AKpédia

Un projet porté par Catherine Lenoble, <stdin>, éditions HYX.

Résumé

AKpedia est une proposition pour le concours Sémanticpédia qui s'inscrit dans le projet *Anna K, mythologies personnelles,* une fiction développée à partir d'une collection de *data* autour de la vie et l'œuvre de la romancière britannique Anna Kavan (cf. Annexe).

Cette proposition vise le développement d'outils et méthodes complémentaires pour aider à la création ou à la complétion d'articles Wikipédia reposant sur :

- 1. l'exploration du web sur Anna Kavan;
- 2. l'extraction de ces savoirs et la production de métadonnées et de code wiki par des agents logiciels ;
- 3. sur la production de pages Wikipédia, optimisée pour la réutilisation par DBPédia, à l'occasion de la mise en place et l'animation d'un wiki sprint.

Contexte

L'objet du concours est la réutilisation des données de Wikipédia, hors le problème que nous rencontrons est le faible développement de la page francophone d'Anna Kavan. De manière générale, beaucoup de pages sont incomplètes et représentent un potentiel contributif surtout dans le domaine de la littérature et des femmes.

À travers ce concours, la fondation Wikimédia France et le ministère de la Culture et de la Communication souhaitent favoriser la ré-utilisation de l'immense base de connaissances constituée par les contributeurs de Wikipédia. En effet, Wikipédia est devenu incontournable grâce à la qualité de ses articles, soutenue par des dispositifs contributifs innovants (chartes, guides d'écritures, gestion des versions, discussions, robots, etc.). Si Wikipédia brille dans de nombreux domaines souvent délaissés dans les encyclopédies traditionnelles (articles scientifiques, culture populaire, etc.), certains articles, qui mériteraient pourtant d'être mis en valeur, sont manquants ou incomplets.

C'est ce à quoi nous avons été confronté avec le projet *Anna K. mythologies personnelle*s, lors de la consultation de la page de l'auteure. Le travail d'Anna Kavan, auteure culte outre-Atlantique, est étrangement peu étudié par le monde académique, qui reconnaît pourtant l'importance de sa contribution à la littérature de XX° siècle. Malgré ce vide académique, son travail et sa personne font l'objet d'un grande nombre de commentaires sur le Web, en particulier à travers des blogs de fans et critiques amateurs, dont la relecture révèle des changements de « valeur littéraire » (telle ou telle œuvre ayant gagné en popularité au fil des années).

Dès lors se pose la question : comment valoriser un contenu qui n'existe pas encore sur Wikipédia, ou bien de manière très partielle, mais qui pourtant existe de manière fragmentée sur le web ?

Objectifs

Nos objectifs sont les suivants :

- favoriser la complétion et création de pages pour les auteurs manquants de visibilité, en particulier les femmes ;
- explorer les sources « mineures » comme sources faisant autorité sur le sujet ;
- faciliter l'ébauche de pages Wikipédia de manière semiautomatisées, structurées pour pouvoir enrichir DBpédia.

Notre proposition consiste à mettre en place, explorer et expérimenter différentes méthodes complémentaires pour aider à la création ou à la complétion d'articles Wikipédia. Ces méthodes reposent d'une part sur l'extraction et la production de métadonnées et de code wiki par des agents logiciels, d'autre part sur l'utilisation de ces contenus lors d'un *wiki sprint* favorisant la participation et notamment l'inclusion de femmes au processus d'édition et d'écriture collaborative, à l'instar des Art & Feminism Edit-a-thon dont le premier a eu lieu le 1er février 2014, ou l'Edit-a-thon organisé par Catherine Lenoble pour célébrer l'Ada Lovelace Day 2013.

Développement d'outils

AAsniff

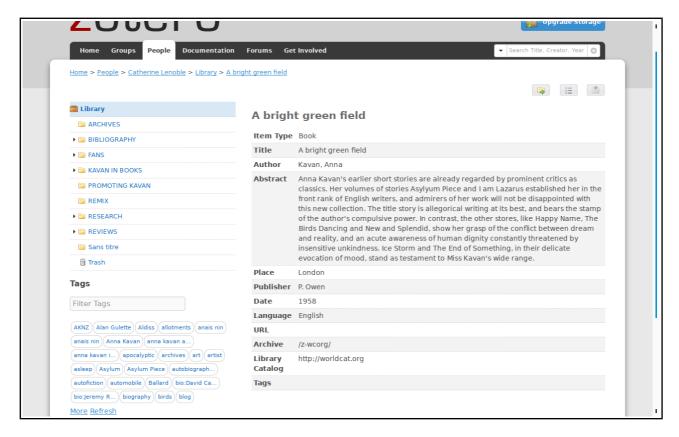
AAsniff est une bibliothèque logicielle écrite en python (sous licence libre AGPL), actuellement en cours d'écriture, mais déjà fonctionnelle. Initiée dans le cadre du projet Active Archives, AAsniff permet de produire des métadonnées RDF à partir de ressources web (pages web, images, sons, vidéos, PDF, etc.). AAsniff repose sur une chaîne d'agents logiciels spécialisés qui s'activent ou non selon la ressource HTTP analysée. Typiquement, la condition d'activation est basée sur le format de la ressource mais peut s'appuyer sur des critères plus complexes (par exemple, une page Youtube ou Archive.org activera des agents supplémentaires dédiés). Un effort particulier porte sur la ré-utilisation de logiciels existants (notamment des outils UNIX comme lmagemagick, ffmpeg, etc.) et l'aisance de la création de nouveaux agents.

INTL http://video.constantvzw.org/LGRU/toonloop.ogv		
OK RELOAD		
media:audio_bits	media:audio_channels	media:audio_codec
<u>16</u>	stereo	vorbis
media:audio_sampling_rate	media:bit_rate	media:duration
44100	499	709.08
media:framerate	media:height	media:video_codec
25.0	512	theora
media:width	http:content_length	http:content_type
<u>640</u>	<u>297244225</u>	video/ogg
http:etag	http:hostname	http:last_modified
"dc000d-11b79641-4ef75a21f1240"	video.constantvzw.org	<u>2014-01-08</u>
Links		
aa:sniffer		
http://example.com/ffmpeg/media/770/		

Exemple de AAsniff: métadonnées extraites d'un fichier vidéo sur le web.

Nous souhaiterions développer de nouveaux agents en partant de la spécificité du cas Anna Kavan. Nous avions utilisé jusqu'à maintenant Zotero, qui propose aussi l'extraction automatique de données des pages collectées mais de manière limitée et conventionnelle (auteur, titre et date).

Zotero semble ne s'appuyer que sur les balises Meta pour extraire les données des pages Web. Nous souhaiterions analyser plus finement les ressources collectées et tenir compte de leurs spécificités pour en tirer plus d'informations (structures spécifiques à la plate-forme de blog, sémantique HTML, microdatas, RDFa, etc.)



Métadonnées récupérées par Zotero.

Au delà de l'extraction de métadonnées, nous souhaiterions expérimenter leur production, notamment avec la librairie Python NLTK (Natural Language ToolKit) pour pouvoir par exemple repérer les noms propres ou des couples de mots particuliers dans des billets de blogs (cf. Annexe).

Aabrowser

AAbrowser est une interface web complémentaire à AAsniff, dont le but est d'explorer le RDF produit, d'en étudier la provenance (quadruplets, contexte) et d'ajouter manuellement, amender ou uniformiser les données. C'est en quelque sorte un wiki pour les données sémantiques. Ce projet, très partiellement entamé, nous permettrait de parcourir les données indexées et de repérer des patterns pertinent à l'élaboration de pages Wikipédia.

L'Oracle

Comment les données RDF collectées peuvent-elle supporter, contredire ou nuancer des intuitions quant à la vie et l'œuvre même d'Anna Kavan et la lecture qui en est faite par ses lecteurs, ses fans, ses critiques ? Peuvent-elles nous permettre de découvrir cette production sous une nouvelle lumière ? Plusieurs questions nous viennent à l'esprit, notamment :

- Dans quelles langues lit-on Kavan ?
- Dans quelles villes lit-on Kavan ?
- Quelle est l'œuvre de Kavan la plus discutée sur Internet ?
- Quelle est la citation de Kavan la plus populaire sur

Internet ?
Quels sont les écrivains à qui Kavan est associée ?
Quels sont les mots-clés que l'on associe le plus à Kavan ?
Anna Kavan est-elle la sœur de Kafka ?
Kavan a-t-elle remixé Kafka ?
Kavan s'est-elle auto-plagiée ?
Kavan a-t-elle été remixée ?

• Où est l'héroïne dans l'œuvre de Kavan ?

• Qui est l'héroïne dans Neige ?

En recoupant des données extraites à l'aide de Zotero, il devient possible de répondre à un certain nombre de ces questions, sous peine qu'elles soient reformulées dans un langage compréhensible par la machine. Nous traduisons ces interrogations et postulats en langage naturel vers le langage formel SPARQL. Nous pouvons ainsi interroger la base de données sémantique de Zotero, et l'augmenter par l'ajout manuel de références, notes et faits. En retour, nous obtenons des listes de résultats (par exemple : classements des pays/villes des *bloggers*) ou des réponses par l'affirmative ou la négative.

Interrogée dynamiquement à chaque lecture, la base de données s'enrichit au fur et à mesure de nouveaux postulats comme autant de prophéties ou la prédiction de mythologies.

Par exemple la question « dans quelles langues lit-on Kavan? » peut se traduire par « dans quelle langue sont écrit les livres répertoriés dans Zotero? ». On peut apporter une réponse à cela en reformulant ainsi : « Sélectionner tous les items de type "book" et renvoyer leur langue . »

Ce qui nous donne en langage de requête SPARQL :

```
SELECT DISTINCT ?title ?lang
WHERE {
          ?book z:itemType "book" .
          ?book dc:title ?title .
          ?book z:language ?lang .
}""", initNs=dict(z=Z, dc=DC))
```

Ce qui nous donne:

[u'A bright green field, English', u'Book Reviews Anna Kavan, en', u'Guilty by Anna Kavan, en', u'The parson, English', u'My madness: the selected writings of Anna Kavan, English', u'I am Lazarus, short stories, English', u'A scarcity of love: a novel, English', u'Ice, Undetermined', u'Unveiling Anna Kavan, en', u'Who are you?, English', u'Rich get rich, English', u'Julia and the bazooka, and other stories, English', u'Guilty, English', u'My soul in China: a novella and stories, English', u"The horse's tale, English", u'A charmed circle, English', u'The Dark sisters, English', u'The Case of Anna Kavan by David Callard, en', u'Mon \xe2me en Chine - Anna KAVAN, fr', u'Mercury, English', u'Asylum piece, English', u'Goose Cross, English', u'Neige - Anna KAVAN, fr', u'Let me alone, English']

Résultats que nous pouvons transformer en syntaxe Médiawiki:

```
from rdflib import graph, Namespace
DC = Namespace("http://purl.org/dc/elements/1.1/")
```

```
Z = Namespace("http://www.zotero.org/namespaces/export#")
graph = Graph()
result = graph.parse("Zotero-AK.rdf")

book_list = graph.query("""\
    SELECT DISTINCT ?title ?lang
    WHERE {
        ?book z:itemType "book" .
        ?book dc:title ?title .
        ?book z:language ?lang .
    }""", initNs=dict(z=Z, dc=DC))

wikitext = ["== Bibliographie ==", ""]

for r in book_list:
    wikitext.append(u"* ''[[{0}]]'', ({1})".format(*r))

print('\n'.join(wikitext))
```

Voici le résultat de ce script (Aabrowser nous permettrait d'harmoniser les noms de langues):

```
== Bibliographie ==
* ''[[Julia and the bazooka, and other stories]]'', (English)
* ''[[A bright green field]]'', (English)
* ''[[The parson]]'', (English)
* ''[[A scarcity of love : a novel]]'', (English)
* ''[[Guilty]]'', (English)
* ''[[My soul in China : a novella and stories]]'', (English)
* ''[[Goose Cross]]'', (English)
* ''[[Rich get rich]]'', (English)
\star ''[[My madness : the selected writings of Anna Kavan]]'', (English)
* ''[[Mercury]]'', (English)
* ''[[Neige - Anna KAVAN]]'', (fr)

* ''[[Guilty by Anna Kavan]]'', (en)

* ''[[Unveiling Anna Kavan]]'', (en)
* ''[[Let me alone]]'', (English)
* ''[[Mon âme en Chine - Anna KAVAN]]'', (fr)
* ''[[Ice]]'', (Undetermined)
\star ''[[I am Lazarus, short stories]]'', (English)
* ''[[Who are you?]]'', (English)
* ''[[A charmed circle]]'', (English)
* ''[Book Reviews Anna Kavan]]'', (en)
* ''[[The Dark sisters]]'', (English)
* ''[[Asylum piece]]'', (English)
* ''[[The Case of Anna Kavan by David Callard]]'', (en)
* ''[[The horse's tale]]'', (English)
```

Ébauches Médiawiki

À partir de l'Oracle, nous souhaitons créer un outil pour générer des ébauches Médiawiki à partir des informations disponibles dans la base de données RDF (fournie par Zotero et AAsniff). Les gabarits alimentés par des résultats de requêtes SPARQL génèrent du code Médiawiki que l'on peut transposer dans Wikipédia comme ébauche servant de base à compléter manuellement par la communauté de Wikipédia.

Propositions de contenu généré:

- bandeau (« ceci est une ébauche générée »)
- infobox
- gabarits de pages (sections principales)
- · bibliographie
- œuvres principales
- citations (celles les plus citées sur le web)
- liens externes/internes
- nouvelles entrées (création de nouveaux articles à partir de sa bibliographie)

Voici un bandeau signalant que l'ébauche a été générée automatiquement :



Ce <u>modèle</u> est une <u>ébauche</u> concernant le livre <u>Neige</u>.

Cette ébauche a été générée à partir de données RDF sur ce sujet. Vous pouvez partager vos connaissances en complétant ces informations.

Voici ce que pourrait être une ébauche automatique du livre *Nouvelles d'une vie : Julia et son bazooka*, à partir des données RDF collectées :

```
{{ébauche|livre}}
     {{titre mis en forme|''Nouvelles d'une vie: Julia et son
bazooka''}}
     {{Infobox Livre
     | titre = Nouvelles d'une vie: Julia et son bazooka
     | auteur =[[Anna Kavan]]
     | préface =
     | directeur =
     | genre =
     | dateparution = [[1990 en littérature|1990]]
     | pays = {{Royaume-Uni}}
     | lieuparution = Paris
     | éditeur = [[Éditions Complexe]]
     | traducteur = [[Claire Malroux]]
     | série =
     | collection =
     | type_média =
     | pages = 210
     | isbn = 978-28702-734-25
     | lieuparution_orig = Londres
     | langue = anglais
     | éditeur_orig = Peter Owen
     | dateparution_orig = 1970
     | isbn orig = 978-07206-034-08
     | précédent = [[Neige]]
     | suivant = [[Laissez-moi ma solitude]]
     } }
      ''''Nouvelles d'une vie : Julia et son bazooka'''' est un
[[Recueil de nouvelles|recueil de nouvelles]] écrit par [[Anna
Kavan]] et paru en [[1990 en littérature|1990]] aux [[Éditions
Complexe]].
```

Exemple d'infobox générée pour le roman Neige:

```
{{ébauche|livre}}
     {{titre mis en forme|''Neige''}}
     {{Infobox Livre
     | titre = Neige
     | auteur =[[Anna Kavan]]
     | préface =
     | directeur =
     | genre = [[Science-fiction]]
     | dateparution = [[2013 en littérature|2013]]
     | pays = {{Royaume-Uni}}
     | lieuparution = Paris
     | éditeur = [[Camourakis]]
     | traducteur = [[Ronald Blunden]]
     I série =
     | collection = Littérature
     | type_média =
     | pages = 208
     | isbn = 978-23662-404-98
     | lieuparution_orig = Londres
     | langue = anglais
     | éditeur_orig = Peter Owen
     | dateparution_orig = 1967
     | isbn\_orig = 978-07206-034-08
     | précédent = [[L'Oiseau qui es-tu ?]]
     | suivant = [[Julia et son Bazooka]]
     } }
     ''''Neige'''' est un [[roman (littérature)|roman]] écrit
par [[Anna Kavan]] et paru en [[2013 en littérature|2013]] aux
éditions [[Cambourakis]].
```

Résultats

Si le projet éditorial est ambitieux, nous pouvons cependant promettre des réalisations concrètes au niveau des outils :

- design et expérimentations logicielles sous licences libres (GPL/AGPL)
- proposition de méthodes et bibliothèques réutilisables
- documentation et valorisation de la démarche sur nos espaces de travail respectifs (blogs, wikis)
- organisation et animation d'un wiki sprint

Wiki sprint « Héroïnes »

Le web sémantique est « un web de données qui peuvent être traitées directement et indirectement par des machines pour aider leurs utilisateurs à créer de nouvelles connaissances » selon l'expression consacrée par Tim Berners-Lee.

L'une des principales portes d'entrée pour quiconque lance une requête « Anna Kavan » sur un moteur de recherche est la page Wikipédia anglophone (traduite par ailleurs en français, allemand, espagnol, polonais, chinois, japonais, hébreu et russe). Or, les informations contenues sur cette page sont loin d'être à jour et pourraient faire l'objet de modifications : améliorations (structuration de l'article, ajout/recherche de sources visuelles en Creative Commons), nouveaux développements (extension de la biographie), ajouts de références (liens internes Wikipédia, citations, ressources externes) et nouvelles entrées (création de nouveaux articles à partir de sa bibliographie).

Aujourd'hui, la page rassemble une <u>Biographie</u> (cinq paragraphes), une rubrique <u>Influence</u> (un paragraphe), une <u>Bibliographie</u> ainsi qu'une liste de <u>Liens externes</u>. Une nouvelle structuration de l'article Anna Kavan sur Wikipédia pourrait s'organiser à titre d'exemple de la façon suivante:

- 1. Biographie
 - 1.1 Jeunes années
 - 1.2 Mariages et débuts littéraires
 - 1.3 Addiction
 - 1.4 Changement de nom
 - 1.5 Deuxième Guerre Mondiale
 - 1.6 Voyages
 - 1.7 Peinture
 - 1.8 Dernières années
 - 1.9 Mort de Kavan
- 2. Œuvres
 - 2.1 Romans et recueils de nouvelles
 - 2.2 Romans et recueils de nouvelles posthumes
 - 2.3 Autres textes
- 3. Héritage
- 4. Analyse de l'œuvre
 - 4.1 Influences reçues
 - 4.2 Style et structure narrative
 - 4.3 Thèmes
 - 4.4 Critiques
- 5. Autour d'Anna Kavan
 - 5.1 Musique
 - 5.2 Théâtre
 - 5.3 Arts visuels
 - 5.4 Blogosphère et réseaux sociaux
 - 5.5 Recherche universitaire
- 6. Iconographie
- 7. Bibliographie
- 8. Notes et références
- 9. Liens externes

Le principe de Wikipédia étant par essence ouvert et collaboratif, il s'agit ici non pas d'avancer seuls mais bien au contraire de mettre en place un

temps d'atelier, façon *wiki sprint*, pour réunir *on-* et *off-line* des participants ayant une connaissance du sujet.

Un wiki sprint consiste à regrouper durant un court laps de temps, une équipe de contributeurs motivés, qui unissent leurs efforts de manière concertée afin de parvenir à un ou des articles aussi aboutis que possible. L'objectif est d'améliorer de manière substantielle, dans plusieurs langues, l'article sur Anna Kavan qui souffre de lacunes affectant une large partie des femmes écrivains sur Wikipédia. La littérature et plus largement la place des femmes et des contenus dédiés aux femmes est un véritable enjeu : 90 % des contenus sur Wikipédia sont édités par des hommes selon une étude de Wikimédia de 2012¹.

Suite aux diverses interviews et visites de terrain, Catherine Lenoble a pu rencontrer des personnes-clés, déjà informées et favorables à cette initiative. La méthode sur laquelle repose cette proposition est celle du *book sprint*, dérivé du *code sprint* pratiqué dans la communauté du logiciel libre, et adapté (notamment par l'artiste néo-zélandais Adam Hyde et son projet FLOSS Manuals) à la production écrite et culturelle.

Portée par l'expérience en octobre dernier d'un atelier d'édition et d'écriture collaborative animé par Catherine Lenoble pour la manifestation internationale Ada Lovelace Day (*ALD2013 edit-a-thon*) et dans le cadre du concours Sémanticpédia, il s'agit de mettre en place selon les mêmes principes, un atelier d'édition et d'écriture collaborative sur Anna Kavan à l'automne prochain à Paris (partenariat en cours avec la BnF), en amont de la diffusion de la création en ligne.

En ce sens, cette proposition de *wiki sprint* permet d'augmenter le projet DBpédia et de mettre en valeur la ré-exploitation de données ayant trait au patrimoine littéraire au féminin et produites collectivement.

Scénario

Événement

- Date: automne 2014, deux journées, de 9h à 13h et de 14h à 17h
- Lieu: Bibliothèque Nationale de France (sous réserve)
- Coût:gratuit
- Co-animation: Catherine Lenoble, <stdin>, volontaire Wikimédia France, équipe BnF.
- Participants: l'événement est ouvert à toutes et tous, aucune expérience d'édition sur Wikipédia n'est demandé, un temps de découverte des fondamentaux en publication/écriture Mediawiki sera proposé dans la première matinée et des tutoriels seront mis à disposition. La participation des femmes est particulièrement encouragée.
- Inscription et informations : par email et en ligne sur une page Wikipédia événement dédiée.

Invitations et participation en ligne

Personnes ressources autour d'Anna Kavan: Victoria Walker (fondatrice Anna Kavan Society), Colin Gale (archiviste), Marc Carlson (archiviste), Emily Hill (journaliste), Nate Dorr (blogger), Matthew Weinstein (artiste), Barbara Woodman (doctorante).

Auteures : An Mertens (Belgique), Lola Robles (Espagne), Nina Allen (Grande-Bretagne), Mary Ockerbloom (USA) - liste en cours ; certaines sont déjà éditrices sur Wikipédia.

Références

Catherine Lenoble | http://litteraturing.net

Catherine Lenoble explore l'écriture dans un environnement numérique, à l'épreuve du réseau, du temps réel, de l'évaporation de la notion d'auteur. L'ubiquité est le héros de son premier roman *petitBain*, une méta-narration publiée aux Éditions à la Criée en 2010 et déployée en ligne sur le site *http://petitbain.net*.

Depuis 2012, elle développe des propositions autour de la lecture et de l'écriture numérique (cycle de rencontres pour le Centre de Ressources du Livre en Pays de la Loire), des interventions croisant écriture collaborative et espace public à travers des ateliers de création et résidences (3D Fiction à Dakar, Futurs Écrits à Orléans, Ctrl-Alt-Kids à Tremblay-en-France) et des espaces d'expérimentation autour de la littérature et des algorithmes (groupe de recherche Algo-lit).

En 2013, le site http://petitbain.net fait partie des œuvres retenues dans la galerie virtuelle de l'exposition Littératures numériques d'hier à demain inaugurée le mardi 24 septembre 2013 au Labo BNF: http://gallery.dddl.eu/fr/accueil/

Stéphanie Vilayphiou et Alexandre Leray | http://stdin.fr

Stéphanie Vilayphiou et Alexandre Leray travaillent au sein de <stdin>, un studio bruxellois de design graphique qu'ils ont initié en 2009. Ils sont également membres du collectif Open Source Publishing. Leur pratique combine la programmation et une approche visuelle du design, pour l'écran comme pour l'imprimé. Leurs principales préoccupations sont la pensée du design à l'époque du numérique, les pratiques de lecture et d'écriture, les software studies et la culture libre.

Ils sont les auteurs de différents essais sur le design graphique, notamment « Écrire le design – vers une culture du code », article paru dans le numéro 4 du magazine *Back Cover*, et travaillent actuellement à l'édition d'un recueil autour de la relation entre outils, pratiques et culture libre dans le cadre du projet européen Libre Graphics Research Unit (LGRU). Ils collaborent de manière régulière avec Constant, association bruxelloise pour l'art et les nouveaux médias.

L'une de leur dernière création est l'œuvre net.art intitulée *Blind Carbon Copy: la Carte ou le Territoire* sur invitation d'Alessandro Ludovico, curateur de l'exposition *Erreur d'impression: publier à l'ère du numérique* de l'espace virtuel du Jeu de Paume. Cette pièce reconstitue le dernier roman de Michel Houellebecq en cherchant les fragments de ses écrits sur Google Books. Elle s'inscrit dans la continuité de l'œuvre *Blind Carbon Copy* qui questionne les contours juridiques de la propriété intellectuelle en appliquant différents filtres au roman *Fahrenheit 451* de Ray Bradbury.

Éditions Hyx | http://editions-hyx.com

Maison d'édition associative, les Éditions HYX entendent participer, en France et à l'international, au développement d'une édition culturelle prospective. Depuis 1994, HYX publient des ouvrages -pour la plupart bilingues- dans les domaines de la culture numérique, des nouvelles formes d'écritures et de lectures, de l'architecture, de l'art contemporain et de la philosophie.

Les Éditions HYX diffusent et distribuent leurs ouvrages en France auprès de 350 lieux (librairies, centres d'art, universités, écoles...) et s'appuient pour l'international sur Idea Books (Pays-Bas).