# Le projet et la plateforme "Millefeuille" : recherches et outils informatiques pour de nouveaux usages des almanachs

Jean-Daniel Fekete, Denise Ogilvie

# Le projet et la plate-forme « Millefeuille » recherches et outils informatiques pour de nouveaux usages des almanachs.

Jean-Daniel Fekete et Denise Ogilvie.

Le projet Millefeuille a été initié en 2003 dans le cadre du programme de recherche « Archéologie des savoirs administratifs, **Construction, conservation et circulations des corpus** XVIIIe XIXe siècle » financé par le CNRS. Un des objectifs de ce programme – comprendre et restituer les logiques de production et de transmission des archives - nécessitait la mise en relation des informations concernant les structures administratives – par exemple le bureau d'un ministère à une date donnée – et la description des documents qu'il a produit et qui sont conservés ses productions administratives

Il fallait donc disposer d'une description de l'administration française avec une précision et sur une durée suffisantes pour pouvoir répondre aux objectifs du projet, ce qui n'avait jamais été fait. C'est pourquoi ce programme a été conçu, dès l'origine, en collaboration avec l'équipe AVIZ de l'INRIA (Institut national de Recherche en Informatique et en Automatique)[1]. Outre le travail interdisciplinaire de réflexion sur la description formelle de l'administration française, il a fallu saisir informatiquement cette structure pour pouvoir l'exploiter.

Le projet consistant à décrire formellement cette administration sous forme informatique a été appelé « projet Millefeuille ». L'outil informatique utilisable pour saisir la structure administrative est la « plateforme Millefeuille ».

La description de l'administration française repose sur la collection des almanachs (royaux, nationaux, impériaux), dont l'usage comme instrument de référence est très répandu parmi les archivistes et les historiens modernistes ou spécialistes du XIXe siècle. Ces almanachs, ancêtres de notre Bottin administratif, décrivent l'administration française de manière relativement complète dans une logique de service (à quel bureau s'adresser pour telle type de demande).

La première des étapes du projet Millefeuille a été de définir un nouvel usage des almanachs qui permette de représenter et de comprendre l'évolution des structures administratives à travers une description formalisée du texte et des informations qu'il contient.

La plateforme Millefeuille – un ensemble logiciel –a été conçue pour la saisie collaborative (plusieurs personnes appartenant à plusieurs organismes travaillant pour enrichir un même projet), le contrôle scientifique des données préalablement à leur utilisation à des fins d'analyse et de visualisation. Cette plate-forme Millefeuille a été également configurée pour permettre une consultation confortable et hypertextuelle des almanachs traités et accueillir une première version de l'outil de visualisation des structures et de leur évolution.

[1] Le projet AVIZ, dirigé par Jean-Daniel Fekete, vise à améliorer les méthodes d'analyse de grandes masses de données par intégration de méthodes d'analyse (statistiques ou autres) et de visualisation d'informations. Voir www.aviz.fr.

## I. Une première étape du projet Millefeuille : représenter les mutations administratives.

### *Décrire l'administration : le traitement des almanachs, problèmes de méthodes.*

Comment décrire l'administration française et son évolution ? Le choix de la collection des almanachs s'est imposé très tôt dans le projet. Cependant, son traitement ne pouvait manquer de soulever d'importants problèmes de méthodes. En premier lieu, la nature du projet éditorial de l'almanach, sa continuité, sa régularité et la relative stabilité de sa mise en forme constituaient, en soi, une représentation formalisée de l'administration par - et pour les contemporains. Au fil du travail sur le statut de l'almanach et celui de ses éditeurs, sur ses conditions de fabrication, sur les interventions du pouvoir, ses censures et ses directives, s'accumulaient les preuves que tout le para-texte de l'ouvrage, et notamment le péri-texte, devaient faire l'objet d'une étude et d'une description attentives[2]. Les filets séparant les unités administratives, les alternances typographiques étaient utilisées en effet comme autant de manières, transparentes pour les contemporains, d'indiquer - ou d'esquiver- les liens hiérarchiques entre les différentes parties de l'administration décrite. Les marques d'impression, comme les signatures des feuilles d'impression, signalant la présence de "cartons", pouvaient souvent signifier une intervention de la censure. Il était donc indispensable de considérer ces éléments comme des données à traiter au même titre que les données textuelles.

En second lieu, la réflexion sur le traitement des informations décrivant les entités administratives (la dénomination des services, les noms, titres, qualités et fonctions du personnel, les adresses des services et leurs attributions), a conduit à prêter une attention particulière au traitement des listes d'« attributions »[3] présentées à la suite de l'intitulé d'un bureau (ou d'une division ou d' un département) pour définir les domaines et les types d'interventions qui lui sont « attribués ». Ces énoncés souvent complexes, toujours clairement séparés, sont apparus comme étant le niveau de description le plus pertinent pour croiser les observations sur l'évolution des structures administratives et le destin des ensembles documentaires qu'elles ont produit. La mise à jour du « voyage » des attributions au fil de l'évolution de l'administration constitue un des bénéfices attendus de l'outil de visualisation en cours de construction. La nécessité d'analyser, dans cet objectif, les modifications de leur énoncé, a été un des motifs principaux de la mise au point de la plate-forme Millefeuille (voir chapitre II).

Enfin, ne traitant qu'une partie des domaines administratifs présentés dans l'almanach[4], nous avons fait le choix, pour préserver les possibilités d'évaluer la place de cet échantillon dans l'architecture d'ensemble de

[2] Voir l'article de Nicole Brondel.

[3] Voir l'article d'Igor Moullier.

[4] La sélection de l'échantillon nécessaire à cette construction s'est faite dans le cadre des limites chronologiques et thématiques du projet « Archéologie des savoirs administratifs » : les administrations chargées de l'organisation et du contrôle de l'Intérieur et des Ponts et chaussées, pour la période 1750-1850.

l'ouvrage, de faire figurer, pour chacune des années sélectionnées, l'intitulé des chapitres qui n'avaient pas été traités.

***Le choix d'un langage de description.***

Comment représenter concrètement l'évolution de l'administration française ? Depuis plusieurs dizaines d'années, la question de la représentation de documents et de connaissance s'est posée en informatique et une réponse récente à cette question a été la définition du format XML. Ce format facilite la description de documents et d'information en fixant un certain nombre de conventions qui simplifient grandement l'échange et l'interprétation d'informations. Le projet Millefeuille a tout naturellement opté pour ce format pour décrire le contenu des almanachs. Le principe de ce langage de description est d'encadrer les fragments du texte que l'on veut annoter par un système de balises qui en définit la fonction dans le texte. Par exemple, le nom d'un bureau est indiqué par la balise **`<div type="bureau">`**.

Cependant, des règles de format informatique ne suffisent pas à résoudre le problème de la description de l'administration française dans le temps. Il existe plusieurs façons de décrire cette administration. Deux méthodes s'opposent : la méthode que nous appellerons « structurelle » et la méthode « textuelle ».

Dans la méthode structurelle, l'administration est idéalisée comme une organisation généralement hiérarchique composée de ministères, de services ayant des attributions, de personnes ayant des responsabilités dans les services. La représentation informatique est donc le reflet de cette organisation structurelle.

Dans la méthode textuelle, la structure de l'administration est interprétée comme une annotation structurelle du texte des almanachs. Ce texte est saisi aussi précisément que possible et des fragments de ce texte sont qualifiés structurellement. Par exemple, lorsque le texte indique le nom d'un service, une annotation XML indique qu'il s'agit du nom d'un service. Les fragments n'ayant pas de fonction structurelle ne sont pas annotés mais sont présents. De plus, les fragments dont la nature structurelle est sujette à caution sont marqués comme telle.

Nous avons choisi cette dernière méthode de description de l'administration car les almanachs sont des documents complexes dont l'interprétation est parfois difficile ou incertaine. Elle permet en outre de combiner les données provenant de l'analyse diplomatique du texte et les informations structurelles.

Des recommandations spécifiques pour la description des textes en langage XML (répertoire de balises et mode d'emploi) ont été mises au point par un groupement d'universitaires, la Text Encoding Initiative (TEI). Ce sont ces recommandations[5] que nous avons utilisées. Elles ont permis la description d'une double palette de niveaux de structure: la description des structures éditoriales des almanachs (alternances typographiques, haut-de-page et bas-de-page, marques d'impression, vignettes, frises, filets[6]) et la description des structure administratives elles-mêmes. Une description analytique fine des éléments de contenu du texte a permis la construction de

[5] dans la version TEI-P5.

[6] L'encodage des vignettes, frises, ou filets, souvent indispensables à la compréhension des liens entre les entités décrites, ont fait l'objet d'une description et d'une illustration à l'aide d'images associées.

typologies (types de services, types de fonctions, types de lieux, types de qualités, etc). Ce sont sur ces typologies que sont construits les index de la plate-forme Millefeuille (voir chapitre II). L'ensemble des choix d'encodage a été réuni dans un manuel rédigé par Nicole Dufournaud[7].

## II. Les outils du projet : le visualiseur et la plate-forme Millefeuille

### *Voir pour comprendre : un outil de visualisation dynamique des structures administratives et de leur évolution.*

Une fois effectué le travail d'encodage de l'almanach, on obtient une série d'instantanés, année après année, des structures administratives. Mais la vie de l'administration, sur la longue durée, les évolutions des services et de leurs relations, restent difficiles à saisir. Les scissions ou regroupements de services, les migrations ou dédoublement d'attributions, le rôle de personnalités dans la condensation ou l'éclatement des responsabilités, sont des événements complexes, le plus souvent asynchrones, et il est plus facile pour les comprendre d'en produire une représentation graphique et dynamique. C'est ce que permet le visualiseur arborescent conçu par Jean-Daniel Fekete dans le cadre du « projet Millefeuille » et qui est accessible à partir de la plateforme « Millefeuille ».

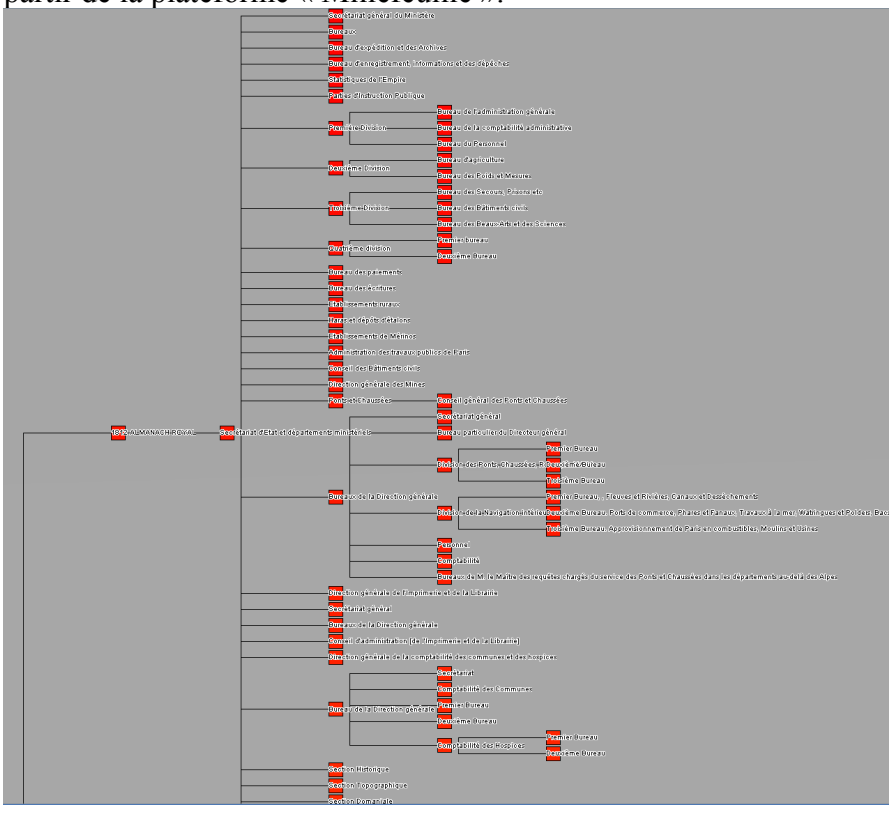

*Figure 1Aperçu de la hiérarchie des services de  l'année 1812*

Dans le cas des structures administratives, représentées par exemple sous forme d'arbres si elles sont hiérarchiques (voir figure 1), il est facile de détecter immédiatement, visuellement, la profondeur et la largeur de l'organisation. Il est aussi possible de comparer un ministère ou un service année après année (voir figure 2). On observe alors les changements dans

[7] N. Dufournaud, Manuel d'encodage XML-TEI des almanachs royaux, nationaux et impériaux (1750 – 1850) dans le cadre du projet Millefeuille, janvier 2008.

l'organisation des services. La visualisation arborescente permet, beaucoup plus rapidement, de comprendre comment fonctionne et évolue une structure. C'est une représentation compacte sur laquelle on peut raisonner très rapidement et qui permet à la fois de se faire une représentation d'ensemble et d'accéder directement aux détails significatifs.

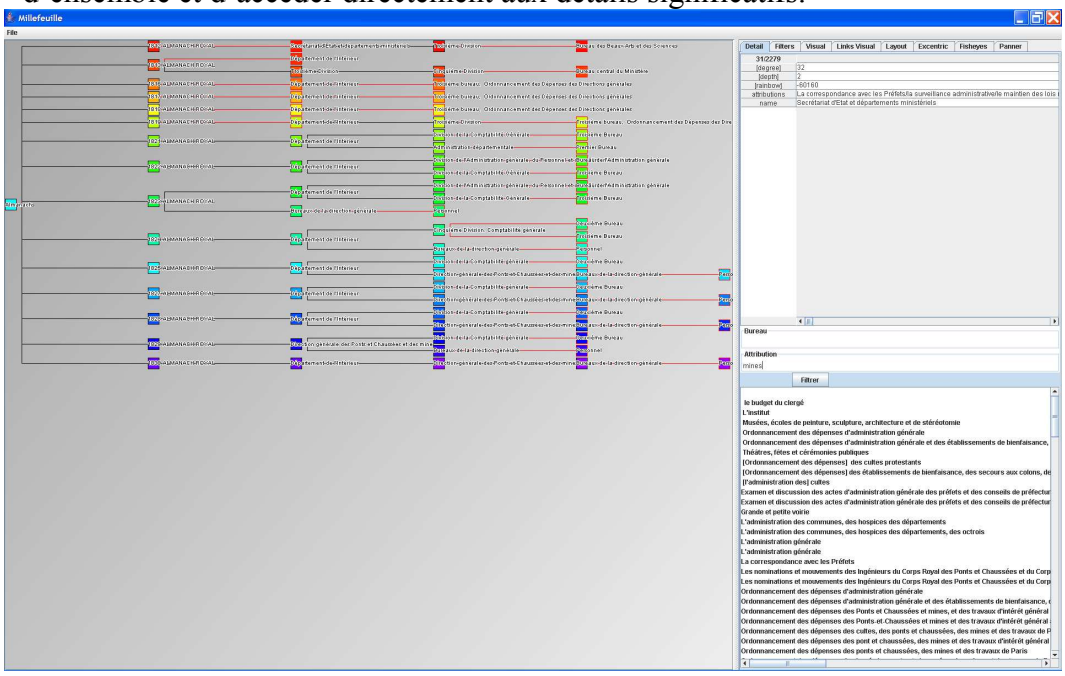


*Figure 2: Représentation d'un même service sur plusieurs années*

Le choix du mode de navigation est dynamique et extrêmement rapide ; il permet de voir se dessiner le parcours suivi par un objet d'intérêt en inscrivant une partie de son nom le champ d'interrogation (une personne ou un service, etc). Tout ce qui ne concerne pas cet objet est alors caché.

***Un outil pour analyser et préparer les données : la plate-forme collaborative « Millefeuille » :***

Pour suivre la trace des mutations des services, des parcours individuels, des transferts d'attributions, des changements de locaux, etc.., -- c'est-à-dire pour construire un outil d'analyse - il faut pouvoir indiquer avec précision, pour chaque année, si les dénominations des services, les noms des personnes, l'énoncé des attributions, les adresses des services ont ou non subi un changement. En terme informatique, on associe un identifiant à chaque fragment d'intérêt (nom de personne ou de lieu) et on lie les fragments identifiés lorsqu'on considère qu'ils sont identiques. Cette validation doit être rapide et scientifiquement fondée. Il est indispensable de pouvoir d'une part comparer rapidement les occurrences liées d'un même terme pour toutes les années encodées, d'autre part d'examiner le terme d'une année dans son contexte pour éviter des contre-sens. La plate-forme Millefeuille facilite ce travail en présentant des index servant à la structuration (on peut ajouter des termes ou lier des termes) et à la navigation.

Le principe de la plateforme est de générer autant d'index que nécessaire (toute séquence de texte encadrée par des balises est susceptible de faire l'objet d'un index), présenté en regard du texte sous forme de liste, interrogeable – et immédiatement modifiable – par n'importe quel terme. Chaque type d'index (noms de services, attributions, fonctions, noms de

personne, noms de lieux, adresses, etc..) existe pour chaque année et pour toutes les années cumulées. Chaque occurrence renvoie directement (en cliquant) au texte sous sa forme encodée et sous une forme restituée directement lisible.

La plateforme Millefeuille se présente comme une extension de la plateforme de programmation « Eclipse ». Le processus suivi par les programmeurs pour aboutir à un programme qui fonctionne correctement est très proche du processus suivi par des historiens pour aboutir à des documents annotés et enrichis en XML. Ils travaillent à plusieurs, doivent partager leurs fichiers, les vérifier de plusieurs manières et partager des outils de vérification et d'indexation. La plateforme Eclipse permet tout cela et surtout l'intégration de modules complémentaires sous la forme d'ajouts nommés « plugins ». Millefeuille est donc un plugin de la plateforme Eclipse qui permet de saisir des documents comme les almanachs[8]. Eclipse repose sur des principes de programmation qui s'appliquent à des projets individuels aussi bien qu'à des projets collectifs avec plusieurs milliers de collaborateurs. Le fait d'utiliser Eclipse comme base offre un nombre de services important. Par exemple, le projet Millefeuille a été réalisé par plusieurs personnes dans des lieux éloignés. Eclipse a permis le partage à distance des documents XML de manière très simple en se reposant sur un système de gestion de version[9].

Ainsi, chaque collaborateur peut avoir accès à chaque document, regarder qui l'a créé ou modifié, la nature des modifications et, le cas échéant, revenir sur une modification antérieure ; toutes les versions successives sont conservées. Cette infrastructure de travail évite toute perte de temps et d'information rencontrée dans d'autres projets ou les documents sont envoyés par courrier électronique d'une personne à une autre et où il est difficile d'être sûr d'avoir la dernière version et d'éviter que deux personnes ne modifient en même temps le même document. En utilisant une infrastructure de gestion de projet associée à la plateforme (mais optionnelle, la plateforme est utilisable de manière autonome), le projet Millefeuille a permis une répartition du travail d'analyse selon les compétences et les spécialités de chacun des membres de l'équipe.

Cette plate-forme se présente donc comme un outil inédit d'aide au travail scientifique de préparation des données à encoder en XML, quelque soit le répertoire de balises utilisé[10]. Plusieurs groupes de recherches l'on déjà adoptée pour l'édition de texte[11] ou le catalogage de manuscrits[12]. C'est à partir de cette plate-forme que pourront être expérimentés de novelles exploitations de l'almanach.

---

[8] Le plugin Millefeuille a été programmé par Félicien François, sous la direction de Jean-Daniel Fekete. Il est accessible gratuitement à l'adresse Web suivante : http://millefeuille.gforge.inria.fr

[9] Le choix s'est porté sur le logiciel libre de gestion de version « Subversion » publié sous licence Apache/BSD.

[10] La cohabitation entre des données décrites dans plusieurs environnements XML a déjà été expérimentée : almanachs encodés en TEI-P5 et inventaires encodés en EAD (Encoding Archive Description).

[11] CESR (Centre d'Etudes Supérieures de la Renaissance) pour le programme BVH (Bibliothèque virtuelle humaniste).

[12] Laboratoire CEMAf-CNRS pour la base de données Zekrä Nägär « Corpus électronique des archives manuscrites éthiopiennes » du programme Cornafrique soutenu par l' Agence nationale pour la Recherche.

## III. Les premiers bénéfices du projet Millefeuille et de sa plate-forme : de nouveaux usages de l'almanach.

L'usage de l'almanach, nouveau ou pas, n'a de sens qu'à condition de disposer d'un échantillon suffisant, tant en terme de tranches chronologiques qu'en terme de parties de l'appareil administratif concernées. L'entreprise de saisir une partie important – voire la totalité – du corpus, pour être ambitieuse, n'est pas du tout irréalisable. Grâce a la plateforme Millefeuille, il est maintenant possible d'organiser la saisie en masse de la partie des ouvrages concernant la description de l'administration et d'en produire un premier encodage.

Installée sur la plate-forme Millefeuille, munie des index directement issus du premier encodage du texte, cette première version de travail offrira tout d'abord l'intérêt de faciliter l'usage de cette collection qui n'est pas toujours accessible, qui le plus souvent est incomplète et d'une manipulation difficile. L'accès direct au texte et la multiplicité des index représente déjà un service qui peut favoriser une plus grande diffusion de la collection et en permettre une pratique plus systématique – ce qui constitue en soit une avancée importante. Cette pratique reste néanmoins dans la stricte continuité de l'usage qui en a été fait jusqu'à maintenant. Une saisie et un encodage en masse augmenteront cependant les chances de la mise en oeuvre d'usages nouveaux. La constitution de ce corpus de données encodées est une condition indispensable à la poursuite du travail scientifique en cours, sur l'almanach abordé comme objet scientifique et non comme instrument de référence déjà constitué d'une part, d'autre part à partir de l'almanach, pour la connaissance du fonctionnement de l'administration et de l'origine des documents conservés.

### *Un almanach « consolidé »*

Le premier bénéfice du projet Millefeuille est de permettre une description scientifique de la collection. L'usage de l'almanach souffre en effet de l'absence des éléments d'information qui en permettent la critique : dates de parution, mises à jour - sous la forme d'insertion de cartons, de rééditions, de la publication de suppléments ou d'abrégés. Ces informations, ainsi que les variantes lorsque le texte de deux éditions diverge, sont progressivement intégrées aux analyses en cours, contribuant à l'élaboration d'un catalogage de référence des collections.

Le travail a permis également de mieux comprendre les modalités de collecte des informations par les éditeurs, ainsi que celles des interventions du pouvoir. Il a permis d'évaluer les difficultés d'interprétation des indices typographiques – parfois volontairement évasifs – utilisés pour traduire les liens existant entre différentes entités administratives. Il s'avère impossible, au vu de ces résultats, d'utiliser directement les informations de l'almanach pour alimenter un système de données d'autorité sur les services administratifs tel qu'il est défini par la norme archivistique ISAAR-CPF traduite, pour la constitution de fichiers électroniques, par les recommandations du standard EAC (Encoding Archival Context). La source principale d'une telle entreprise reste les textes législatifs ou les décisions qui portent explicitement création ou modification d'une entité administrative. Il n'est pas du tout impossible d'ailleurs de créer des liens entre les deux systèmes de représentation: notices d'autorité décrivant les

services à un temps t présentées en XML-EAC et description de l'almanach en XML TEI.

***Nouveaux usages : des usages anciens cumulés***

Un bénéfice voisin peut être retiré de la possibilité de déposer progressivement sur la plate-forme Millefeuille les textes législatifs qui encadrent, préparent et définissent l'action de l'administration : derrière les termes utilisés dans la définition d'une « attribution » transparaissent le plus souvent les modalités d'application d'une loi ou les nouvelles dispositions d'un décret. La recherche de ces textes constitue depuis toujours une des approches privilégiées par l'archiviste responsable d'un classement pour identifier les procédures de travail à l'origine des documents qu'il traite. Ces informations, exposées et diffusées dans les introductions scientifiques des inventaires publiés, sont difficilement cumulables, et parfois négligées. Leur intégration à une plate-forme de travail qui permette de confronter les textes à la description des bureaux chargés de leur mise en oeuvre enrichit considérablement la connaissance du cadre du travail administratif nécessaire au traitement des fonds et à leur exploitation.

L'intérêt du dispositif s'accroît si l'on envisage d'y adjoindre les informations concernant les documents conservés. Une étape importante de cette nouvelle déclinaison d'un environnement de travail autour des almanachs concerne la description des versements d'archives effectués aux Archives nationales par les ministères (leurs bureaux ou leurs services d'archives) au cours du XIXe siècle. La description[13] et l'exploitation des bordereaux et des registres de versements conservés à la section du XIXe siècle est en cours[14] dans le cadre d'un projet qui prolonge en partie le projet Millefeuille.

***Un environnement de travail pour la construction du visualiseur.***

Il est maintenant possible de constituer progressivement, sans préjuger des évolutions techniques à venir, un environnement de travail qui permette de rassembler les données nécessaires à l'exploitation des fonds archivistiques. Cet environnement complexe bénéficie de l'association de deux puissants supports : en premier lieu celui de l'almanach, qui, par sa continuité, sa régularité et son formalisme originel, offre une colonne vertébrale solide propre à accueillir des informations ponctuelles ou discontinues dont il assure la cohésion; en second lieu un langage de description commun, propre à mettre en forme des informations dont on peut organiser le dialogue : le langage XML, décliné selon plusieurs standards adaptés mais compatibles et exploitables de façon durable.

Le passage dans les technologies du XXI[e] siècle de ces sources importantes et utiles, tels les almanachs, leur donnera une nouvelle jeunesse et facilitera, nous en sommes certains, leur exploitation dans des contextes extrêmement variés et créatifs dans les années voire …les siècles à venir.

---

[13] Cette description en XML se fera en TEI-P5 (Text Encoding Initiative, version P5).

[14] Ce programme de travail a reçu en 2008 le soutien de l'Agence Nationale de la Recherche dans le cadre du projet MOSARE présenté en réponse à l'appel d'offre Gouverner et administrer.