Rôle des normes dans une démarche de numérisation du patrimoine

Les politiques institutionnelles liées à la gestion du patrimoine matériel et notamment documentaire ont été petit à petit transformées par la place grandissante qu'ont pris des projets de numérisation ambitieux portés notamment par les grandes institutions patrimoniales un peu partout dans le monde. En France, la Bibliothèque Nationale de France a très tôt montré le mouvement avec la démarche exemplaire portée par le projet Gallica[1] qui, partant d'une ouverture initiale de ses fonds en texte intégral, met maintenant à disposition de ses usagers tout un ensemble de sources numérisées, qu'il s'agisse de textes, documents sonores, vidéos ou même artefacts. Tout récemment encore, la mise en ligne de l'édition numérique des Testaments de Poilus[2], issus de fonds des archives nationales ainsi que des archives départementales des Yvelines et du Val-d'Oise, a montré que l'on pouvait adopter des démarches originales où la technologie vient compléter un travail de transcription distribué (*crowd-sourcing*) pour fournir *in fine* une base documentaire riche et précise utilisable à la fois par les chercheurs et le grand public.

Ce dernier projet est d'autant plus exemplaire qu'il offre, au-delà des versions images ou textes lisibles en ligne, la possibilité d'accéder à une représentation source encodée suivant les directives de la *Text Encoding Initiative*[3], la principale démarche de normalisation existante dans le domaine des données textuelles. On met le doigt ici sur un élément central de tout projet de numérisation en contexte patrimonial, à savoir la nécessité de dépasser la simple mise en ligne d'objets numériques que l'on peut consulter directement sur le site des projets concernés, pour créer un véritable patrimoine numérique ouvert et réutilisable pour toute une variété d'usages, notamment académiques. Dans ce cadre, la patrimonialisation

[1] https://gallica.bnf.fr

[2] https://edition-testaments-de-poilus.huma-num.fr

[3] https://tei-c.org

numérique ne pourra être effective que si l'on adopte, pour la représentation des données correspondantes, des formats pérennes dont l'utilisation ne dépendra pas de plates-formes logicielles ou matérielles particulières. C'est ici que les normes viennent jouer leur rôle de médiateur neutre entre émetteur et récepteur d'un processus technique.

Dans ce contexte, le présent chapitre a un double objectif. D'une part, il vise à donner une idée précise de la façon dont les normes en général sont définies, publiées et maintenues. D'autre part, par le biais de différents exemples de normes au cœur des projets de numérisation de fonds patrimoniaux, nous allons essayer d'ouvrir quelques débats qui animent les réflexions autour de l'usage des normes. Pourquoi une telle variété de normes, et d'organisations de normalisations ? Pourquoi existe t'il parfois des normes aussi différentes entre les projets portés par des archives ou des bibliothèques et les travaux des chercheurs, par exemple dans la création d'une édition numérique critique ? Quelle peut être à terme la bonne stratégie pour créer de véritables fonds numériques patrimoniaux pérennes ?

Comprendre les normes et le processus normatif

Le discours sur les normes n'est pas toujours à la hauteur des enjeux. Leur existence est souvent ignorée de ceux qui en auraient besoin. Des codes langues (*vide infra*) jusqu'à la représentation des dates, il n'est pas rare de découvrir qu'un développement informatique a conduit à une sur-créativité désastreuse en la matière.

Les normes sont aussi parfois raillées, notamment parce qu'elles seraient complexes ou parce qu'il y en aurait trop. On voit ainsi souvent citée la petite bande dessinée d'XKCD[4], parangon du *standards bashing*, qui trahit souvent une méconnaissance des processus de normalisation, ainsi que du rôle de complémentarité que peuvent jouer les normes entre-elles.

Je vais donc m'attacher ici à décrire les fondements de la normalisation pour que chacun puisse mieux en comprendre les attendus, mais aussi pour identifier comment intégrer ces éléments dans la perspective de leur application au domaine du patrimoine.

[4] https://xkcd.com/927/

La première chose à garder en tête est qu'une norme n'est pas un texte légal, c'est un instrument de contrat social qui permet à l'utilisateur d'un dispositif, principalement technique, d'en connaître un certain nombre de propriétés sur la base d'une déclaration de conformité à une certaine norme de la part du producteur de ce dispositif. Ainsi, si un logiciel signale sa conformité à la norme ISO 8601-1, l'acheteur de ce logiciel est en droit de s'attendre à ce que les formats des dates utilisés par celui-ci sont conformes à ladite norme, mais rien légalement n'oblige le vendeur à utiliser la norme. Au pire, l'acheteur pourra se retourner juridiquement contre le vendeur pour tromperie concernant les propriétés de son logiciel. Bien sûr, les normes peuvent ensuite être utilisées dans un cadre plus officiel, comme par exemple l'attribution d'un label, par exemple NF[5], ou même quand un organisme de normalisation travaille la main dans la main avec une organisation internationale qui publie des règlements reposant sur une famille de normes donnée[6].

Pour que le contrat social fonctionne, il est nécessaire qu'un minimum de confiance soit attaché à une norme de sorte qu'elle soit effectivement comprise et appliquée. Cette confiance peut passer par la reconnaissance qu'a l'organisme de normalisation où elle a été développée (nous reviendrons là-dessus), ou simplement parce qu'elle s'est déployée peu à peu avec succès au sein d'une communauté. Dans tous les cas, l'expérience prouve qu'il y a trois critères fondamentaux qui doivent être vérifiés pour qu'une norme s'ancre de façon solide et pérenne dans son domaine technique.

Tout d'abord, la norme doit être le résultat d'un travail de *consensus* au sein d'une communauté d'experts reflétant au mieux les fondements techniques couverts par la norme. Une norme ne peut être le fruit d'un travail individuel, ou un travail théorique déconnecté du monde qui l'entoure. Le consensus doit permettre de couvrir le mieux possible l'état de l'art, de prendre en compte les éventuelles variations techniques qui peuvent exister entre différentes sous-communautés, et surtout permettre à la norme d'avoir un cran d'avance sur les productions techniques existantes de sorte qu'elle soit susceptible de couvrir les inévitables évolutions que sa publication même va déclencher. En effet, la production d'une

[5] https://fr.wikipedia.org/wiki/Liste_de_normes_NF

[6] L'organisation maritime internationale travaille ainsi en collaboration avec différents comités de l'ISO pour définir ses réglementations (https://www.iso.org/fr/news/ref2227.html)

norme technique et son utilisation par une communauté large d'utilisateurs qui la mettent en œuvre conduit quasi automatiquement à la création d'un fonds de compétence vis-à-vis de la norme qui conduit les utilisateurs à dériver de nouveaux usages et de nouvelles constructions techniques. On l'observe ainsi depuis toujours au sein de la TEI, mentionnée en introduction, où dès qu'un mécanisme que l'on croit spécifique à un usage (par exemple l'attribut @status associé à un document pour indiquer son degré de validation) est repris très rapidement par d'autres utilisateurs pour en étendre l'usage (pour reprendre notre exemple, pour déterminer le degré de finalisation éditoriale d'une entrée de dictionnaire[7]).

La deuxième contrainte associée à une norme est que celle-ci soit effectivement disponible. Même si la chose peut paraître évidente, elle exclue par principe les « normes » industrielles ou communautaires dont la documentation n'est pas accessible à tous. Bien sûr, suivant le modèle économique associé à l'organisme de normalisation, les conditions d'accès libre ou payant, de citation ou de droits d'auteurs peuvent varier. Mais il doit exister un moyen simple d'accéder au document. Là encore, c'est un élément essentiel du contrat social associé à la norme puisque le concepteur d'un objet technique, comme l'utilisateur de celui-ci doivent tous deux pouvoir vérifier la conformité au cadre normatif. Si l'on creuse plus finement ce critère, il est bien sûr possible d'envisager différents niveaux d'exigence qui peuvent avoir un impact sur l'utilisabilité de la norme. Ainsi, un simple document PDF — comme fournit par l'ISO — ne permet pas facilement d'effectuer des tests de conformité techniques sur un format de données par rapport à une norme qui serait associée à un schéma de validation XML[8] comme c'est le cas pour l'EAD[9] ou la TEI.

Enfin, un élément essentiel du cycle de vie d'une norme est sa *maintenance*. Les contextes techniques et les connaissances évoluant, notamment comme on l'a vu de l'usage des

---

[7] Le lecteur intéressé pourra parcourir le fil de discussion correspondant sur la liste d'échange de la communauté des utilisateurs de la TEI : https://listserv.brown.edu/cgi-bin/wa?A1=ind2007&L=TEI-L&X=O32C5BE40517CE26FAC#7

[8] eXtended Markup Language ; Métalangage de représentation de données textuelles structurée (https://www.w3.org/XML/)

[9] Encoded Archival Description : norme pour la description de collections d'archives (cf. https://www.bnf.fr/fr/ead-encoded-archival-description)

normes elles-mêmes, une norme doit pouvoir s'adapter en intégrant dès sa conception un processus de mise à jour. Il peut s'agir d'un processus d'évolution en continue, comme c'est le cas pour la TEI ou les codes langue de la série ISO 639, ou un processus déclenché à intervalle régulier, comme c'est le cas pour l'ISO par défaut.

Si un tel processus de maintenance n'est pas à l'œuvre, la norme risque de devenir rapidement obsolète ou pire, de subir des évolutions non contrôlées issues de communautés particulières d'utilisateurs qui ne se sont pas vu offrir un espace de contribution à sa révision. C'est exactement ce que nous avons pu observer au fil des années pour les normes archivistiques[10], notamment EAD, qui bien que développées dans le cadre d'une association professionnelle, la société des archivistes américains, n'ont jamais été véritablement gérées dans le cadre d'une structure spécifique assurant son développement et sa maintenance. Au bilan, on assiste maintenant à l'émergence de travaux directement concurrentiel avec la nouvelle initiative *Records in Context*, gérée par le conseil international des archives (*International Council on Archives*, ICA) qui risquent de contribuer à fragmenter un peu plus le paysage existant.

On perçoit sur ce dernier exemple le rôle essentiel que joue le cadre organisationnel lié au travail de normalisation et en particulier pourquoi la plupart des standards sont de fait développé au sein de ce qu'on appelle une organisation de développement de normes (*SDO – Standards Development Organization*). Une telle organisation est en général indispensable pour garantir la mise en œuvre des trois piliers de la normalisation (consensus, diffusion, maintenance) évoqués ci-dessus, et ce en assurant l'organisation du travail de normalisation autour de trois fonctions principales :

- La gestion de la participation d'experts à la définition des normes ;
- L'organisation des étapes du cycle de vie du document, de la proposition initiale à la publication finale ;
- Le recueil de l'acquiescement de ses membres en fonction de sa structure et de son modèle économique.

[10] Le bilan historique dressé par les acteurs eux-mêmes est édifiant : https://www.loc.gov/ead/eaddev.html

Le dernier point est essentiel car la normalisation coute cher et nécessite qu'une organisation de normalisation soit viable sur le long terme de façon à garantir la pérennité des travaux correspondants. Ainsi derrière toute organisation de normalisation il y a un consortium de membre, individuels ou institutionnels, publics ou privés qui contribue au financement de l'activité, en complément de l'éventuel vente des normes produites.

Quel que soit l'organisme de normalisation concerné, il est une distinction qu'il est utile de garder en tête lorsqu'on parle de normalisation, à savoir la différence entre normes horizontales et verticales. Une norme horizontale va couvrir un large spectre de domaines applicatifs et, dans le cas du domaine numérique, va pouvoir être considéré comme une brique indispensable pour que tout système d'échange de données puisse fonctionner. C'est typiquement le cas de normes génériques telles que l'ISO 10646-Unicode ou les codes de langue (la série ISO 639). A l'inverse, une norme verticale est dédiée à un domaine particulier en tentant de couvrir précisément un besoin technique d'une communauté. EAD en est un bon exemple.

Pour les domaines liés au patrimoine qui sont l'objet du présent ouvrage, il faut prendre conscience de l'existence d'une opposition supplémentaires entre normes institutionnelles, privilégiées par les établissements patrimoniaux d'un côté, et les normes qui vont plutôt accompagner le travail académique des chercheurs s'appuyant sur les contenus numérisés. Pour simplifier, au risque d'être caricatural, les premières sont en général extrêmement conservatrices alors que les secondes seront en général plus proches des derniers développements technologiques. Les raisons sont faciles à analyser : le déploiement en largeur de projet de numérisation au sein des établissements patrimoniaux coute cher et la nécessaire pérennisation des résultats implique que les cadres normatifs restent particulièrement stables.

A l'opposé, les communautés de chercheurs, dispensées de travailler en quantité et en largeur sur les corpus disponibles, peuvent avoir une exigence de granularité plus fine et surtout vont plus être tenté de voir les normes qu'ils utilisent évoluer rapidement afin d'intégrer au mieux les derniers concepts issus de leurs recherches. Cette dynamique peut avoir un prix, à savoir l'émergence trop rapide de cadre normatif concurrentiels qui risquent

de remettre en cause la pérennité des contenus numériques ainsi représentés[11]. Pour prendre un exemple concret, on a vu comment la communauté épigraphique a développé sa propre variante des directives de la TEI (Epidoc[12]), pour qu'ensuite il faille des années de travail de réintégration dans le cadre de référence.

Cette tension entre normes des institutions patrimoniales et normes des chercheurs est un réel problème pour que ces deux communautés puissent facilement échanger leurs contenus numériques et donc ce faisant leurs cadres conceptuels. On voit ainsi une tension constante entre les descriptions bibliographiques fines exigées par la création d'éditions scientifiques associées à des manuscrits anciens, qui seront quasi exclusivement décrite dans le cadre des directives de la TEI et les informations plus génériques obtenues au sein des instruments de recherche en EAD fournis par les archives hébergeant les sources correspondantes.

De façon presque plus anecdotique, on peut mentionner ici combien le déploiement de normes liées à la numérisation des contenus dans les établissements patrimoniaux peut se heurter à des réticences liées à l'ouverture même de ces contenus qui risque d'en résulter. Tout d'abord, la mise en ligne des contenus nécessite de porter une attention plus grande aux contraintes de diffusion et de réutilisation des fonds existants, qu'il s'agisse d'aspects liés aux droits d'auteurs ou à la protection de données personnelles (dans le cas par exemple de fonds récents d'archives personnelles). Bien que la directive européenne PSI, et son implémentation dans les lois Valter et LPR cadrent relativement bien l'ouverture par défait des contenus, de nombreux établissements, notamment les plus petits, n'ont pas les moyens d'accompagner le déploiement de fonds numérique en ligne d'une vérification des contraintes légales qui s'imposent éventuellement. Pour les mêmes raisons, il y a souvent une réticence supplémentaire qui s'exprime de la part de certaines archives d'ouvrir leurs fonds et que nous avons rencontré au cours du projet Cendari portant sur les archives Européennes liées à la première guerre mondiale. Certaines institutions ont en effet peur que l'ouverture numérique, favorisée par le déploiement de normes telles qu'EAD, n'attire

[11] En novlangue scientifique, on y verrait une remise en cause des principes FAIR (https://www.ouvrirlascience.fr/fair-principles/).

[12] https://epidoc.stoa.org/gl/latest/intro-eps-fr.html

des communautés d'utilisateurs plus nombreuses pour lesquels elles n'auraient pas les personnels d'accueil suffisants pour répondre aux demandes de consultation des fonds.

Les normes pour le patrimoine dans la pratique

Pour sortir quelque peu d'un cadre qui pourrait paraître un peu trop théorétique, nous souhaitons montrer, au travers d'exemples concrets, comment les normes se créent et évoluent en lien avec les activités de numérisation et de recherche sur le patrimoine. La sélection ci-dessous est forcément biaisée : d'une part, je n'ai choisi que des normes que j'ai pratiquées, à différents niveaux de précision, et d'autre part, le choix vise explicitement à montrer la variété des communautés et donc des besoins quand il s'agit de représenter des contenus patrimoniaux numérisés.

Le mille-feuille du codage des langues

Le meilleur exemple par lequel nous pouvons commencer est bien sûr le codage des langues. Il s'agit en effet d'un domaine central pour la gestion du patrimoine numérique à caractère linguistique (écrit, oral), mais aussi pour tout ce qui touche au catalogage et à la documentation de n'importe quel objet patrimonial physique ou numérique. Le domaine du codage des langues est aussi l'une des activités normatives ayant eu le plus d'impact en technologies de l'information, probablement juste derrière la norme ISO/IEC 10646[13]. Enfin, l'ensemble du mille-feuille normatif autour du codage des langues illustre parfaitement les rôles complémentaires que peuvent jouer certaines normes et leurs agences de normalisation respectives pour offrir un cadre solide et évolutif de représentation d'objets numériques.

L'ensemble de l'édifice normatif relatif au codage des langues repose sur la série de normes ISO 639 développée conjointement par les comités techniques 37 (Langue et terminologie) et 46 (Information et documentation) de l'ISO. Son histoire, qui remonte à une première norme publiée en 1967 (ISO 639/R) ainsi que sa nature pluridisciplinaire à la croisée entre technologies de l'information, sciences des bibliothèques et linguistique se retrouve dans la

[13] Que l'on connaît surtout par son déploiement industriel dans le cadre du consortium Unicode (https://home.unicode.org)

structure actuelle de la série normative en fait maintenant une norme extrêmement solide et utile dans tous les milieux allant du patrimoine aux industries de l'information.

A un premier niveau, la norme ISO 639-1 introduit des codes à deux lettres pour un noyau de 210 langues correspondant aux priorités du monde des bibliothèques, qui fût à l'origine du projet. Les codes sont accompagnés de noms en anglais, en français et dans la langue elle-même (autoglossonymes). Ces codes sont de fait figés ;

Les parties 2 et 3 de la série ISO 639 définissent des codes à trois lettres. La partie 2 définit une sous-série de 500 langues[14] auxquelles sont parfois (pour 25 d'entre elles) associés 2 codes, afin d'assurer une compatibilité avec les pratiques de la bibliothèque du congrès américain. C'est d'ailleurs cette même bibliothèque du congrès (Library of Congress) qui gère l'autorité d'enregistrement correspondante.

La partie 3 de l'ISO 639 définie elle aussi des codes à trois lettres mais visent à une couverture la plus exhaustive possible des langues vivantes ou anciennes. Son autorité d'enregistrement est le SIL (*Summer Institute of Linguistics*), une agence confessionnelle de documentation linguistique, qui a apporté un fond initial descriptif important avec son registre Ethnologue[15].

Deux parties viennent compléter ce portefeuille : la partie 5 qui couvre les familles de langues et la partie 4, en cours de révision qui agrège les principes généraux concernant l'identification, le nommage et la codification des langues.

Au-delà de la richesse descriptive de la série de normes ISO 639, l'essentiel de leurs utilisations se fait par combinaison et intégration dans un mille-feuille de normes particulièrement efficace à couvrir une large gamme d'usages de qualifications linguistiques. Tout d'abord, au sein même de l'ISO différentes normes complètent l'ISO 639 dans les domaines du codage des noms d'écriture (arabe, cyrillique, latine, syriaque etc.) avec l'ISO

[14] L'expression qui me vient est « ayant pignon sur rue » : il s'agit de langues pour lesquelles il existe au moins 50 documents dans une agences officielles, et qui sont confirment à une série de critère prouvant leur rôle officiel (étendue de la littérature correspondante, soutien d'un état ou d'une région, existence d'une agence de normalisation, éducation, etc.).

[15] https://www.ethnologue.com

15924 ou des codes de noms de pays (série ISO 3166), pour localiser les langues référencées par exemple.

Pour un nouveau venu, ces différentes normes offrent une combinatoire assez complexe, mais qui est compensé, comme illustré figure 1, par le travail d'autres institutions normatives qui créent des cadres applicatifs de plus en plus précis. Ainsi, l'IETF[16] propose un cadre précis et univoque [BCP 47, Phillips et Davis, 2009] pour coder une langue particulière en combinant les normes ISO disponibles, mais aussi en offrant la possibilité de faire référence à des normes ou variations dialectales qui n'ont pas été normalisées. L'IETF maintiens en particulier un registre de codes qui sert de référence pour tout usage en technologies de l'information et de la communication. Ce registre, couplé à la norme BCP 47 est devenu la base normative pour la représentation de données linguistique au sein du W3C, notamment pour ses recommandations XHTML ou XML. Enfin, le consortium TEI a décidé de s'appuyer entièrement sur les recommandations du W3C en la matière et a ainsi repris l'attribut XML @xml:lang pour qualifier tout contenu textuel d'un document encodé suivant ses directives.

Ce que montre l'exemple du codage des langues, c'est que pour un domaine aussi « simple », il est nécessaire de mobiliser tout un ensemble de briques normatives issues d'organisations de normalisation aux modes de fonctionnement eux-mêmes très différents, et qu'aborder le sujet nécessite en amont une réelle éducation en la matière. C'est encore plus vrai quand on se retrouve impliqué dans une activité de normalisation verticale et qu'il faudra s'assurer que l'on n'y réinvente pas le savoir technique correspondant.

Cette stratification technologique peut aussi avoir ses dangers quand il s'agit de planifier des changements stratégiques dans un domaine particulier. L'évolution de la représentation des données vers le web sémantique et les données liées implique que l'on dispose d'un référencement des codes langue sous la forme d'URI cohérentes et pérennes. A quel niveau une telle évolution doit-elle s'opérer et comment les différents acteurs peuvent ils se coordonner ?

[16] Internet Engineering Task Force (https://www.ietf.org)

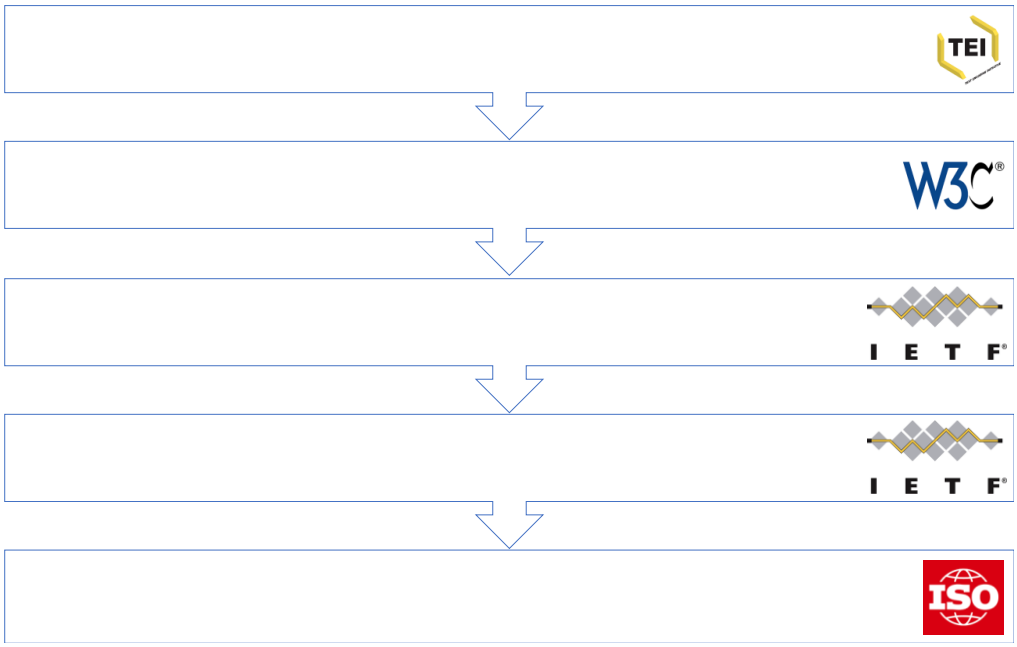

Figure 1 : le mille-feuille normatif du codage des langues

La TEI – une réussite à tous les étages

Alors que la série ISO 639 est devenu une norme horizontale incontournable au niveau international, on peut qualifier la *Text Encoding Initiative* de plus beau succès de normalisation verticale. Initiée il y a un peu plus de 30 ans avec comme objectif de fournir un cadre de représentation commun à plusieurs projets de bases textuelles qui se mettaient en place dans les années 80, elle est devenue le cadre de référence incontournable pour tout projet de numérisation de données textuelles, dès qu'il comporte une dimension d'interopérabilité ou de préservation (Romary, 2009).

La TEI repose sur une infrastructure solide articulée autour de l'usage systématique de la recommandation XML pour la sérialisation de ses modèles de données, associé à un vocabulaire extrêmement riche couvrant toute une palette de formes textuelles (prose, théâtre, dictionnaires, transcription de données orales), mais aussi son propre langage de spécification[17]. C'est dans ce langage, sous-ensemble des directives de la TEI, qu'est décrit l'ensemble de la documentation et des spécifications de la TEI suivant le principe de la programmation lettrée issue des travaux de D. Knuth (1984).

[17] Appelé ODD, pour *One Document Does it all* et documenté dans un chapitre spécifique de la TEI (cf. https://tei-c.org/release/doc/tei-p5-doc/de/html/TD.html)

Ce même langage de spécification est disponible à chaque utilisateur ou projet pour documenter la façon dont la TEI a été utilisée dans un contexte donné. On peut ainsi décrire le sous-ensemble des éléments XML utilisés, contraindre les valeurs d'un élément ou d'un attribut voire même ajouter des objets particuliers au vocabulaire de la TEI. Ce dispositif donne des outils aux utilisateurs de la TEI qui permettent de garantir un peu plus la réutilisabilité à long terme des contenus ainsi produits.

Mais la grande force de la TEI est avant tout son modèle de standardisation ouvert qui permet de couvrir le plus étroitement possible les besoins de la communauté des utilisateurs. Tout d'abord, l'ensemble des productions de la TEI, principalement la spécification technique et les documentations qui l'accompagne, sont librement accessibles en ligne sous une licence double CC-BY[18] et BSD 2-Clauses[19], qui en offre la réutilisation maximale. Le travail de maintenance de la TEI s'appuie sur un comité technique d'experts élus par la communauté (les membres individuels ou institutionnels du consortium) et surtout opérant à partir des propositions faites par les utilisateurs sur le serveur GitHub du consortium TEI[20].

Par ailleurs, la TEI intègre dans ses fondements même une dimension patrimoniale qui s'exprime à différents niveaux :

- Tout document TEI intègre de façon obligatoire un entête contenant les métadonnées associées au contenu (le *teiHeader*) et qui font du document un objet numériquement autonome, y compris du point de vue de son archivage ;
- L'entête TEI intègre l'ensemble des étapes liant le document à sa source, la création de l'objet numérique, les versions et ses conditions de publication (licence etc.) ;
- La structure du document TEI intègre les mécanismes nécessaires pour faire référence, quand c'est approprié, aux images sources (facsimilés).

De fait, la TEI développe une vision de l'objet numérique qui à la fois le relie fortement à sa source, qu'elle soit physique ou même une version antérieure d'un contenu que l'on a par la

[18] https://creativecommons.org/licenses/by/4.0/

[19] https://opensource.org/licenses/BSD-2-Clause

[20] Sans entrer dans les détails, je signale au passage que l'ensemble des données et des services de la TEI sont hébergés au sein de l'infrastructure Huma-Num du CNRS.

suite enrichi, et fait du document lui-même un objet numérique patrimoniale qui peut directement s'intégrer dans un fonds plus complet. Cette démarche est ainsi à l'origine de différentes initiatives de bibliothèques numériques de référence, comme les BVH[21] par exemple.

Au total, bien que la TEI ait connu un succès centré sur des projets principalement académique d'édition numérique de sources, on peut observer un intérêt croissant des institutions patrimoniales de maîtriser cette norme pour diffuser ses propres contenus, notamment dans le cadre de la mise en ligne de large fonds de texte comme à la BNF, la Bibliothèque Royale des Pays bas, ou même l'administration en charge des documents historiques des États-Unis d'Amérique (Wicentowski, 2011). De façon plus prospective, la TEI tient la corde pour devenir un format de référence quand l'essentiel de ces institutions basculeront d'une politique numérique de métadonnées, et parfois d'images, à la publication intégrale des contenus en texte intégral. La puissance et la flexibilité offert par le cadre normatif de la TEI sera une base idéale pour développer des contenus qui, tout en reposant sur des démarches éditoriales affirmées par les institutions elles-mêmes, leur permettra de rester interopérable avec leurs consœurs.

EAD – un processus laborieux de normalisation

A l'opposé de la *success story* que représente la TEI, nous avons déjà fait allusion à une autre norme importante pour le patrimoine, vu dans sa dimension numérique, et qui a connu un développement un peu moins harmonieux : EAD (*Encoded Archival Description*).

La conception d'EAD résulte de la nécessité pour les établissements archivistique de représenter leurs instruments de recherche, c'est à dire le catalogue des collections, jusque-là gérés sous forme papier. De tels catalogues intègre à la fois des informations sur l'histoire des fonds, leurs contenus, leur description physique et bien sûr leur localisation au sein de l'archive elle-même. Leur rôle d'interface entre les fonds de l'archive et leur appropriation par des utilisateurs extérieurs est donc essentiel et l'on comprend alors comment tôt des communautés de chercheurs se sont approprié l'EAD au fur et à mesure que la norme se déployait dans les institutions patrimoniales.

[21] http://www.bvh.univ-tours.fr

Au départ, il est intéressant de constater que les travaux autour d'EAD ont été fortement influencés par ceux de la TEI, en particulier de par le rôle moteur de Daniel Pitti (Pitti, 1997) membre des deux communautés du texte et des archives. Le format EAD intègre ainsi très tôt une structure intégrant un entête reprenant plusieurs caractéristiques de l'entête TEI et permettant d'intégrer des métadonnées caractérisant l'instrument de recherche numérique. Bien plus, la charte de développement d'EAD[22], telle que définit par le groupe d'experts dès 1995, exprime la contrainte de reprendre les éléments pertinents de la TEI au sein d'EAD dès qu'il s'agit de représenter le même type de contenu.

Mais contrairement à la TEI l'EAD n'a pas connu de vrai processus de normalisation géré dans le temps. Porté par un groupe d'experts enthousiaste au sein de la société des archivistes américains dès 1993, les développements successifs à partir de la V1 officielle publiée en août 1998 se sont fait au hasard des financements et des volontaires pour porter les évolutions, comme le traduise la variété des dénomination (EAD 2002, EAD3). La différence la plus notable avec le maintien en continue des directives de la TEI est le fait qu'EAD a subi un développement par strate, privilégiant ainsi les refontes en profondeur plutôt que les corrections à la marge. Cette tendance s'est trouvée même exacerbée avec le changement de cap complet introduit pat l'International Council of Archives se lançant dans l'aventure du linked open data archivistique avec Ric-O dont le lien avec les implémentations existantes d'EAD est pour le moins ténu.

A moins que les organisations archivistiques ne mettent en place une vraie organisation de leur activité normative, intégrant compétence technique et masse critique d'expertise, ainsi qu'une vraie architecture de spécification de modèle de données (qui pourrait s'inspirer de la TEI, comme on l'a montré dans (Romary et Riondet, 2018), il y a un vrai risque de déboucher sur une situation de blocage entre les implémentations existantes et des propositions trop expérimentales pour être utilisables concrètement.

Un exemple de petite norme qui facilite la vie : IIIF

Pour finir, nous pouvons mentionner la norme IIIF qui illustre bien le rôle de « petites » normes de traverse qui peuvent profondément modifier le paysage technique de

[22] https://www.loc.gov/ead/eaddev.html

la diffusion de données patrimoniale. La suite de normes IIIF (*International Image Interoperability Framework*) a été principalement portée par un réseau de grosses bibliothèques (dont la BNF) ayant investi lourdement dans des programmes de numérisation et qui ont compris la nécessité d'offrir des points d'entrée (ce que techniquement on appelle des API - *Application Programming Interface*) à leur base d'images. Le problème technique est simple : comment faire en sorte qu'un service externe — visualisation de fonds, transcription automatique, déploiement d'une édition numérique d'une œuvre — puisse s'appuyer sur un fonds d'images d'un établissement patrimoniale sans avoir à télécharger les images correspondantes avec toutes les conséquences néfastes que l'on peut imaginer (duplication, décalage avec le référentiel descriptif, dédoublement des citations etc.). Ces bibliothèques donc, constituées en consortium dont la structure rappelle un peu celui de la TEI, ont mis en place une vraie démarche de normalisation pour offrir des interfaces d'accès simples, reposant sur les protocoles de bas niveau du web (e.g. HTTP), qui permettent par exemple de pouvoir accéder à tout ou partie d'une image disponible sur un serveur, accompagnée des métadonnées correspondantes.

De fait, IIIF peut devenir un instrument essentiel d'interface entre le monde du patrimoine et celui de la recherche. Il est ainsi possible de diffuser des éditions critiques dans un environnement choisi par le chercheur (sur un serveur de laboratoire ou hébergées par une infrastructure telle que Huma-Num) tout en pointant vers les images accessibles depuis l'institution d'origine, comme le fait la BNF par exemple. Bien plus, des plates-formes de reconnaissance automatique d'écriture manuscrite telles qu'eScriptorium/Kraken[23] peuvent fonctionner par simple référencement IIIF aux images correspondante, facilitant ainsi de nouveau la réutilisation de contenus patrimoniaux numérisés.

Normes et patrimoine – un constant compromis cap-vitesse

J'espère avoir fourni au cours de ce survol rapide de différentes notions liées à la normalisation en contexte patrimonial une bonne idée des enjeux de disposer de normes à la fois solides techniquement et fiables dans le temps. Dans un contexte où toute opération de numérisation devient très vite couteuse si l'on combine les coûts de l'opération elle-

[23] https://escripta.hypotheses.org

même et la préservation à long terme, il est nécessaire que les communautés et les institutions en charge de la gestion du patrimoine numérique correspondant puisse s'appuyer sur des formats de données stables dans le temps pour que chaque évolution technologique ne nécessite pas une reprise complète du travail effectué. Bien qu'il soit parfois difficile d'échapper aux effets de mode et aux illusions de nouveauté (Poupeau, 2019), nous avons vu qu'il est possible dans un certain nombre de cas de disposer d'environnements normatifs offrant à la fois une stabilité organisationnelle et conceptuelle, qui seul permettant de disposer d'un patrimoine numérique solide et pérenne.

Références académiques

Donald Ervin Knuth, 'Literate Programming', *The Computer Journal*, 27.2 (1984), 97–111.

Poupeau, Gautier, 'Why I Don't Use Semantic Web Technologies Anymore, Even If They Still Influence Me? | Les Petites Cases', 2019 <http://www.lespetitescases.net/why-I-dont-use-semantic-web-technologies-anymore-even-if-they-still-influence-me>

Pitti, Daniel. "Encoded archival description: The development of an encoding standard for archival finding aids." *The American Archivist* 60.3 (1997): 268-283.

Romary, Laurent, 'Questions & Answers for TEI Newcomers', *Jahrbuch Für Computerphilologie 10*, 2009 <https://hal.archives-ouvertes.fr/hal-00348372>

Romary, Laurent et Charles Riondet. EAD-ODD: A solution for project-specific EAD schemes. *Archival Science*, Springer Verlag, 2018, ⟨10.1007/s10502-018-9290-y⟩. ⟨hal-01737568v2⟩

Wicentowski, Joseph, 'history.state.gov: A case study of Digital Humanities in Government', *Journal of the Chicago Colloquium on Digital Humanities and Computer Science*, 1(3), 2011.

中村覚, 佐治奈通子, & 永崎研宣. (2019). TEI と IIIF をベースとしたオン/オフライン併合型史料研究支援システムの開発-オスマン・トルコ語文書群を対象として. *じんもんこん 2019 論文集*, *2019*, 293-300.

Liste des normes citées

ISO/R 639:1967 Symbols for languages, countries and authorities [retirée]

ISO 639-1:2002 Codes pour la représentation des noms de langue — Partie 1: Code alpha-2

ISO 639-2:1998 Codes pour la représentation des noms de langue — Partie 2: Code alpha-3

ISO 639-3:2007 Codes pour la représentation des noms de langues — Partie 3: Code alpha-3 pour un traitement exhaustif des langues

ISO 639-4:2010 Codes pour la représentation des noms de langue — Partie 4: Principes généraux pour le codage de la représentation des noms de langue et d'entités connexes, et lignes directrices pour la mise en œuvre

ISO 639-5:2008 Codes pour la représentation des noms de langue — Partie 5: Code alpha-3 pour les familles de langues et groupes de langues

ISO 8601-1:2019 Date et heure — Représentations pour l'échange d'information — Partie 1: Règles de base

ISO 3166-1:2020 Codes pour la représentation des noms de pays et de leurs subdivisions — Partie 1: Codes de pays

ISO 3166-2:2020 Codes pour la représentation des noms de pays et de leurs subdivisions — Partie 2: Code pour les subdivisions de pays

ISO 3166-3:2013 Codes pour la représentation des noms de pays et de leurs subdivisions — Partie 3: Code pour les noms de pays antérieurement utilisés

ISO/IEC 10646:2017 Technologies de l'information — Jeu universel de caractères codés (JUC)

ISO 15924:2004 Information et documentation — Codes pour la représentation des noms d'écritures

Phillips, A., Ed., and M. Davis, Ed., "Tags for Identifying Languages", BCP 47, RFC 5646, September 2009.

TEI Consortium, eds. *TEI P5: Guidelines for Electronic Text Encoding and Interchange.* TEI Consortium. http://www.tei-c.org/Guidelines/P5/.

Biographie

Laurent Romary est directeur de recherche à Inria au sein de l’équipe ALMAnaCH. Ancien directeur général de l’infrastructure européenne DARIAH, il conduit actuellement des

recherches en humanités numériques sur la modélisation de données en sciences humaines et sociales, avec un intérêt particulier pour les données lexicales. Il a contribué au fil des années à de nombreuses action normatives, présidant en particulier le comité ISO/TC 37/SC 4 (ressources linguistiques, 2002-2014), le comité ISO/TC 37 (langue et terminologie, 2016-) et la *Text Encoding Initiative* en présidant son conseil technique de 2008 à 2011. Engagé dans la démarche de sciences ouverte, il a défini la politique en la matière de plusieurs institutions de recherche.