Accès en ligne aux données de recherche en sciences sociales : l'exemple de "NESSTAR"

Autor(en): **Hadorn, Reto**

Objekttyp: **Article**

Zeitschrift: **Geschichte und Informatik = Histoire et informatique**

Band (Jahr): **10 (1999)**

PDF erstellt am: **16.07.2024**

Persistenter Link: https://doi.org/10.5169/seals-8100

Accès en ligne aux données de recherche en sciences sociales: l'exemple de «NESSTAR»

Reto Hadorn, SIDOS

Le développement du réseau Internet conduit les archives de données pour les sciences sociales, telles le SIDOS,[1] à élaborer des techniques plus efficientes pour donner accès à l'information sur les données disponibles et pour faciliter l'accès aux données elles-mêmes. NESSTAR,[2] un projet conduit par trois archives de données européennes membres du CESSDA,[3] financé par l'UE, est un bel exemple de cet effort. Il mérite l'attention d'un large public: les instruments en préparation peuvent être mis en œuvre par toute agence diffusant des données.

Les pages qui suivent invitent le lecteur à effectuer une brève incursion dans un domaine en plein développement. Les adresses Web permettront au lecteur intéressé d'approfondir les informations et de consulter des exemples.

Après une brève description des intentions de NESSTAR, quelques paragraphes seront consacrés à la présentation du format de documentation des données mis en œuvre. On reviendra ensuite au dispositif de NESSTAR et à quelques indications utiles pour en faire l'essai. Les références techniques sont là pour bien situer les enjeux.

1. Nesstar – en quelques lignes

NESSTAR est la suite logique d'un premier dispositif mis en place par le CESSDA en 1995 déjà: l'IDC ou Integrated Data Catalogue[4] qui permet d'interroger en ligne en une seule opération les catalogues de données de plusieurs archives. Le réseau Internet est utilisé ici pour donner un accès

1 SIDOS: Service suisse d'information sur la recherche et d'archivage de données pour les sciences sociales, Ruelle Vaucher 13, 2000 Neuchâtel. URL: www-sidos.unine.ch. Pour une présentation générale du service, voir Hadorn, Reto: «Le SIDOS, une archive de données pour les sciences sociales». In*: Histoire et Informatique*, Vol. 9, 1998, pp. 69-78.

2 NESSTAR: Networked Social Science Tools and Resources. URL: http://www.nesstar.org. Si le lecteur a un PC avec une connexion Internet sous la main, il peut consulter ce serveur et faire très rapidement des essais, moyennant l'installation sur sa machine d'un programme client gratuit, disponible en ligne.

3 CESSDA: Council for European Social Science Data Archives. URL: http://www.nsd.uib.no/cessda/.

4 Ce catalogue peut être consulté pour quelques mois encore à l'adresse http://www.nsd.uib.no/cessda/IDC/. Il sera remplacé par NESSTAR dans le courant de l'an 2000.

centralisé à une information distribuée. Les catalogues sont gérés par les archives sur leur propre serveur, le formulaire d'interrogation est unique.

L'IDC présente les informations généralement présentes dans un catalogue: informations de type bibliographique (auteur, titre, dépôt, conditions d'accès), complétées par une description plus ou moins sévèrement résumée du contenu thématique des données (nature du projet de recherche et principaux thèmes abordés dans l'enquête par exemple). Or, les archives disposent la plupart du temps d'informations bien plus détaillées, et ceci sous forme informatique: codebooks, labels de variables et de valeurs incorporés aux fichiers SPSS ou SAS; pourquoi ne pas donner accès à ces informations en ligne au lieu de les distribuer sur disquette?

Par ailleurs, la transmission des jeux de données aux utilisateurs passe de plus en plus par le réseau (attachement à un message électronique, transfert par FTP depuis un serveur dédié): il est logique de chercher à simplifier et automatiser les manipulations qu'impliquent ces techniques, de manière à réduire le travail de préparation du côté des archives.

Le rêve conduit plus loin encore: est-il bien nécessaire de disposer des données en local? Comment les exploiter si l'on ne dispose pas d'un programme d'exploitation statistique? En les exploitant à distance sur le serveur de l'archive …

Ainsi se constitue une première idée de ce que fait NESSTAR; ce système:

– donne accès par le biais d'une interface unique à l'information sur les jeux de données disponibles dans l'ensemble des archives membres du réseau;
– donne accès à l'information détaillée sur les données jusqu'au niveau des variables;
– permet de télécharger des jeux de données entiers, d'un choix de variables ou d'une sélection de cas;
– permet l'analyse statistique à distance et la production des représentations graphiques associées.

L'accès aux données est-il pour autant totalement libre? Non: toutes les archives ont dans leurs fonds des jeux de données dont l'accès est soumis à des conditions plus ou moins restrictives, le plus souvent définies par les producteurs des données. NESSTAR gère donc aussi les droits d'accès.

2. Un nouveau format de documentation: le codebook du DDI au format XML

Le premier enjeu pour NESSTAR a été de trouver un format de documentation des données qui puisse être produit à partir des matériaux existants et qui soit reconnu par les serveurs et les browsers Web. Sur ce point, le projet a pu s'appuyer sur deux développements récents, la définition d'un nouveau format de codebook basé sur SGML, qui lui-même profite du développement d'une version réduite de SGML pour le Web: XML.

Points de repère

Traditionnellement, l'information détaillée au niveau des variables est présentée sous la forme de codebooks, c'est-à-dire de livres, imprimés ou électroniques, qui présentent, variable après variable, les informations suivantes: nom et étiquette de la variable, formulation de la question, codes et signification des codes, éventuellement la distribution des fréquences, des remarques d'ordre méthodologique sur les conditions d'exploitation de la variable. Le codebook est souvent complété par une introduction qui décrit dans le détail les méthodes de collecte des données (population de référence, échantillonnage, non-réponse etc.).

Les archives de données ont entrepris dès les années 70 de produire des codebooks électroniques (imprimables), produits à partir des fichiers de données selon une procédure qui facilite la correction d'erreurs dans les fichiers et la diffusion de la documentation. Elles se sont appuyées pour cela sur le format de codebook proposé par un software d'exploitation statistique courant à l'époque, OSIRIS. Ce programme est depuis tombé en désuétude, sans que les particularités qui faisaient sa force pour la préparation de codebooks soient reprises par les programmes qui tiennent actuellement le haut du pavé, SPSS et SAS. Ces derniers permettent bien d'intégrer des éléments d'information tels que labels de variables et de valeurs, mais obligent qui veut produire un codebook complet, avec questions et remarques méthodologiques, à un bricolage sans lendemain: combien d'assistants n'ont-ils pas passé des heures et des jours à insérer les questions dans les tableaux de fréquences, pour ensuite recommencer si des corrections devaient être apportées aux données?

En 1995, un groupe de travail a été constitué dans le cadre de IASSIST[5], avec pour mandat d'élaborer un nouveau format de documentation des jeux de données:

---

5 IASSIST: International Association for Social Science Information Service & Technology, URL: http://datalib.library.ualberta.ca/iassist/

– propre à réactualiser le standard des années septante;
– indépendant des programmes d'exploitation statistique;
– mobilisant les technologies actuelles tout en restant suffisamment basique pour être largement accessible pour des utilisateurs aux moyens toujours limités.

Le groupe de travail, connu sous l'étiquette "Data documentation initiative" ou DDI[6], a élaboré un premier projet sur la base de SGML. La définition par le W3C d'une version réduite de SGML pour le Web, XML, intervenue en cours de projet, a ouvert la porte à de nouvelles applications, dont NESSTAR est précisément un exemple.

XML – La technique

SGML[7], Standard generalised markup language, est un langage générique (ou générateur de langages) pour la structuration et la présentation des informations dans un document. Son application la plus largement connue est certainement le HTML, Hypertext Marking Language, le langage qui permet aux serveurs Web de communiquer avec les browsers tels Internet Explorer ou Netscape.

HTML donne cependant une idée un peu courte du potentiel de SGML. HTML est une application centrée sur la mise en forme de documents textuels et même sous cet angle, il montre très vite ses limites. L'intérêt majeur de SGML est de *générer des langages structurant l'information*, par la définition d'éléments caractérisés par des attributs et reliés entre eux par des relations hiérarchiques. C'est de cette propriété structurante que le groupe de travail du DDI a fait usage lorsqu'il a développé le premier format de codebook sur une base SGML. C'est encore la capacité de structurer l'information (et pas seulement de la mettre en page) qui intéresse les nombreux utilisateurs de SGML qui produisent avec cet instrument des bases de données de pièces détachées comme des bibliothèques électroniques.

XML[8] est la réponse du W3C aux Netscape et autres Microsoft qui n'ont jamais voulu intégrer dans leurs browsers les instruments nécessaires à la consultation sur Internet des informations enregistrées en format SGML. Il conserve l'essentiel des fonctionnalités du SGML dans une spécification beaucoup plus courte – donc maîtrisable. Le propos de XML est de rendre utilisable dans des applications Web les fonctions structurantes de

6 URL: http://www.icpsr.umich.edu/DDI/

7 URL: http://www.oasis-open.org/

8 XML: eXtended markup language; URL: http://www.w3.org/XML/

SGML. Comme SGML, XML est un générateur de langages qui permet de définir des structures d'information appropriées à diverses applications.

De petits producteurs de software sont dès lors en mesure de développer les ressources en question; aussi les grands suivent-ils le mouvement: Microsoft annonce une nouvelle génération de softs capables de générer des versions XML des données traitées et Internet Explorer5 est en principe en mesure d'afficher des informations au format XML.

Le codebook du DDI – le moule

Le groupe de travail de IASSIST chargé de définir le nouveau format de codebook (the Data Documentation Initiative – DDI) a donc converti son codebook SGML au standard XML. Le format de codebook est ainsi une application particulière du langage XML. Les types d'information attendus dans un codebook, leurs relations et leurs propriétés sont définis dans un document qui porte, en anglais, le nom de Document Type Definition (DTD). Il est dès lors possible de programmer des applications capables de reconnaître, gérer et distribuer toute information conforme aux définitions données dans le DTD du DDI.[9]

Le codebook du DDI comporte les chapitres suivants: description du codebook, description du projet qui a donné lieu au relevé de données, description du fichier de données, description des variables, documentation complémentaire. Chacun des chapitres comprend plusieurs éléments, parfois eux-mêmes précisés par des sous-éléments. Aux éléments sont associés des attributs; les relations logiques entre éléments sont de type hiérarchique (structure en arbre).

Le format type défini par le DDI n'est qu'une structure, c'est-à-dire un ensemble de règles concernant des types d'éléments, des types de rapports entre ces éléments et les propriétés de ces éléments. Il ne devient réellement utile que s'il est reconnu par des applications capables de l'interpréter. A cet égard, NESSTAR présente un intérêt tout particulier, puisque c'est la toute première application grandeur nature du nouveau format de codebook.[10] D'autres applications seront certainement développées dans le futur – on pense notamment à des programmes permettant aux chercheurs d'éditer eux-mêmes des codebooks, dont le

---

9 Le DTD pour le codebook peut être consulté sur le site du DDI (http://www.icpsr.umich.edu/DDI/CODEBOOK.TXT) de même qu'un schéma simplifié qui en donne une vue d'ensemble (http://www.icpsr.umich.edu/DDI/ddischem.html) et la définition explicite des éléments et de leurs propriétés (http://www.icpsr.umich.edu/DDI/codebook/codedtd.html).

10 Qui veut avoir un aperçu de l'utilisation du format défini par le DDI fera donc un essai de NESSTAR (voir ci-dessous), tout en tenant compte du fait que les exemples qu'on peut y trouver en ce début d'an 2000 ne donnent qu'une petite idée du potentiel du nouveau format de documentation.

caractère standardisé facilitera l'échange, la correction et l'importation dans diverses bases de données – dont celles des archives de données, naturellement.

Le format proposé par le DDI a subi en 1999 un test étendu, auquel l'équipe de NESSTAR a participé. Une première version officielle est attendue pour le début de l'an 2000. Cela dit, il est d'ores et déjà évident que l'effort entrepris devra être poursuivi, notamment parce que le nouveau format n'est pour le moment utilisable que pour décrire les jeux de données les plus simples: les données d'enquêtes transversales non répétées; une deuxième phase de développement est planifiée, qui vise notamment à étendre le format aux jeux de données complexes (données hiérarchiques, séries temporelles, panels etc.).

3. NESSTAR: un réseau de serveurs au service d'une multitude de clients

Les programmes constituant NESSTAR comprennent deux composantes principales: le serveur, Nesstar Publisher, et le client (browser), Nesstar Explorer.

Chaque distributeur de données intègre données et documentation dans un serveur NESSTAR et gère localement l'ensemble des informations mises à disposition (données, catalogue, information détaillée). Par le biais d'Internet, plusieurs distributeurs de données utilisant NESSTAR peuvent être réunis en un réseau. La localisation des distributeurs de données est indifférente, un réseau mondial aussi facile à réaliser qu'un réseau local.

L'utilisateur de NESSTAR installe sur son ordinateur le programme client et se connecte à un réseau de distributeurs de données.[11] Il a dès lors accès de manière simultanée à l'information disponible chez tous les distributeurs du réseau. Ainsi NESSTAR apparaît-il comme une *archive de données virtuellement internationale*, qui donne un accès global à des fonds entretenus localement par chaque archive.

Dans le premier catalogue international, l'IDC, toute l'information substantielle sur un jeu de données était placée sous une rubrique unique. Les catalogues des données proposés par les archives sur leur serveur local limitent eux aussi les possibilités d'interrogation à quelques champs tels le titre, le résumé du projet et le contenu thématique du jeu de données. Avec NESSTAR, la recherche des jeux de données pertinents peut faire usage de

---

11 Pour le moment, seul le réseau constitué autour du CESSDA pour le test final du dispositif est actif. Avec le temps, il est probable que d'autres réseaux se constituent.

toutes les rubriques définies dans le format standard du DDI, notamment les formulations de questions et les descriptions de variables. Les méthodes de collecte ou les modes d'échantillonnage pourront également être utilisés comme critères. La recherche d'un jeu de données s'appuie donc sur des critères bien plus précis que précédemment.

Lorsque l'utilisateur identifie un jeu de données intéressant, il peut accéder à l'ensemble des métadonnées: consulter la description du projet, la description du relevé des données, la description détaillée de chaque variable. La structure hiérarchique du nouveau format de codebook est interprétée dans NESSTAR par une arborescence analogue à celle qu'on trouve dans les gestionnaires de fichier: l'utilisateur navigue dans les métadonnées en développant ou réduisant le niveau de son choix.

Si les autorisations nécessaires ont été obtenues auprès de l'archive dont on veut obtenir les données, l'utilisateur de NESSTAR peut choisir d'analyser les données à distance ou de les télécharger. L'exploitation à distance est basée sur le programme d'exploitation statistique développé par les archives norvégiennes pour être utilisé dans les écoles; elle est limitée à des techniques simples mais suffisantes pour une première exploration des données. L'analyse à distance a notamment pour objectif de limiter le trafic sur le réseau et de donner des possibilités d'analyse statistique à des utilisateurs qui ne disposent pas des programmes nécessaires. Des représentations graphiques appropriées sont disponibles.

Essayer Nesstar

NESSTAR achève cet hiver une phase de test. La version 1 du client et du serveur NESSTAR est attendue pour la fin du mois de janvier 2000. Pour le moment, seul un petit nombre de jeux de données ont été intégrés au système, dans le cadre des tests de mise au point. La consultation de NESSTAR a donc aujourd'hui un intérêt technique avant tout, pour qui veut suivre le développement de l'outil ou avoir une idée des ressources à venir. Ceci dit, il vaut désormais la peine de se connecter régulièrement au réseau pour suivre son développement. Au milieu de l'année, le catalogue international actuel (IDC) devrait être transféré sur NESSTAR.

Un essai de NESSTAR est possible dès maintenant, moyennant téléchargement du programme client depuis la page www.nesstar.org. L'installation du programme est aisée et les ressources disponibles suffi-

samment «évidentes» pour que l'utilisateur parvienne rapidement à un résultat.[12]

NESSTAR a été développé en priorité pour donner aux archives de données pour les sciences sociales un instrument de distribution qui utilise de manière efficace et rationnelle les ressources apportées par les nouvelles technologies (Internet, Web, Java, XML). Le cercle des utilisateurs potentiels est bien plus grand, ce qui conduit les producteurs de NESSTAR à mettre leur produit sur le marché.[13] On peut donc imaginer que d'ici quelques mois ou années, le péquin à la recherche de données aura le choix entre plusieurs réseaux basés sur NESSTAR.

4. Quelques difficultés

Pour réaliser NESSTAR, il fallait bien sûr tout d'abord … en rêver et ce n'est pas un hasard si les auteurs du projet aiment à parler de leur «social science dream machine». La concrétisation du projet à une échelle internationale et avec des exigences minimales à l'égard des participants n'est cependant pas sans poser quelques problèmes.

La langue

L'anglais fonctionne entre archives comme la langue internationale. Certaines archives autres que l'anglaise produisent des descriptions de jeux de données en anglais, notamment les archives danoise et hollandaise. L'archive allemande a également entrepris de traduire ses descriptions de jeux de données en anglais.

Pour la plupart des autres archives, il est tout simplement impossible d'assurer un tel travail; seule une petite partie de la description est parfois traduite, afin de permettre l'interrogation du catalogue. Par contre, l'information détaillée fournie à l'utilisateur des données (description détaillée du projet, de la méthode et des variables) restera le plus souvent dans une langue nationale autre que l'anglais. Même si NESSTAR permet techniquement de passer de la recherche à l'exploitation sans rupture, la langue peut faire obstacle.

Le groupe NESSTAR est conscient du problème et travaille à un thésaurus multilingue basé sur le thésaurus développé au fil des ans par l'archive

---

12 A noter cependant qu'une machine puissante est recommandée. NESSTAR est particulièrement gourmand en matière de mémoire. S'il tourne avec 64Mb, il ne se sent à l'aise qu'à partir de 128Mb …

13 Si le serveur est payant, le browser (client) restera gratuit.

de données anglaise.[14] Un tel instrument permettrait au moins d'interroger le catalogue de données sur la base de descriptions homogènes. Encore faudra-t-il que les archives concernées investissent dans une indexation appropriée de leurs jeux de données et que le niveau de détail auquel l'indexation a lieu soit relativement homogène et, sur ce plan, l'archive anglaise a mis la barre relativement haut.

L'hétérogénéité des descriptions

Les catalogues de données des diverses archives ont été constitués de manière indépendante, même si la plupart se réfèrent avec plus ou moins de rigueur au standard défini par les premières archives au cours des années 70. Non seulement le niveau de détail est-il très variable d'un catalogue à l'autre, mais en plus il n'est pas toujours simple d'établir une correspondance entre les rubriques existantes et les rubriques proposées aujourd'hui par le DDI. La conséquence est prévisible: la structure d'information proposée par le DDI sera très inégalement remplie et certaines rubriques seront par la force des choses détournées par l'une ou l'autre archive en fonction de l'information dont elle se trouve disposer.

Le catalogue virtuel résultant de l'intégration des divers catalogues nationaux sera donc nécessairement hétérogène quant à l'information disponible. Un jeu de données décrit par le titre et un court résumé aura beaucoup moins de chances d'être repéré lors d'une interrogation du catalogue virtuel qu'un jeu de données qualifié par une description détaillée et une liste de mots clés substantielle. L'interrogation à l'aide des champs très différenciés du codebook du DDI, en soi un instrument remarquable, n'a de sens que si l'on sait exactement comment les différentes archives ont complété chacune des rubriques.

Les archives qui développent aujourd'hui un système d'information interne, par exemple sous la forme d'une base de données relationnelle, tendent tout naturellement à prendre appui sur la structure d'information développée par le DDI. Il faudra certainement du temps – et une volonté politique claire de la part des partenaires concernés – pour que le standard nouvellement défini par le DDI se traduise par un catalogue virtuel d'homogénéité acceptable.

Ces difficultés sont connues des promoteurs du DDI et de NESSTAR. Ils ont voulu que l'utilisation du standard reste peu contraignante parce qu'il leur paraît plus important de permettre à toutes les archives de données de

14 URL: http://biron.essex.ac.uk/searching/zhasset.html

participer, quelles que soient les conditions nationales et les ressources à disposition.

L'illusion de la documentation complète

Nous allons terminer par une remarque plus générale concernant la politique de documentation des jeux de données des sciences sociales. La question est de savoir si la documentation généralement mise à disposition des utilisateurs est suffisante ou non pour une exploitation correcte des données. Cette question a un corollaire: l'utilisateur des données fait-il vraiment usage de l'information mise à disposition?

Le plus souvent, l'information substantielle sur le projet à l'origine des données doit être cherchée dans des rapports ou des publications qui sont simplement référencés dans l'information sur les jeux de données. On peut imaginer qu'une transmission «lente» des données, avec les techniques actuelles, laisse un peu de temps à l'utilisateur des données pour acquérir ces documents. Cela ne veut naturellement pas dire qu'il le fasse. La quasi-immédiateté de l'accès aux données peut malheureusement transmettre le message implicite que la documentation informatique attachée aux données (le questionnaire dans le meilleur des cas, parfois les labels de variables seulement) est suffisante pour une interprétation pertinente des données.

Un chercheur familier d'un domaine de recherche dont il connaît la littérature et les rites peut estimer avoir une compréhension suffisante des données relevées par des collègues pour passer directement à l'exploitation. Les archives de données ne s'adressent cependant pas à une clientèle *d'insiders*. Les autres utilisateurs ont besoin de plus d'information et l'on peut considérer qu'il appartient aussi aux archives de données d'insister là-dessus. Si l'accès aux données est facilité, il faut de même faciliter l'accès à l'information contextuelle sur la recherche.

5. Le SIDOS

Le SIDOS se joindra certainement au réseau NESSTAR des archives de données pour les sciences sociales. Le premier niveau de participation consistera à intégrer son catalogue de données dans le réseau. La traduction en anglais de certaines rubriques permettra la recherche des données suisses dans le catalogue international – une traduction complète de toutes les rubriques ne peut cependant pas être réalisée; l'utilisateur ne recevra donc une information complète que dans une des langues nationales.

Il est prévu que les jeux de données issus d'enquêtes nationales soient intégrés complètement dans NESSTAR, avec les données. Là encore, l'in-

formation détaillée au niveau des variables s'affichera le plus souvent dans une des langues nationales. Il est à prévoir que la plupart des archives de données non anglophones procèdent de manière analogue.

Les progrès réalisés par le SIDOS dans cette direction seront régulièrement signalés sur le site Web de l'archive.

Leere Seite
Blank page
Page vide