PCR Confluence : Plan de Gestion de Données initial

Plan de gestion de données créé à l'aide de DMP OPIDoR

Créateurs du PGD : Quentin Verriez, Emilie Dubreucq, Matthieu THIVET

Affiliation du créateur principal : Other Organisation

Modèle du PGD : Horizon 2020 FAIR DMP (français)

Dernière modification du PGD : 18/02/2020

Financeur : Région Bourgogne-Franche-Comté

Résumé du projet :

Le PCR : *La confluence Saône-Doubs à l'âge du Fer (VIe s. av. J.-C. au Ier siècle de notre ère)*, coordonné par Matthieu Thivet et Émilie Dubreucq, vise à relancer une dynamique de recherche sur un secteur géographique clé : la confluence de la Saône et du Doubs, située en plein coeur de la Bourgogne, à une vingtaine de kilomètres au nord de Chalon-sur-Saône.

Nous souhaitons travailler d'une part, sur les interactions hommes/environnement dans la longue durée (VIe-Ier siècle), afin de mieux percevoir le rôle joué par la zone de confluence dans l'émergence d'un important pôle de peuplement à l'âge du fer. D'autre part nous aborderons la question de la proto-urbanisation et des modalités de développement des agglomérations ouvertes à spécialisation artisanale et commerciale au cours de l'âge du Fer, dont le Verdunois peut s'enorgueillir de posséder deux cas emblématiques : Bragny-sur-Saône « Sous Moussière » et Verdun-sur-le-Doubs « Petit chauvort ».

Le document qui suit décrit le cycle de vie de la donnée numérique produite par le PCR, soit son processus de création, de documentation, de partage et d'archivage. Avant tout, nous souhaitons que ce Plan de Gestion de Données (PGD), ou *Data Managment Plan* (DMP), ne reste pas un document figé, mais au contraire qu'il évolue avec les pratiques de ses utilisateurs d'année en année. En effet, la démarche est encore peu courante en SHS et nous la proposons tant dans une volonté d'ouvrir la donnée que de partager notre expérience. Ce PGD « initial » représente donc une première version, qui sera mise à jour chaque année avec le rapport scientifique du PCR.

Le PGD s'adresse autant aux chercheurs membres du PCR qu'aux autres utilisateurs des données. Pour les membres, une première étape importante est l'identification des données (créer par qui, pourquoi et comment), qui facilite ensuite le travail collaboratif par la mise à disposition d'outils adaptés et des normes d'échange. A plus large échelle, l'objectif est d'automatiser au maximum la documentation afin de pouvoir diffuser largement une donnée qui respecte la propriété intellectuelle du créateur tout en étant intelligible et exploitable par le plus grand nombre. Autrement dit, il s'agit de s'inscrire pleinement dans la démarche des *FAIR Data* (*Findable, Accessible, Interoperable, Reusable*).

Chercheur Principal : Emilie Dubreucq, Matthieu Thivet

Contact pour les Données : Quentin Verriez

PCR Confluence : Plan de Gestion de Données initial

1. Résumé descriptif des données

Pour répondre aux multiples problématiques soulevées, la force du programme repose sur son interdisciplinarité, en compilant des approches à la fois documentaires, archéologiques, géophysiques et spatiales. Les données collectées et/ou générées par les différents participants du PCR et leurs institutions de rattachement sont ainsi d'ordre multiple :

Type de données	**Format**
Textuelle	.doc, .docx, .txt, .odt, .pdf
Tabulée	.xls, .xlsx, .csv, .txt, .odf
Base de données	.fmp(12-14-16), .accdb, .csv, .xml
Vectorielle	.ai, .svg
Images	.jpg, .tiff, .png. .bmp, .pdf
Spatiale	.shape, .json
Géophysique	.grd, .dat, .csv, .bin
3D	.ply, .obj, .stl

Dans un premier temps, les données seront accessibles et exploitables par les participants du PCR, mais nous espérons qu'elles pourront par la suite profiter à un plus large public de chercheurs travaillant essentiellement en relation avec les disciplines archéologique, historique, géophysique et paléoenvironnementale.

Il est difficile d'anticiper précisément le volume de données partagées produit à ce stade du PCR, mais il ne devrait pas dépasser le seuil maximum de 500 Go/an.

2. Données FAIR

Métadonnées

Les données seront documentées par des métadonnées au format Dublin Core. Le tableau suivant présente les différents éléments de ce modèle, ainsi qu'une courte description et les informations attendues à remplir (tab. 1). Certains éléments de cette liste sont normalisés selon les recommandations du Dublin Core, et présentés sous forme de listes fermées (tab. 2 à 6).

Tableau 1 : Modèle Dublin Core exploité par le PCR Confluence

Elément Dublin core	Description	Informations attendues	Commentaire
Title	Nom donné au fichier.	Titre normalisé	Voir convention de nommage
Creator	Créateur du contenu du fichier.	Créateur 1 (Nom, Prénom), Créateur 2, etc.	Autant de créateurs que nécessaire
Subject	Thème du contenu du fichier.	Sujet 1, Sujet 2, etc.	Voir tableau 2
Description	Présentation du contenu du fichier (résumé, exposé du contenu en texte libre).		Texte libre
Publisher	Entité responsable de la mise à disposition ou diffusion du fichier (personne, organisation ou service).	PCR : La confluence Saône-Doubs à l'âge du Fer (Vie s. av. J.-C. au Ier siècle de notre ère) ; UMR 5608 TRACES ; UMR 6249 Chrono-environnement ; Université de Franche-Comté	Informations répétées systématiquement et non modifiables
Contributor	Entité responsable de contributions au contenu du fichier (personne, organisation ou service).	Contributeur 1 (Nom, Prénom ou raison sociale), Contributeur 2, etc.	Autant de contributeurs que nécessaire
Date	Date de la version définitive (date de création du fichier dans sa version partagée)	AAAA/MM/JJ	
Type	Nature ou genre du contenu du fichier.	Type 1, Type 2, etc.	Voir tableau 3
Format	Matérialisation du fichier (type de medium, dimensions, supports informatiques).	Format du fichier, type MIME (si disponible)	Pour type MIME, voir tableau 4. Il sera utilisé seulement s'il correspond à un format pris en charge par le CINES.
Identifier	Référence univoque au fichier.	Digital Object Identifier (DOI)	
Source	Référence à une ressource dont le fichier décrit est dérivé.	Digital Object Identifier (DOI)	
Language	Langue du contenu intellectuel du fichier.	Code langage	Voir tableau 5
Relation			Elément non utilisé dans le cadre du PCR
Coverage	Couverture spatio-temporelle du contenu du fichier.		Voir tableau 6
Rights	Informations sur les droits associés au fichier.	Creative Commons BY NC	Information répétée systématiquement et non modifiable

Tableau 2 : Element "Subject"

Termes issus des notices d'autorité RAMEAU du catalogue de la BNF

Terme	URI de la notice
Archéologie	https://catalogue.bnf.fr/ark:/12148/cb13318444z
Géophysique	https://catalogue.bnf.fr/ark:/12148/cb119395915
Culture matérielle	https://catalogue.bnf.fr/ark:/12148/cb133189670
Données géospatiales	https://catalogue.bnf.fr/ark:/12148/cb167314005
Paléoenvironnement	https://catalogue.bnf.fr/ark:/12148/cb12149208x
Archéométrie	https://catalogue.bnf.fr/ark:/12148/cb12093753j
Télédétection	https://catalogue.bnf.fr/ark:/12148/cb119336434
Bioarchéologie	https://catalogue.bnf.fr/ark:/12148/cb17793413c.public

Tableau 3 : Element "Type"

Liste recommandée par le Dublin Core et simplifiée pour l'utilisation du PCR.

Type	Sous Type	Définition	Commentaire
Collection		Une compilation de ressources	Une collection décrit un groupe de données ; ses parties peuvent également être décrites séparément.
Dataset		Données encodées dans une structure définie.	Il s'agit par exemple de listes, de tableaux et de bases de données.
Text		A resource consisting primarily of words for reading.	Il s'agit par exemple de livres, d'articles, d'archives de listes de diffusion. Notez que les fac-similés ou les images de textes sont toujours du genre Texte, comme les données 3D.
Image	Still Image	Une représentation visuelle statique.	Il s'agit par exemple de peintures, de dessins, de conceptions graphiques, de plans et de cartes. La meilleure pratique recommandée est d'attribuer le type Texte aux images des documents textuels.
Image	Moving Image	Une série de représentations visuelles donnant une impression de mouvement lorsqu'elles sont montrées successivement.	Les exemples comprennent les animations, les films, les programmes de télévision, les vidéos, les résultats visuels d'une simulation.

Tableau 4 : Element "Format"

Liste de référence des types MIME maintenue par l'organisation IANA (Internet Assigned Numbers Authority), recommandée par le Dublin Core et simplifiée pour l'utilisation du PCR.

Extension	Type de document	Type MIME
.aac	fichier audio AAC	audio/aac
.abw	document AbiWord	application/x-abiword
.arc	archive (contenant plusieurs fichiers)	application/octet-stream
.avi	AVI : Audio Video Interleave	video/x-msvideo
.bin	n'importe quelle donnée binaire	application/octet-stream
.css	fichier Cascading Style Sheets (CSS)	text/css
.csv	fichier Comma-separated values (CSV)	text/csv
.doc	Microsoft Word	application/msword
.docx	Microsoft Word (OpenXML)	application/vnd.openxmlformats-officedocument.wordprocessingml.document

.gif	fichier Graphics Interchange Format (GIF)	image/gif
.htm .html	fichier HyperText Markup Language (HTML)	text/html
.jar	archive Java (JAR)	application/java-archive
.jpeg .jpg	image JPEG	image/jpeg
.js	JavaScript (ECMAScript)	application/javascript
.json	donnée au format JSON	application/json
.mpeg	vidéo MPEG	video/mpeg
.odp	présentation OpenDocument	application/vnd.oasis.opendocument.presentation
.ods	feuille de calcul OpenDocument	application/vnd.oasis.opendocument.spreadsheet
.odt	document texte OpenDocument	application/vnd.oasis.opendocument.text
.png	fichier Portable Network Graphics	image/png
.pdf	Adobe Portable Document Format (PDF)	application/pdf
.ppt	présentation Microsoft PowerPoint	application/vnd.ms-powerpoint
.pptx	présentation Microsoft PowerPoint (OpenXML)	application/vnd.openxmlformats-officedocument.presentationml.presentation
.rar	archive RAR	application/x-rar-compressed
.rtf	Rich Text Format (RTF)	application/rtf
.svg	fichier Scalable Vector Graphics (SVG)	image/svg+xml
.swf	fichier Small web format (SWF) ou Adobe Flash	application/x-shockwave-flash
.tif .tiff	image au format Tagged Image File Format (TIFF)	image/tiff
.ts	fichier Typescript	application/typescript
.vsd	Microsoft Visio	application/vnd.visio
.wav	Waveform Audio Format	audio/x-wav
.xhtml	XHTML	application/xhtml+xml
.xls	Microsoft Excel	application/vnd.ms-excel
.xlsx	Microsoft Excel (OpenXML)	application/vnd.openxmlformats-officedocument.spreadsheetml.sheet
.xml	XML	application/xml
.zip	archive ZIP	application/zip
.3gp	conteneur audio/vidéo 3GPP	video/3gpp audio/3gpp dans le cas où le conteneur ne comprend pas de vidéo
.3g2	conteneur audio/vidéo 3GPP2	video/3gpp2 audio/3gpp2 dans le cas où le conteneur ne comprend pas de vidéo
.7z	archive 7-zip	application/x-7z-compressed

.ai	Adobe Illustrator file	application/postscript
.txt	Text File	text/plain

Tableau 5 : Element "Language"
Liste de référence ISO 639-3 (code à 3 lettres), recommandée par le Dublin Core et simplifiée pour l'utilisation du PCR.

Ref_Name	**Id**
Albanian	sqi
Bosnian	bos
Bulgarian	bul
Croatian	hrv
Czech	ces
Danish	dan
Dutch (Netherlands)	nld
English	eng
Estonian	est
Finnish	fin
French	fra
German	deu
Hungarian	hun
Icelandic	isl
Irish	gle
Italian	ita
Latvian	lav
Lithuanian	lit
Modern Greek	ell
Norwegian	nor
Polish	pol
Portuguese	por
Romanian	ron
Russian	rus
Serbian	srp
Slovak	slk
Slovenian	slv
Spanish	spa
Swedish	swe
Swedish	swe
Swiss German	gsw
Turkish	tur
Ukrainian	ukr

Tableau 6 : Element "Coverage"
Termes issus des notices d'autorité RAMEAU du catalogue de la BNF

Terme	Sous terme	URI de la notice d'autorité RAMEAU du catalogue de la BNF
Protohistoire		https://catalogue.bnf.fr/ark:/12148/cb119707216
	Civilisation de La Tène	https://catalogue.bnf.fr/ark:/12148/cb11944010z
	Civilisation de Hallstatt	https://catalogue.bnf.fr/ark:/12148/cb11942233p
Antiquité		https://catalogue.bnf.fr/ark:/12148/cb11975677r
	1er siècle av. J.-C.	https://catalogue.bnf.fr/ark:/12148/cb12044924q
	2e siècle av. J.-C.	https://catalogue.bnf.fr/ark:/12148/cb12044923c
	3e siècle av. J.-C.	https://catalogue.bnf.fr/ark:/12148/cb120449221
	4e siècle av. J.-C.	https://catalogue.bnf.fr/ark:/12148/cb11976859m
	5e siècle av. J.-C.	https://catalogue.bnf.fr/ark:/12148/cb11977422d
	6e siècle av. J.-C.	https://catalogue.bnf.fr/ark:/12148/cb16770060h
	1er siècle	https://catalogue.bnf.fr/ark:/12148/cb119938537
Verdun-sur-le-Doubs (Saône-et-Loire, France)		https://catalogue.bnf.fr/ark:/12148/cb15273536r
Bragny-sur-Saône (Saône-et-Loire, France)		https://catalogue.bnf.fr/ark:/12148/cb15273028r
Les Bordes (Saône-et-Loire, France)		https://catalogue.bnf.fr/ark:/12148/cb152730173
Allerey-sur-Saône (Saône-et-Loire, France)		https://catalogue.bnf.fr/ark:/12148/cb152729776
Verjux (Saône-et-Loire, France)		https://catalogue.bnf.fr/ark:/12148/cb152735400
Ciel (Saône-et-Loire, France)		https://catalogue.bnf.fr/ark:/12148/cb152731041
Saunieres (Saône-et-Loire, France)		https://catalogue.bnf.fr/ark:/12148/cb152734741

Le modèle de métadonnée employée est illustré ici par plusieurs tableaux, afin de l'exposer de la manière la plus complète et la plus transparente possible. Dans la pratique, des formulaires simplifiés seront créés pour que chaque créateur puisse documenter les métadonnées Dublin Core et minimiser au maximum les risques d'omission ou d'erreur d'inattention.

Convention de nommage

Par son caractère pluridisciplinaire, il est compliqué d'anticiper les types de données susceptibles d'être produits dans le cadre du PCR. Afin de ne pas créer inutilement des « cases vides » ou « manquantes », il est nécessaire de rassembler l'ensemble des données brutes et traitées déjà existantes avant de définir une convention de nommage optimale et évolutive. Néanmoins, il est déjà possible de formaliser quelques éléments indispensables pour la bonne exploitation des fichiers :

- Numéro d'action :
  Une action correspond à l'intervention d'un ou plusieurs opérateurs sur une durée définie pour répondre à un ou plusieurs objectifs. Une emprise géographique peut être associée en option.

Ex : opération archéologique, étude de spécialiste, reprise/analyse documentaire

Les actions seront numérotées chronologiquement de 1 à N. L'action 0, plus particulière, regroupera les fichiers qui normalisent ou régissent les autres actions, par exemple l'inventaire des actions ou le PGD.

- Catégorie de données :

Chaque donnée sera rattachée à une catégorie. Une liste fermée (mais évolutive) des catégories sera définie selon les premières données produites.

- Titre court

Un titre court pour le fichier résumant son contenu et son but. Il sera possible de le normaliser en fonction des différentes catégories de fichier.

Quelques règles d'édition sont aussi à observer :

- Pas de caractères spéciaux

Ex : à, é, ï, @, *, &

- Pas d’espace, préférer le underscore

Ex : « mon_fichier » et pas « mon fichier »

- Pas de majuscule

S’il peut sembler fastidieux de répéter systématiquement certains éléments, dans la pratique, des solutions logicielles telles que Ant Renamer permettent de modifier semi-automatiquement par lot des noms de fichier.

Identifiant pérenne

Les métadonnées Dublin Core et la convention de nommage sont nécessaires pour déposer les données sur un entrepôt sécurisé qui fournira un DOI (*Digital Object Identifier*) par fichier sélectionné.

Gestion des versions

Pour les données traitées, il est nécessaire de conserver l’historique des modifications majeures d’un fichier. À partir de la seconde version, l’élément « V2 » sera ajouté à la suite du nom du fichier, « V3 » pour la troisième version, etc. Les métadonnées seront mises à jour pour chaque fichier.

Par définition, les données brutes et pérennes ne peuvent pas avoir plusieurs versions. Les données brutes deviennent traitées si elles sont utilisées. Les données pérennes représentent des versions définitives. Il est possible de les modifier, mais le fichier sera considéré comme une nouvelle donnée citant sa source dans ses métadonnées.

Les données transiteront par deux espaces, mis à disposition par la Très Grande Infrastructure de Recherche (TGIR) Huma-Num, qui s’adaptent à deux étapes de leur cycle de vie :

- Un espace de stockage par le service ShareDocs, qui propose une gamme d’outils collaboratifs pour faciliter le partage et l’édition de fichiers communs. Il sera réservé aux membres du PCR et accueillera les données brutes et de travail que les chercheurs souhaitent partager.
- Un espace de diffusion par le service Nakala, qui permet de déposer, de documenter et d’exposer des données pérennes. Un lot de données sélectionnées (brutes et pérennes) et déposées en même temps que le rendu du rapport annuel du PCR auprès des services de la DRAC (Service Régional d’Archéologie). La propriété intellectuelle de chaque fichier diffusé sera renseignée, afin de permettre sa réutilisation dans les meilleures conditions tout en augmentant la visibilité des résultats du PCR.

Les données (et leurs métadonnées) déposées sur Nakala respecteront la liste des formats validés par le CINES (Centre Informatique National de l’Enseignement Supérieur) pour un archivage pérenne. Il s’agit là de formats libres et ouverts pouvant être réexploités par la plupart des logiciels.

Le vocabulaire utilisé dans le modèle de métadonnées Dublin Core présenté plus haut a été normalisé chaque fois que cela a été possible. Il se base sur le langage d’indexation RAMEAU, exploité sous forme de notice par la Bibliothèque National de France (BNF). Nous avons fait le choix d’employer un vocabulaire précis, mais sans trop le spécialiser, pour ne pas limiter la visibilité de la donnée.

L’entrepôt de données Nakala est structuré selon les principes, méthodes et technologie du Web de données (entrepôt RDF (Resource Description Framework) de type *Triple Store*), qui permet de moissonner les informations par des moteurs de recherches spécialisés tels que ceux proposés par ISIDORE, Europeana ou encore Gallica.

Comme déjà mentionné plus haut (cf. 2.2), un lot de données sélectionnées, vérifiées et renseignées sera déposé annuellement sur Nakala, en appui à la publication du rapport transmis aux services de l’état. Chaque fichier de ce lot sera diffusé sous licence Creative Commons Attribution + Pas D’Utilisation Commerciale (CC BY NC) : « le titulaire des droits autorise l’exploitation de l’œuvre, ainsi que la création d’œuvres dérivées, à condition qu’il ne s’agisse pas d’une utilisation commerciale (les utilisations commerciales restant soumises à son autorisation). » (http://creativecommons.fr/licences/).

L’attribution oblige celui qui réutilise la donnée à créditer son créateur. Dans le cas présent, il sera fait mention du nom et prénom du ou des créateurs, systématiquement présent dans les métadonnées accompagnant le fichier. La mention devra se faire soit dans des métadonnées structurées, soit le cas échéant directement dans le contenu du fichier.

La reproduction, la modification et la diffusion de nouvelles données sont possibles, à condition que l’attribution soit respectée et que l’utilisation ne soit pas à des fins commerciales.

Les données de travail ne seront accessibles que par les membres du PCR, sur l’espace dédié ShareDocs.

3. Allocation de ressources

Le PCR a intégré pleinement le développement du travail collaboratif, le principe d'interopérabilité et l'importance de la diffusion des données (notamment en respectant les principes FAIR) en proposant d'en faire un axe de recherche, intitulé « Bancarisation des données et SIG ». S'il devient aujourd'hui indispensable de systématiser ce genre d'approche dans les programmes de recherche, nous souhaitions à travers ce PCR proposer un cadre pour expérimenter, optimiser et partager les protocoles et méthodes que nous mettons en place.

Le recollement de la donnée sera effectué par Quentin Verriez, qui proposera ensuite des normes de gestion et d'échange adaptées aux types de données. Une fois les normes validées, Quentin Verriez proposera une formation, afin que chacun assure la mise aux normes et de la documentation des données qu'il produit. Le dépôt, la gestion et la diffusion des données en ligne seront sous la responsabilité de Matthieu Thivet et Quentin Verriez.

En ce qui concerne le cout financier de la conservation à long terme des données, les services du CINES et d'Huma-Num indiquent une valeur approximative de 700 € par an et par Téraoctet (1000 Go) (communication Consortium 3D SHS). Cela devrait donc représenter une fourchette comprise entre 200 et 350 € par an pour les données du PCR, assumer par les deux structures qui offrent leurs services.

4. Sécurité des données

En ce qui concerne la sécurité des données sur la durée du programme, le PCR s'appuie sur les compétences et les outils de la TGIR Huma-Num, avec ses plateformes ShareDocs et Nakala. Pour l'archivage à long terme, la TGIR propose un service en lien direct avec la Plateforme d'Archivage au CINES (PAC). Cette démarche sera automatisée au maximum par l'ensemble des normes de création, de nommage et de validation des données produites par le PCR.

5. Aspects éthiques

Chaque fichier validé par le PCR sera accompagné de sa métadonnée qui décrit précisément et clairement la propriété intellectuelle de la donnée (créateur(s), contributeur(s), éditeur(s) et droits). Dans ces conditions, toute personne qui citera ou réexploitera des informations partagées par le PCR sans respecter les conditions de propriété intellectuelle (créditer le ou les créateurs, réutilisation non commerciale) s'exposera à de possibles poursuites judiciaires.

Les seules données à caractère personnel qui seront partagées par le PCR représenteront les informations nécessaires pour définir la propriété intellectuelle des fichiers, soit les nom et prénom du ou des créateurs, leurs institutions de rattachement et un mail de contact.

Si une donnée est renseignée, validée par son ou ses créateurs et déposée sur les entrepôts identifiés par le PCR, les membres du programme acceptent de fait, selon les conditions définies par la propriété intellectuelle, le partage de cette donnée.

6. Autres

Glossaire :

Données brutes :

Données primaires acquise par un ou plusieurs opérateurs, selon une méthode, à l'aide d'un dispositif et n'ayant subi aucun traitement (autre que ceux que le dispositif d'acquisition peut réaliser automatiquement).

Données traitées :

- Données de travail : données traitées, mais au caractère non définitif, reflétant une étape intermédiaire du processus de recherche, avec un intérêt de diffusion à large échelle faible. (ex : plan de publication, schéma, etc.).
- Données pérennes : données exploitées et renseignées (métadonnées) par un ou plusieurs opérateurs selon un protocole défini afin de répondre à un objectif scientifique.

Lot de données :
Ensemble de données (brutes ou traitées) cohérent et uniforme, produit ou traité par un ou plusieurs mêmes opérateurs « en bloc » pour répondre à un objectif scientifique (ex : photographies, planches, données géophysiques, etc.).