Construire les Digital humanities en France. Des Cyber-infrastructures pour les Sciences humaines et sociales (rapport)

Marin Dacos, Jean-Paul Caverni

Construire les *Digital humanities* en France

Des Cyber-infrastructures pour les Sciences humaines et sociales

30 octobre 2009

Jean-Paul Caverni

Président de l'Université de Provence

jean-paul.caverni@univ-provence.fr

Marin Dacos

Directeur du Centre pour l'édition électronique ouverte (Cléo | Revues.org)

marin.dacos@revues.org

La construction de cyber-infrastructures en sciences humaines et sociales est une nécessité pressante. Elle répond à des impératifs majeurs concernant la recherche sur l'homme et sur la société. L'enjeu n'en est pas seulement le rayonnement de la recherche française dans le monde, mais aussi la pérennité de l'accès aux résultats des recherches ainsi que l'apparition de nouveaux paradigmes d'articulation entre le texte scientifique et l'exercice d'administration de la preuve.

Or, depuis plus de 10 ans désormais, l'ensemble des données de la recherche en sciences humaines et sociales est numérique. Ce matériau est, pour l'heure, largement laissé en jachère, soumis aux aléas de la structuration, de la diffusion et de la conservation par chaque chercheur ou par son laboratoire. Les programmes de recherche sont financés pour une durée déterminée, sans politique de conservation ou d'accès concernant les résultats et les données collectées. La fragilité d'un tel non-dispositif est évidente. Il n'existe pas de forte alternative à la mise en place de cyber-infrastructures permettant de gérer ces données, qui concernent autant les données primaires que les résultats de la recherche, les données secondaires que les éléments de démonstration, les identités numériques des chercheurs que les logiciels qu'ils développent.

Table des matières

CONSTRUIRE LES *DIGITAL HUMANITIES* EN FRANCE DES CYBER-INFRASTRUCTURES POUR LES SCIENCES HUMAINES ET SOCIALES

I. Cyberinfrastructures. Faire entrer la recherche dans le paradigme numérique

1. Programmes de recherches, plateformes, très grands équipements

2. Accès, vie des communautés, données, édition

3. Principes de pilotage

4. Assurer la pérennité

II. Cinq grandes infrastructures

1. TGE Adonis

2. SHS 2.0

3. Édition scientifique numérique (ÉSN)

4. CORPUS-SHS

5. PROGEDO

I. CYBERINFRASTRUCTURES. FAIRE ENTRER LA RECHERCHE DANS LE PARADIGME NUMERIQUE

On reprendra à notre compte la définition de cyberinfrastructure proposée par l'American Council of Learned Societies (ACLS) en 2006[1] : *« a layer of information, expertise, standards, policies, tools, and services that are shared broadly across communities of inquiry but developed for specific scholarly purposes : cyberinfrastructure is something more specific than the network itself, but it is something more general than a tool or a resource developed for a particular project, a range of projects, or, even more broadly, for a particular discipline.* »[2]. On le complètera par la définition donnée plus récemment par Françoise Thibault et Philippe Casella : « *Une très grand infrastructure de recherche est un outil établi en vue de mener une recherche propre d'importance et pouvant assurer une mission de service pour une ou plusieurs communautés scientifiques de grande taille. Son coût de construction et d'exploitation est tel que cela justifie un processus de décision et de financement concerté au niveau national, et éventuellement international, et une programme pluriannuel. Son évaluation et sa visibilité sont assurées par des comités scientifiques de haut niveau, le plus souvent internationaux* ». On préfèrera, en revanche, ne pas distinguer les infrastructures pilotées par les chercheurs des infrastructures pilotées par d'autres types de personnels (que nous désignerons comme ingénieurs dans ce texte[3]). La dichotomie entre recherche et accompagnement de la recherche a produit des projets faiblement structurés scientifiquement ou techniquement et ne convient pas à la dimension majeure qui est celle des cyber-infrastructures, dans laquelle une coopération entre les préoccupations de recherche et d'accompagnement de la recherche s'impose au plus haut niveau. Nous appelons donc à une *alliance* entre tous les métiers de la recherche qui permettent de se garder des deux

[1] http://www.acls.org/programs/Default.aspx?id=644

[2] Cité par Pierre Mounier dans « Une cyberinfrastructure pour les sciences humaines et sociales », *Blogo-numericus*, 2007 : http://blog.homo-numericus.net/article130.html

[3] Dans ce texte, ce terme désignera, au sens large, tous les métiers de l'accompagnement de la recherche concernés par le numérique : ingénieurs spécialisés en information scientifique, ingénieurs informatiques, documentalistes, conservateurs, …

écueils du pilotage par la technique ou d'un pilotage « hors-sol ».

Pour faire face à de tels enjeux scientifiques, une discipline a émergé en Angleterre et aux Etats-Unis : les *Digital humanities*. Ce terme désigne la discipline transversale qui qui structure des dizaines de centres de recherches aux USA et qui est en train d'émerger avec force à l'échelle internationale, notamment dans le cadre d'une fédération de centres de recherches appelée « CenterNet »[4]. Ce réseau, créé par le Maryland Institute for Technology in the Humanities (MITH[5]), associe plus de 200 centres de recherches dans le monde. Il organise à l'échelle internationale les centres de Digital humanities, qui sont en général autant des centres de recherches que de services, notamment autour de conférences annuelles « Digital humanities »[6] organisées par « The alliance of Digital humanities »[7], mais aussi autour de projets spécifiques, notamment autour de l'encodage de texte (TEI consortium, né à Oxford[8]), de structuration des bibliographies des chercheurs (Center for history and new media, CHNM, University George Mason, Washington[9]). En France, la discipline se structure également[10].

1. Programmes de recherches, plateformes, très grands équipements

Nous proposons d'adopter la typologie suivante :

1) *programmes de recherches* : il s'agit de programmes de recherche, qu'ils soient régionaux, nationaux ou même internationaux, qui peuvent avoir une grande ampleur problématique, géographique ou chronologique, et mobiliser de nombreux chercheurs ainsi que de nombreuses ressources. Ils portent les problématiques et les innovations scientifiques. Financés en général par projet, ils ne disposent pas d'infrastructures leur permettant de pérenniser leurs méthodes et leurs *corpus* sur le long terme (les laboratoires de recherche n'ont pas cette vocation). Ils publient leurs résultats une fois

[4] http://digitalhumanities.org/centernet/
[5] http://mith.umd.edu
[6] http://www.tge-adonis.fr/?Conference-Digital-Humanities-2009
[7] http://www.digitalhumanities.org/
[8] http://www.tei-c.org
[9] http://chnm.gmu.edu/ , http://www.zotero.org et http://zotero.hypotheses.org
[10] http://www.digitalhumanities.cnrs.fr/

le programme achevé.

2) *plateformes* : les plateformes sont des centres spécialisés en *digital humanities* qui mènent des missions de long terme à forte dimension technologique. À cette échelle, un effort de généricité est nécessaire pour transposer dans la longue durée des projets dont les modalités sont initialement pensées comme des prototypes uniques et spécifiques. Cet effort de généricité doit parvenir à respecter l'intégrité du questionnement scientifique de chaque projet, tout en se préoccupant :

- de réduction drastique des idiômes, grâce à l'adoption de normes internationales,

- de factorisation de l'ensemble des éléments technologiques mutualisables,

- de diffusion la plus large possible, dans le respect d'une politique d'accès contrôlée à chaque fois que cela s'impose (données nominatives, données confidentielles, ...),

- de conservation et d'accès à long terme.

Les plateformes ont donc des compétences fortes en ingéniérie et s'appuient sur des dispositifs puissants, afin de stabiliser l'effort d'innovation scientifique issu des projets de recherches.

3) *très grandes infrastructures*, qui assurent le financement des plateformes et arbitrent sur les priorités stratégiques. Leur rôle est également d'assurer l'interconnexion et la mise en cohérence de l'ensemble des dispositifs des plateformes. Ils s'assurent que l'évolution des plateformes est en phase avec l'évolution des enjeux mis en évidence par la communauté scientifique.

2. Accès, vie des communautés, données, édition

Nous proposons de mettre en place un ensemble de grands équipements, qui seront coordonnés par un *Conseil des grands équipements en Digital humanities,* qui pourrait être rattaché au programme "*e-science*". Ce conseil serait chargé d'assurer la cohérence entre les grands équipements. Ces grands équipements portent sur les quatre dimensions majeures de la recherche en sciences humaines et sociales :

- ***l'accès***, la stabilisation et la mise en relation des données numériques entre elles,

- ***la vie des communautés scientifiques*** (débat scientifique, identité numérique),

- ***les données*** sur lesquelles s'appuient les chercheurs (archives historiques, enquêtes orales, statistiques diverses, données archéologiques...) *mais aussi les méthodes et outils* qui permettent d'en extraire des découvertes scientifiques,

- ***l'édition*** des résultats de la recherche (livres, revues, archives ouvertes).

3. Principes de pilotage

Le pilotage de ces équipements devrait s'appuyer sur quelques principes simples :

1. L'échelle des cyberinfrastructures impose des stratégies très ouvertes, qui associent grands organismes, universités, fondations, etc.

2. Prévoir des coopérations internationales, permettant la mise en place d'une cyber-infrastructure générale, à l'échelle internationale, par le partage d'expertises et la mise en place de spécialisations.

3. Associer la communauté scientifique (chercheurs et ingénieurs) au pilotage, notamment à travers les comités des utilisateurs des plateformes et des très grandes infrastructures.

4. Enfin, la communauté scientifique devra débattre de la question des éventuelles obligations associées aux financements des projets de recherche. Cela concerne en particulier la question du libre accès à la littérature scientifique (*open access*) ainsi que la mutualisation des données collectées (question du *mandat*, ou dépôt obligatoire).

4. Assurer la pérennité

Garantir la pérennité de tels équipements impose de s'appuyer sur l'expertise technologique là où elle se trouve, en imposant le respect de standards ouverts. Les formats de données et les protocoles sont devenus des enjeux de société qui ne peuvent être considérés comme des contingences techniques. Leur normalisation et leur

ouverture doivent être obligatoires.

Il faut également s'appuyer sur une grille puissante, telle que celle qu'élabore actuellement ADONIS au Centre de calcul de l'IN2P3, et une politique d'archivage, telle celle qui est confiée au CINES. L'implantation de ces équipements dans de grands centres universitaires paraît stratégique, afin de les connecter aux communautés scientifiques et de les ancrer dans le maillage institutionnel existant.

Une partie de ces dispositifs pourrait se consolider grâce à des solutions de type *freemium* (le *freemium* donne un accès gratuit aux principaux services, réservant une facturation à certains services complexes).

II. CINQ GRANDES INFRASTRUCTURES

Le chantier qui se trouve devant nous est immense. En revanche, il serait contre-productif de penser que le terrain est totalement vierge. Les travaux des équipes de recherche sont très nombreux et déjà partiellement structurés ; un certain nombre de plateformes existent ou sont déjà préfigurées.

La *Roadmap* française des très grandes infrastructures de recherche (TGIR), parue en 2008, mentionne l'existence du seul TGIR des Sciences humaines et sociales, le TGE Adonis, et propose la création de trois nouveaux TGIR : PROGEDO (Production et gestion de données pour les Sciences humaines et sociales), CORPUS (Corpus pour les Sciences humaines et sociales) et BSN (Bibliothèque scientifique numérique)[11]. Nous reprenons l'essentiel de ces conclusions, en suggérant la création d'équipements intitulés *SHS 2.0* et *Édition scientifique numérique* (la notion d'édition ne devant pas être confondue avec celle de bibliothèque). Nous proposons la convergence de BSN, qui n'est encore qu'un projet, avec le TGE Adonis, leurs missions étant largement convergeantes.

[11] http://www.roadmaptgi.fr/Documents/TGIRs%20en%20STIC%20et%20SHS.pdf

1. TGE Adonis

ADONIS, "Accès unifié aux données et documents numériques des sciences humaines et sociales" est le seul grand équipement existant en SHS aujourd'hui. Il s'agit d'un TGE qui développe un moteur de recherche scientifique, couvrant l'ensemble des types de données qui concernent la communauté scientifique. Il constitue, *de facto*, une bibliothèque numérique de nouvelle génération, fortement distribuée et structurée. Il a pris appui sur le CINES et le Centre de calcul de l'IN2P3 pour se doter d'une architecture numérique solide et mettre en place un programme d'archivage pérenne (OAIS). Son rôle structurant est accentué par la prise en compte des enjeux de l'identification unique des documents scientifiques et de leurs interconnexions (*crosslinking*), enjeux centraux qui ont été largement minorés jusqu'à présent et ont fait l'objet d'une appropriation à dominante économique (DOI, Crossref).

Objet : Réalisation d'une architecture permettant de fédérer autour d'un accès unique l'ensemble des ressources utiles à la recherche.

Partenaires : Toutes les plateformes et centres de recherches relevant des Digital humanities, bibliothèques universitaires.

Usagers : Chercheurs, étudiants, enseignants, citoyens, journalistes.

Enjeux : Accès, bibliothèques numériques, indexation, Identification unique, Crosslinking, visibilité internationale, Archivage garantissant l'accès à long terme.

Maîtrise d'œuvre : CC IN2P3, CCSD, CINES.

Missions : Accès, indexation, archivage pérenne, bus de services.

Date de création : 2007.

Budget à prévoir : 3M€.

2. SHS 2.0

"SHS 2.0" aurait vocation à mettre en place un réseau social universitaire de haut niveau, afin de permettre l'épanouissement en ligne du débat scientifique [12].

- SEMINAIRE VIRTUEL PERMANENT

Il faut offrir un dispositif complet : listes de discussions et Wiki (Universalistes), mais aussi une ou plusieurs plateformes de blogging scientifique afin de permettre la diffusion la plus large possible de la parole scientifique et de favoriser le débat dans les carnets de recherche (*Hypothèses*). Ces "séminaires virtuels permanents" constituent une opportunité historique pour développer la visibilité de la recherche française, pour développer l'articulation science-société et pour favoriser le débat scientifique au-delà des disciplines. Les perspectives heuristiques d'un tel dispositif sont très importantes.

- RESEAUX SOCIAUX ACADEMIQUES

Il est également nécessaire d'offrir une solution alternative aux réseaux sociaux privés généralistes, dont les politiques concernant le respect de la vie privée, la confidentialité et l'usage des informations nominatives et le modèle de développement présentent des inconnues, pour ne pas dire qu'elles sont sources d'inquiétudes. Le mouvement actuel d'inscription des chercheurs dans les différents réseaux sociaux est rapide et manifeste d'une nécessité importante, qui désarticule la visibilité de la recherche française. Celle-ci est très faiblement lisible en dehors de ses murs. Pourtant, elle constitue un gisement d'expertises qui pourrait se déployer fortement si les personnes à la recherche d'experts disposaient d'information à leur sujet. Un tel dispositif ne pourra fonctionner sans une appropriation très forte par la communauté scientifique. Le rendre obligatoire serait contre-productif, tout comme lui donner une allure fortement institutionnelle. En revanche, lui donner une forte utilité concrète sera le meilleur moyen de le faire réussir. Il faut donc que l'outil soit pertinent pour les communautés scientifiques elles-mêmes.

---

[12] Le terme « SHS 2.0 » a été inventé par Pierre Mounier en 2005. http://blog.homo-numericus.net/article32.html

Pour cela, il faudra qu'il soit très souple, qu'il produise une identité numérique scientifique contrôlée par les chercheurs et fournisse des services comme des outils bibliographiques (croisant HAL et Zotero, par exemple). Des alliances avec d'autres réseaux sociaux sont nécessaires, pour éviter les isolats et promouvoir les passerelles, la circulation, l'échange.

- FORGE

Plateforme d'extraction et de traitement des données. Cette plateforme aura pour vocation de s'appuyer sur SourceSup et sur Plume, pour capitaliser l'important travail de développements informatiques réalisés en France autour des différents outils et instruments de traitement et à l'étranger autour des *Digital Humanities*.

- COMPETENCES

Le quatrième axe de cet équipement devra être celui de la diffusion des compétences liées au numérique en sciences humaines et sociales. Cela doit se faire par des formations diplômantes et des formations tout au long de la vie.

Objet : Doter les chercheurs d'outils mutualisés de débats scientifiques et de partages en ligne.

Usagers : Chercheurs, enseignants.

Projets : Réseau social des Sciences humaines et sociales, éclairant notamment les enjeux de société de la science en train de se faire ("Facebook des SHS"), plateforme(s) de carnets carnets de recherches, banques de connaissances sur les bonnes pratiques numériques et forge logicielle destinée à la mutualisation des développements informatiques locaux.

Maîtrise d'oeuvre : CRU[13] (Sympa[14], Universalistes[15]), Cléo[16] (Hypothèses[17], Calenda[18]),

[13] http://www.cru.fr/
[14] http://www.sympa.org/

Paris V (Carnets de Paris Descartes[19]), URFIST[20], Plume[21], Sourcesup[22], Suplibre[23]...

Date de création : à définir.

Budget à prévoir : 3M€.

[15] https://listes.cru.fr/sympa
[16] http://cleo.cnrs.fr
[17] http://hypotheses.org/
[18] http://calenda.revues.org
[19] http://blogs.univ-paris5.fr/
[20] http://urfistinfo.blogs.com/
[21] http://www.projet-plume.org/
[22] http://sourcesup.cru.fr/
[23] http://www.cru.fr/documentation/suplibre/

3. Édition scientifique numérique (ÉSN)

La perspective de mettre en valeur environ 1500 revues en sciences humaines et sociales et quelques centaines de collections d'ouvrages doit être envisagée dans sa totalité. Il s'agit d'une part de numériser le patrimoine scientifique (Persée), de permettre la mise en ligne des revues et collections de livres vivantes (Revues.org), de poursuivre la politique d'auto-archivage menée par le CCSD (archives ouvertes). Il ne s'agit, en revanche, pas d'une bibliothèque scientifique numérique. Cet ambitieux programme d'édition scientifique numérique est une occasion majeure pour offrir aux Presses universitaires, qui sont en proie à des difficultés structurelles, des plateformes dignes des missions qui leur sont confiées. Le développement de l'impression à la demande et les enjeux du libre accès constituent des opportunités à ne pas rater. Enfin, les enjeux associés à l'exploration et à la mise au point de nouveaux paradigmes concernant l'édition électronique sont très importants et la communauté scientifique doit disposer d'ouils innovants pour y prendre sa part et mettre en valeur ses préoccupations heuristiques et ses valeurs.

Maîtrise d'oeuvre : CCSD[24], Persée[25], Revues.org[26] (Cléo).

Partenaires : Presses universitaires, laboratoires de recherches, sociétés savantes, laboratoires éditant des revues ou des ouvrages, éditeurs indépendants.

Usagers : Chercheurs, étudiants, enseignants.

Budget à prévoir : 3M€.

[24] http://www.ccsd.cnrs.fr/
[25] http://www.persee.fr
[26] http://www.revues.org

4. CORPUS-SHS

Coopération des opérateurs de recherche pour un usage des sources numériques en SHS (données qualitatives). La richesse très importante des travaux déjà en cours dans ce domaine doit faire l'objet d'une enquête précise, permettant l'établissement d'un état des lieux. Le CNRS a lancé une première enquête dans ce sens en 2009 et une synthèse des résultats des grandes enquêtes menées dans le passé (enquête des TRA, par exemple) et plus récemment (ACI puis programmes ANR), montreront l'ampleur des corpus existants et auront à définir les efforts de normalisation qui s'imposent. Les Centres de ressources numériques pourront constituer de bons points de départ pour structurer le paysage.

Usagers : chercheurs, étudiants.

Maîtrise d'oeuvre : TELMA[27], CNRTL[28], CRDO[29], CN2SV[30], MUTEC[31]…

Budget à prévoir : 3 M€.

[27] http://www.cn-telma.fr/
[28] http://www.cnrtl.fr/
[29] http://crdo.risc.cnrs.fr/
[30] http://www.cn2sv.fr/
[31] http://www.mutec-shs.fr/

5. PROGEDO

PROGEDO. Production et gestion de données pour les sciences humaines et sociales (données quantitatives).

Centre d'aide à la production de données quantitatives pour l'économie, la sociologie, la démographie, les sciences politiques, le droit, la géographie et l'histoire. PROGEDO participe à des missions de collecte, de diffusion, de promotion, d'aide à la production et de préservation d'un vaste ensemble de données quantitatives pour les sciences de l'homme et de la société. Ainsi, il assure un accès contrôlé aux différentes données issues

de la Statistique publique (INSEE, DARES, DEPP...), des enquêtes et recherches répondant à des objectifs scientifiques et des sondages d'instituts privés. Il permet la production d'enquêtes de sciences sociales notamment sur des grandes cohortes (ELFE/Etude longitudinale depuis l'enfance) ou des enquêtes européennes identifiées sur le roadmap ESFRI (ESS/european social survey, SHARE/ Survey of health, ageing and retirement in Europe).

Partenaires : Centres d'archives, grands organismes produisant des données en série...

Usagers : Chercheurs, centres de recherches.

Maîtrise d'oeuvre : Centre Quetelet, INSEE...

Date de création : à définir.

Budget à prévoir : 3 M€.