Humanités numériques et manuscrits

Aujourd'hui je vais tenter de présenter quelques aspects des humanités numériques, en général, et en particulier en rapport avec les études sur les manuscrits, pour tout d'abord comprendre de quoi on parle, même si on ne parle pas toujours de la même chose. En effet selon les disciplines et même à l'intérieur des disciplines il y a des avis qui divergent. J'ai donc essayé de donner quelques éléments d'historique dans une sélection qui n'est pas exhaustive, et puis j'ai voulu mentionner quelques thématiques et au sein de ces thématiques quelques projets actuels, et enfin finir par quelques concepts qui ont traversé le développement des humanités numériques et qui les traversent encore.

Que sont les humanités numériques ?

(D) Il est impossible de résumer en une phrase ce que sont les humanités numériques, et d'ailleurs les contours et les limites de cette notion sont largement sujet à débat. Lorsque l'on demande à Google ce que sont les humanités numériques, on trouve des images qui évoquent le rapport à la machine : notamment ici le lien entre l'Homme et la machine, via un texte encodé en binaire, ou le lien entre l'ordinateur et le texte manuscrit, grâce à l'action de quelques humains pour le transmettre à un grand nombre d'humains.
(D) Les humanités numériques se sont constituées en tant que telles à la suite d'une évolution longue de plusieurs décennies et correspondant à l'utilisation de plus en plus fréquente, intense et diversifiée de l'outil informatique dans la recherche en sciences humaines et sociales, dans les arts et dans les lettres. Je dis outil informatique, mais évidemment on comprendra que cela ne se limite pas aux outils de bureautique ou à la création de pages web. Il s'agit de l'outil informatique appliqué à différentes phases du processus de gestion des données de la recherche, de la phase de constitution, de formalisation, de traitement, d'analyse, et de publication des données de la recherche, mais aussi de leur conservation. Ce sont donc des champs d'application assez larges dans lesquels on va, de plus en plus souvent, rencontrer le concept d'humanités numériques.

(D) Dans le manifeste des Digital humanities[1] de 2010, les humanités numériques sont définies comme une « transdiscipline ». Il ne s'agit donc pas d'une nouvelle discipline qui serait totalement interdisciplinaire, mais d'une posture active qui a de multiples aspects :

- bien entendu, l'utilisation, et la création d'outils numériques qui peuvent être communs à plusieurs disciplines, et leur configuration pour sa discipline propre
- la définition de méthodes, de dispositifs et de perspectives heuristiques liés aux outils numériques : le numérique est alors envisagé comme un outil pour la recherche
- une réflexion sur les apports et les limites de ces outils en regard des méthodes traditionnelles, ainsi que sur les conséquences de leur utilisation : on est alors dans une perspective critique sur les humanités numériques
- une réflexion également sur la nature et l'évolution des métiers de la recherche dans ce contexte et notamment des rapports entre ingénieur informaticien et chercheur, et l'émergence d'une nouvelle catégorie d'ingénieur, ce qui veut dire une implication sociale

(D) La définition des humanités numériques comme « transciscipline » ne fait pas l'unanimité, notamment en France, où la critique se porte sur le côté « accessoire » de l'aspect « humanités numériques » dans un certain nombre de projets, et sur la nécessité que le chercheur soit lui-même un programmateur « autonome » en quelque sorte ou alors qu'ils se fasse aider par des informaticiens ou des ingénieurs.

1 https://tcp.hypotheses.org/318

Certains appellent même à la fin des humanités numériques et à la refonte complète du cursus de l'historien et du chercheur en SHS, qui devrait intégrer pleinement l'utilisation des outils numériques, c'est le cas de Emilien Ruiz et Paul Bertrand dans les deux premières références que vous avez sous les yeux, tandis que d'autres appellent à l'intégration de l'aspect numérique au sein même de la recherche en sciences humaines et sociales, c'est l'idée du billet de Gautier Poupeau dans la troisième référence, qui est orientée du point de vue de l'ingénieur, que je vous suggère de lire si ce thème vous intéresse. Les trois blogs sont très intéressants à cet égard.

Les humanités numériques et les études sur les manuscrits

Domaines

(D) Comme vous le savez les manuscrits tiennent une place à part, à la fois dans le patrimoine et pour la recherche : en effet ils peuvent être considérés sous plusieurs aspects :

- le manuscrit en tant qu'objet précieux, de collection, que l'on peut vouloir reproduire tel quel, notamment sur les sites d'institutions qui conservent des manuscrits, comme les bibliothèques
- mais aussi l'aspect du contenu intellectuel et artistique
- et enfin en tant qu'objet matériel, objet de collection, de commerce, d'artisanat, signe de prestige social ou outil de travail, ou manuel d'études universitaires, en gros le manuscrit comme objet historique, qui porte la trace et est un témoin de pratiques socio-économiques

(D) En raison de ces différents aspects, on va retrouver les manuscrits au centre de différents champs liés aux humanités numériques, parmi lesquels :

- numérisation des manuscrits et leur mise en ligne, vous connaissez tous les grandes entreprises de numérisation des bibliothèques nationales d'Europe comme la BNF, la British Library, la Bibliothèque Vaticane
- édition électronique, qui n'est pas une numérisation, ni des images, ni d'une édition papier
- études de stemmatologie assistée par ordinateur, qui visent donc à établir le stemma ou arbre généalogique si on veut d'une tradition manuscrite, en générant des arbres de distance entre les textes portés par les manuscrits
- analyses sémantiques et fouille de texte : analyse du vocabulaire, analyse linguistique
- également l'analyse de réseaux, qui a fini par gagner aussi les études sur les manuscrits, après avoir fait son entrée en histoire plutôt par l'histoire quantitative et la prosopographie : on peut analyser le contenu d'un manuscrit sous forme de réseau, ou le voisinage textuel des œuvres par exemple, c'est à dire la présence des œuvres ou des genres dans les mêmes manuscrits, en rapport avec la constitution des grandes collections de livres au Moyen Âge
- codicologie quantitative, c'est à dire une approche d'exploitation statistique de la description des manuscrits

Histoire

(D) Historiquement, les disciplines de l'érudition sont au coeur du développement des humanités numériques, puisque c'est en philologie que les HN se sont d'abord développées

- dans le domaine de l'étude des textes, les travaux du jésuite Roberto Busa autour de l'informatisation de l'index des éditions critiques des œuvres de Thomas d'Aquin, entamé en 1949 avec le financement d'IBM font figure de précurseurs => ce travail est aujourd'hui accessible sur le site du *Corpus thomisticum*
- **(D)** dans le domaine de l'étude des textes bibliques également, on fait les premiers essais d'utilisation de l'ordinateur dans les années 50. John Ellison soutient une thèse en 1957 à Harvard qui s'intitule « The Use of Electronic Computers in the Study of the Greek New

Testament Text » dans laquelle il publie une concordance de la Bible établie avec l'ordinateur.

- Le travail d'Ellison porte aussi sur l'étude des traditions textuelles et le travail de reconstitution du stemma, c'est-à-dire la reconstitution de l'histoire et des relations entre les manuscrits conservés d'une œuvre, à travers l'étude des variantes que l'on relève dans le texte de ces manuscrits, ici en l'occurrence la Bible. En effet il utilise l'ordinateur pour regrouper les manuscrits à partir des variantes et les comparer à ceux établis par ses prédécesseurs.
- En 1926 déjà Dom Quentin avait inventé une méthode de groupement de variantes par triplets pour identifier le manuscrit intermédiaire entre deux autres. En 1968, Dom Froger reprend sa méthode pour la perfectionner et l'adapter à l'ordinateur.
- En 1969, John Griffith fait traiter des masses de variantes à l'ordinateur pour établir une matrice de similarité grâce au traitement de la concordance des manuscrits grecs du Nouveau Testament sur chaque variante.

(D) Les approches « informatiques » dans les disciplines de l'érudition se développent conjointement à la faveur des développements techniques, et parfois s'inspirent de ce qui est fait dans d'autres domaines scientifiques :

- en biologie évolutionniste : l'approche phylogénétique, qui se fonde sur la cladistique, qui vise à reconstituer les rapports de parenté entre les êtres vivants à l'aide d'arbres = cladogrammes
- en linguistique et notamment en traitement automatisé du langage avec une Association for Computational Linguistics fondée aux Etats-Unis en 1962, dans la continuité des efforts de l'armée américaine investis dans la traduction automatique, notamment du russe[2], à partir de la fin des années 40, et un symposium international sur le thème « Computers in Literary and Linguistic Research » depuis 1970 en Angleterre.
- on peut citer aussi les progrès de l'IA et du *Deep learning* qui sont aujourd'hui utilisés dans des projets de reconnaissance automatique des écritures manuscrites

(D) Ce sont aussi des linguistes qui forment le noyau originel de la communauté qui a développé dès 1987 les Recommandations de la TEI, la Text Encoding Initiative, qui est un vocabulaire suivant les règles du langage informatique XML, qui est aussi à la base du XHTML que vous utilisez tous les jours sur internet.
Ces recommandations utilisent le langage XML dans le but spécifique de proposer une norme d'encodage assez souple pour encoder les textes de sciences humaines et sociales.

(D) A quoi sert la TEI ? l'objectif est de faciliter l'échange et le stockage d'informations électroniques, principalement de textes de sciences humaines et sociales.
Pour ce faire, on va encoder ou baliser un certain nombre d'éléments du contenu d'un document, afin de les rendre explicite. On parle alors de balisage sémantique par opposition à un balisage formel qui ne rendrait compte que de l'apparence qu'auraient la structure et les éléments internes du document. Avec XML et en particulier la TEI, on pose en quelque sorte des étiquettes sur les éléments : un titre, une date, une annotation marginale etc.
Pour baliser le texte, on utilise le langage XML, un langage de balises extensible.
Avec les règles d'XML, on définit des éléments, des attributs et des règles pour leur existence et leur agencement.

(D) En 2007 la version P5 de la TEI inclut pour la première fois trois modules spécifiquement consacrés aux manuscrits :

2 John Hutchins: Retrospect and prospect in computer-based translation. Proceedings of MT Summit VII, 1999, pp. 30–44.

- module 10 « Manuscript description »
- module 11 « Representation of primary sources »
- module 12 « Critical apparatus »

Je n'en dis pas plus sur la TEI, j'en parlerai plus avant dans l'atelier de mercredi soir.

(D) A l'IRHT, on est loin d'être en reste puisque la revue « Le médiéviste et l'ordinateur », est éditée dès 1979 sur des thématiques aussi avant-gardistes que le traitement de texte à la bande magnétique, la carte perforée, la saisie sur disquette : ces techniques, qui visaient à établir des index, des concordances ou des listes de citations par exemple, nous semblent aujourd'hui dépassées, mais il faut bien comprendre le progrès que cela représentait à l'époque où le crayon et la machine à écrire constituaient les principaux outils de travail sur les textes. Des thèmes aussi actuels que la nécessaire mutualisation des outils et le manque de moyens sont déjà abordés dans le premier numéro. La revue a paru jusqu'en 2006, après un dernier numéro qui portait sur les formes et couleurs dans les manuscrits du Moyen Age. En 2009, un communiqué annonçait la suspension de la parution, après avoir pris acte des mutations et des évolutions dans le domaine du numérique, l'expansion d'internet et la part croissante des activités en ligne.
Les numéros sont accessibles en ligne sur Persée à cette adresse, et les derniers numéros sont encore accessibles sur le site désormais inactif.
Citons aussi le colloque international sur « La pratique des ordinateurs dans la critique des textes » organisé par l'IRHT en 1978.

(D) On peut citer aussi la communauté internationale « Digital medievalist » qui comporte notamment une liste de diffusion et une revue, on retrouve toutes ces ressources sur leur site.

Domaines et projets

(D) Un dernier point sur les avancées les plus récentes en matière d'humanités numériques appliquées aux manuscrits. Depuis quelques années, l'offre en terme de manuscrits numérisés s'est accrue de manière exponentielle, fournissant aux chercheurs de nouveaux matériaux sur lesquels ils peuvent mener des études de masse, et ce dans plusieurs domaines :

- la reconnaissance automatique de textes manuscrits comme je l'ai déjà mentionné grâce à l'IA
  - renvoi aux travaux de D. Stutzmann, projet Himanis
  - projet comme e-scriptorium

- l'exploitation des métadonnées des sites de bibliothèques numériques comme e-codices, la bibliothèque virtuelle des manuscrits de Suisse, dont les descriptions sont structurées en XML-TEI et qui fait figure de modèle pour la qualité de ses métadonnées, et ce pour alimenter des bases de données qui vont donner lieu à des études d'analyse de réseaux.

Quelques principes et notions qui ont traversé l'histoire des humanités numériques

Je vais maintenant aborder quelques principes et notions qui ont traversé l'histoire des humanités numériques : la notion de libre accès, la notion d'interopérabilité et la notion de données fair ou fair data.

Libre accès

- question qui traverse toute la société qui a accès aux outils numérique depuis des années
- il faut distinguer entre libre, open source et gratuit

(D) Pour qu'un logiciel puisse être appelé libre, il doit respecter 4 critères qui sont des libertés :

La liberté d'exécuter le programme selon les besoins de l'utilisateur
La liberté d'étudier le fonctionnement du programme, et de le modifier comme on veut
La liberté de redistribuer les copies du logiciel initial
La liberté de distribuer des copies des versions modifiées

Ensuite pour ce qui est de l'Open source, on a une perspective un peu différente, qui s'exprime dans une liste d'obligations de faire ou de ne pas faire, notamment :
La redistribution doit être libre ;
Le programme doit être distribué avec le code source
La licence doit autoriser les modifications et les œuvres dérivées, et doit leur permettre d'être distribuées sous les mêmes termes que la licence du logiciel original ;
La licence peut exiger que les œuvres dérivées portent un nom ou un numéro de version différent de ceux du logiciel original, afin de préserver l'intégrité du code source de l'auteur,
=> et ensuite une liste d'interdictions de l'ajout de restrictions sur la licence

D'après Richard Stallman, l'un des initiateurs du mouvement du logiciel libre, la différence entre les deux notions réside dans leur philosophie : « l'open source est une méthodologie de développement; le logiciel libre est un mouvement social ».

=> à noter que ni libre, ni open source n'est synonyme de gratuit, ce n'est pas une obligation spécifique

- **(D)** pour la recherche, la question est cruciale dans le domaine des publications scientifiques
  - aujourd'hui, l'État et donc le contribuable paie deux à trois fois le coût de la recherche en ce qui concerne les publications scientifiques :
    - il paie le salaire du chercheur qui publie
    - il paie l'abonnement des bibliothèques universitaires et des centres de recherche à des bouquets de revue et les commandes de livres, abonnements qui sont souvent dénoncés comme exhorbitants et ce à cause de la position de monopole des grands éditeurs commerciaux
    - **(D)** parfois, le chercheur doit en plus payer des frais de publication à l'éditeur commercial

Ce problème génère des débats depuis des années au sein de la communauté scientifique, dans les institutions qui financent et également avec les éditeurs qui peinent à trouver de nouveaux modèles économiques.

- **(D)** Pour ce qui est des données de la recherche, 4 arguments pour l'ouverture des données de la recherche[3] :
  - => la possibilité de reproduire ou vérifier la recherche et donc de valider les résultats présentés

3 C.L. Borgman, «The conundrum of sharing research data», Journal of the American Society for Information Science and Technology, vol. 63, no 6 (juin 2012

- => l'ouverture au public des résultats de la recherche financée par des fonds publics
- => la possibilité donnée à d'autres de poser de nouvelles questions sur les données
- => l'opportunité de faire progresser l'état de la recherche et de l'innovation grâce au partage de données

Autant vous dire que la situation réelle est encore très loin de ce souhait.

- **(D)** Un exemple de restriction de l'ouverture des données de la recherche par les éditeurs commerciaux. Les éditeurs se servent du droit de la propriété intellectuelle pour restreindre l'accès aux publications scientifiques et même aux éditions critiques avec ou sans leur apparat, alors même que les ayant droit des auteurs des textes sont morts depuis des siècles
  - on appelle cet abus de droit le « copyfraud », c'est-à-dire une revendication infondée, frauduleuse, de droits de la propriété intellectuelle
  - dans le domaine des éditions critiques, peu d'affaires ont donné lieu à des litiges mais il y a un cas devenu célèbre en 2014 : celui de la plainte de l'éditeur Droz contre la maison Garnier qui continuait à publier des textes d'éditions critiques de poésie française du catalogue Droz, sans apparat, après la fin d'un contrat entre les deux maisons pour la publication de ces textes sur CD-Rom.
    - la maison Droz argumentait qu'il y avait contrefaçon et donc qu'elle détenait le droit de propriété intellectuelle sur ces textes qui constituaient de nouvelles œuvres au regard du droit
    - le jugement a précisé la nature particulière du travail de l'éditeur critique, qui utilise les méthodes à sa disposition et son propre jugement critique pour tenter de reconstituer un état d'un texte ancien à partir des diverses copies qu'il a à sa disposition, et sans y ajouter quelque chose de personnel, et donc il ne fait pas pour autant un travail de création d'une nouvelle œuvre, ni un travail d'adaptation ou de traduction, qui sont encadrés par le droit de la propriété intellectuelle
    - la maison Droz a donc été déboutée de sa plainte, ce qui devrait constituer une jurisprudence pour la reprise des textes dépouillés de leur apparat critique ; pour l'instant ce n'est pas le cas car il n'y a pas eu d'autres affaires de ce type à ma connaissance, ce qui veut dire que les chercheurs continuent à se conformer à un prétendu droit de la propriété intellectuelle détenu par les éditeurs commerciaux, ou qu'ils s'en affranchissent sans le dire – sur cette affaire avec des répercussions très intéressantes sur la question de l'originalité du travail du chercheur et du droit de la propriété intellectuelle appliquée aux produits de la recherche, jusque aux postures positivistes assumées par certains éditeurs critiques, vous pouvez aller voir notamment ces deux billets de blog[4]

  - on peut aussi argumenter sur la légalité des restrictions exigées par les bibliothèques dans le domaine de la reproduction et de la diffusion d'images des manuscrits qu'elles conservent : ces restrictions, bien qu'encore souvent appliquées, ne reposent sur aucun fondement juridique puisque les manuscrits appartiennent en principe au domaine public

  - vous allez me dire, cela concerne surtout les publications papier et pas spécifiquement les éditions électroniques, mais en fait, dès que l'on veut utiliser une édition même ancienne pour une base de données ou une édition électronique, on se

4 https://scinfolex.com/2014/04/13/une-victoire-pour-le-domaine-public-un-cas-de-copyfraud-reconnu-par-un-juge-francais/ ; https://apocryphes.hypotheses.org/389

heurte à une revendication de droit de la propriété intellectuelle de la part des éditeurs commerciaux, qui doivent certes trouver un modèle économique qui leur permette de subsister tout en ouvrant l'accès aux publications de la recherche comme le souhaitent maintenant depuis plusieurs années la communauté scientifique et les institutions qui la financent, ce qui n'est pas forcément évident

- **(D)** initiatives, particulièrement en France, pour l'édition ouverte
  - depuis 1999, le Centre pour l'édition électronique ouverte met en place et maintient des plateformes pour la publication en accès ouvert des résultats de la recherche, avec historiquement :
    - Revues.org
    - Hypothèses.org
    - OpenEdition Books
    - Calenda
    - => le tout a été refondu dans le portail OpenEdition en 2017 qui est à présent une infrastructure complète d'édition numérique en sciences humaines et sociales et est portée par le centre OpenEdition qui est une unité mixte de service du CNRS depuis 2009 et une USR unité de service et de recherche depuis 2018

- je signale rapidement Persée pour la publication électronique de numéros anciens de revues scientifiques ou à barrière mobile : Persée

  - C'est vraiment en 2019 que le Ministère s'est saisi de cette problématique avec le Plan national pour la science ouverte et cette déclaration de la Ministre qui souhaite que 100 % des publications scientifiques françaises soient en accès ouvert
  -

(D) On en est encore loin : on atteint apparemment le nombre de 62 % de publications françaises ouvertes en 2020, en tant qu'on peut les observer en 2021.
Il y a un site dédié au baromètre français de la science ouverte, vous pouvez aller voir.

Interopérabilité

(D) Une autre notion que je voulais aborder avec vous, c'est l'interopérabilité

- pendant des décennies, on a fait le constat que des solutions individuelles étaient le plus souvent développées pour subvenir aux besoins d'un projet avec une composante numérique, plutôt que de réutiliser les outils déjà existants
  - c'est une situation qui est loin d'être réglée
  - et ce plusieurs raisons :
    - les outils déjà existants ne sont pas toujours open source même si de plus en plus le sont par conscience du besoin de partage au sein des communautés et suite à l'impulsion des pouvoirs publics qui veulent pousser à la mise à disposition du public des résultats de la recherche, dont font partie les outils logiciels développés à côté des données produites
    - chaque chercheur qui porte un projet voudrait une solution qui soit adaptée à son projet avec le moins d'ajustements possibles de son processus de travail ou workflow pour le traitement de ses données => c'est plus ou moins vrai d'un chercheur à l'autre mais globalement, pour réutiliser un outil déjà existant, il faut que en gros, les deux parties – chercheur et ses données et l'outil à utiliser – se rapprochent d'un standard commun, avec si possible un outil configurable pour pouvoir faire tout de même des ajustements

- or plus on fait d'ajustement, plus on s'éloigne d'un standard commun
- et cela finit par poser des problèmes lorsque quelqu'un d'autre veut exploiter les données publiées par ce projet de recherche : plus on s'est éloigné du standard commun, moins ces jeux de données seront exploitables facilement

- parmi les solutions, l'utilisation de standards :
  - dans le langage d'encodage des données, l'exemple le plus connu est probablement celui de la TEI, même si les possibilités offertes par les recommandations de la TEI sont tellement vastes qu'il est nécessaire d'adopter des standards plus restreints afin de faciliter l'exploitation des données
  - pour cela, on peut passer par la modélisation des données, en s'aidant de :
    - schéma d'encodage communs, par exemple
    - d'ontologies pour un domaine particulier, une ontologie est l'ensemble des termes et concepts représentant les éléments d'un domaine de connaissances, ensemble qui a été structuré et hiérarchisé
  - dans les références, les index : on peut s'aider de thésauri, de listes d'autorités externes, d'outils de gestion de bibliographie comme Zotero

que ce soit au niveau des ontologies, des schémas communs, des thésaurus : ces ressources intermédiaires demandent un gros travail de coopération entre équipes de recherche et aussi avec des institutions patrimoniales : c'est un travail de rassemblement de données, de nettoyage, d'alignement, mais aussi, dans le cas des modélisations comme les ontologies ou les schémas d'encodage, de conceptualisation d'un minimum commun qui peut être accepté par tous les partenaires
=> c'est un travail qui est toujours en train de se faire et qui vraisemblablement durera des années

Je citerai plusieurs initiatives dans ce domaine pour nos disciplines de l'érudition :

- **(D)** l'Equipex Biblissima dont l'objectif était de fournir un observatoire du patrimoine écrit du Moyen Âge et de la Renaissance, via deux entreprises :
  - **(D)** un alignement de données de diverses provenances[5] :
    - référentiels unifiés, notamment sur les noms de personne physique et morales, les noms de géographie, les cotes de manuscrits
    - une ontologie qui permet de définir les relations entre les différentes classes de données
    - un thésaurus qui est formé par un ensemble de termes normalisés et reliés par des relations terminologiques et sémantiques ; ce thesaurus complète l'ontologie et est formé à partir de divers thesauri existants sur la géographie, l'iconographie, l'archéologie, la musique et les archives
  - **(D)** un portail sur l'histoire des textes et des livres, qui permet d'accéder avec un moteur de recherche unifié aux collections de diverses institutions partenaires, grâce à l'utilisation de ce dont je viens de parler, donc des référentiels, de l'ontologie et du thésaurus, mais aussi d'un protocole standard pour la mise en ligne des reproductions qui est le protocole IIIF pour le partage des images.

(D) L'objectif est de définir des schémas et des outils d'encodage et de publication communs. Les premiers travaux en ce sens dans le cadre de Biblissima ont abouti notamment au Corpus d'inventaires anciens de livres manuscrits et imprimés *Thecae* dont vous pouvez consulter les

5 ontologie et thesaurus : https://doc.biblissima.fr/ ; référentiels : https://data.biblissima.fr/w/Accueil#R.C3.A9f.C3.A9rentiels_disponibles

premières réalisations en ligne[6] : ici le début de l'édition électronique de la section des manuscrits du Vatican dans la Bibliotheca Bibliothecarum de Bernard de Montfaucon : ce corpus a été mis en ligne grâce à un travail de modélisation des données des inventaires anciens et de mutualisation du schéma d'encodage pour ces textes, après quoi seulement le texte a pu être encodé, puis validé techniquement et scientifiquement, puis publié.

Ces travaux de collaboration sont appelés à continuer dans le cadre de Biblissima+, qui met l'accent sur les chaînes d'outils numériques et sur l'interopérabilité des données[7], et dont un groupe de travail est consacré à un laboratoire d'édition de sources.

Pour conclure sur l'interopérabilité, on est un peu revenu sur l'idéal d'interopérabilité des données elles-mêmes, pour concentrer les efforts, de manière plus réaliste, sur la mise au point de ressources intermédiaires qui permettent d'espérer à terme une forme d'interopérabilité : notamment ce que j'ai évoqué donc l'utilisation de références standard : thesauri, référentiels, ontologies ; l'utilisation d'outils d'encodage avec un schéma commun pour le noyau dur des commandes, mais aussi des petits programmes qu'on appelle des API ou Interfaces de programmation, qui permettent à un programme que l'on est en train de développer, d'exploiter les fonctionnalités d'un autre programme qui gère ou traite des données de la recherche, et de récupérer des données dans un format que l'on va pouvoir utiliser au sein de son propre projet.

Métadonnées et fair data

(D) Enfin je vais évoquer la notion de données fair ou fair data. Depuis quelques années a émergé dans la communauté scientifique internationale, et pas seulement celle des sciences humaines et sociales, la notion de données FAIR[8] dont les principes ont été énoncés en 2016.
Mais avant cela je dois aussi expliquer le concept de métadonnée puisque elles sont une partie importante des critères FAIR.

Qu'entend-on par métadonnée ? Une métadonnée est une donnée qui décrit une autre donnée. On peut les décrire comme un ensemble structuré d'informations décrivant une ressource donnée. Elles peuvent être externes, comme une notice d'un catalogue, ou internes, comme un ensemble d'éléments dans un document numérique et par exemple, si on parle de la TEI, les métadonnées du document sont contenues dans l'élément *teiHeader*, l'en-tête du fichier.

(D) Des standards de métadonnées ont été créées parmi lesquels le principal pour les données de sciences humaines est sans doute le Dublin Core, dont la norme a été définie en 1999 autour d'un noyau de 15 éléments de base parmi lesquels créateur, titre, date, format, langue etc, qui sont les descripteurs minimaux d'une ressource.

(D) Alors que sont les données FAIR ? FAIR est l'acronyme pour :

- Findable : faciles à trouver
  - métadonnées de qualité, avec des identifiants uniques et pérennes
  - que l'on puisse retrouver via des moteurs de recherche, utilisant des index, mots-clé
  - qui soient conservées grâce à un archivage sécurisé
- Accessibles :

6 https://www.unicaen.fr/services/puc/sources/thecae//accueil
7 https://projet.biblissima.fr/fr/actualites/biblissima-observatoire-cultures-ecrites-argile-a-imprime
8 https://www.force11.org/group/fairgroup/fairprinciples

    - accessibilité des métadonnées grâce à l’identifiant au moyen d’un protocole de communication standardisé qui soit ouvert et libre
    - les métadonnées sont accessibles même si les données elles-mêmes ne sont pas ou plus en accès libre
    - les conditions d’accessibilité doivent être bien visibles
- Interopérables :
    - les (méta)données doivent utiliser un langage standard et accessible
    - les (méta)données utilisent des vocabulaires qui suivent les principes FAIR
- Re-usable : réutilisables
    - les (méta)données sont décrites avec de nombreux attributs
    - elles sont publiées avec une licence d’utilisation claire
    - la publication mentionne leur provenance