Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

Ministère Culture Communication

TRAÇABILITÉ DES DONNÉES NUMÉRIQUES

QUELS MODÈLES POUR LA PROVENANCE DES DONNÉES NUMÉRIQUES ? ÉTAT DE L'ART

Ministère de la Culture et de la Communication
Stratégie « Métadonnées culturelles et transition Web 3.0 »

INTRODUCTION

Le Web sémantique et ses technologies favorisent l'interconnexion de nombreuses sources de données qui n'auront pas toutes la même fiabilité. Dans ce contexte, il est important de pouvoir retracer la provenance des données pour permettre aux utilisateurs de les réutiliser avec confiance. Le consortium W3 a élaboré un modèle, PROV, générique, extensible et interopérable, pour les métadonnées de provenance.

1. CONTEXTE

Le présent document s'inscrit dans la continuité de la feuille de route sur les métadonnées culturelles et la transition Web 3.0. du MCC (http://cblog.culture.fr/projet/2013/11/07/groupe-de-travail-metadonnees-culturelles/), qui s'appuyait elle-même sur les actions 4 et 5 de la feuille de route stratégique ministérielle en faveur de l'ouverture et du partage des données publiques. La feuille de route du MCC identifiait neuf actions opérationnelles, dont l'action 8 visant à « positionner le MCC en tant qu'expert sur la traçabilité des données numériques » et à faire un « état de l'art sur les modèles permettant de reconstituer la provenance des données ». Un groupe de travail spécifique rassemblant des représentants des ministères de la Culture et de la communication (MCC) et des Affaires étrangères et du Développement international (MAEDI), d'une collectivité territoriale (Conseil général de la Gironde) et du Laboratoire d'informatique en image et systèmes d'information (LIRIS) s'est donc réuni d'avril à octobre 2014 (voir la composition du groupe de travail en annexe 1) pour dresser un état de l'art en matière de description de la provenance.

2. ENJEUX ET OBJECTIFS

Le Web sémantique et ses technologies favorisent l'interconnexion de nombreuses sources de données qui n'auront pas toutes la même fiabilité. Dans ce contexte, il est important de pouvoir retracer la provenance des différentes données afin de permettre aux utilisateurs de les réutiliser avec confiance dans leurs propres systèmes de connaissance. Les métadonnées de provenance peuvent répondre à des questions critiques pour des cas d'utilisation différents :
qualité des données : connaître la provenance d'une donnée permet de juger de sa qualité, de savoir si elle répond aux exigences techniques et légales d'un métier donné. Toute erreur ou approximation introduite en amont se retrouve en aval et peut vicier des raisonnements apparemment bien construits. Le « data journalisme » ou « journalisme de données », par exemple, utilisant de grandes masses de données pour faire naître des informations nouvelles, peut conduire à des conclusions erronées si le contexte de validité des informations n'est pas mûrement évalué ;
reproduction de processus : la provenance n'est pas seulement une trace de ce qui a été fait mais aussi un moyen de le reproduire. Pour une expérience scientifique, par exemple, les détails sur les conditions de son déroulement, les instruments utilisés, les logiciels utilisés ainsi que leurs versions doivent être enregistrés pour pouvoir reproduire l'expérience et aboutir aux mêmes résultats que l'expérience initiale ;
conformité légale : un historique précis des traitements effectués sur des données peut servir de preuve de conformité aux normes et réglementations en vigueur. En cas de conflit entre deux entités sur l'authenticité d'une donnée, la provenance peut servir comme un élément de preuve légale.

La gestion de la provenance des données numériques révèle les problématiques suivantes :
la difficulté d'identifier toutes les sources de provenance, distribuées sur de multiples systèmes et gérées par différentes entités,
l'hétérogénéité des données et des modèles de provenance,
la difficulté pour différents acteurs d'interroger les mêmes données de provenance avec des besoins métier différents.

Cette problématique n'est pas nouvelle puisqu'elle se retrouve dans l'architecture du Web sémantique, représentée traditionnellement comme un empilement de couches technologiques et/ou fonctionnelles dont les plus élevées sont les couches « Preuve » (Proof) et « Confiance » (Trust) qui assurent la vérification des déclarations effectuées dans le Web 3.0. Ces deux couches permettent d'avoir un environnement Web plus fiable et mieux sécurisé dans lequel les tâches d'authentification, de sécurité et de sûreté peuvent être assurées.

3. PRINCIPAUX AVANTAGES OU BÉNÉFICES ATTENDUS

La traçabilité des données numériques représente tout d'abord un enjeu économique important pour le secteur culturel, en lien avec l'identification, la protection des droits et la rémunération des créateurs. Les informations de provenance peuvent en effet faciliter la « remontée des recettes » en direction des auteurs, artistes interprètes et producteurs ou éditeurs. Elles peuvent également servir à l'identification des titulaires de droits dans les demandes de retraits d'œuvres exploitées sans autorisation. Du côté des sociétés de perception et de répartition des droits (SPRD), les carences en matière de métadonnées de provenance peuvent entraîner des coûts induits. Comme l'a rappelé le rapport Lescure sur l'Acte 2 de l'exception culturelle, « le traitement manuel des erreurs (ex : orthographes différentes d'une même œuvre), doublons (ex : attribution de deux codes différents pour un même contenu) ou incohérences (ex : deux titulaires du même droit pour un même contenu) est un frein récurrent à leur mission ». L'une des suites du rapport Lescure est le pilotage, par la direction générale des médias et des industries culturelles du ministère de la Culture et de la communication (DGMIC), d'une étude visant à analyser la faisabilité technique, économique, financière et réglementaire de la mise en place de registres ouverts de métadonnées pour le livre, la musique, la photographie, l'audiovisuel et la presse, pour permettre notamment une amélioration de la rémunération des auteurs, artistes interprètes et producteurs ou éditeurs pour les morceaux exploités sur les plates-formes.

Les informations de provenance peuvent également permettre de mieux vérifier que les conditions d'usage sont bien respectées et qu'on a le droit d'accéder à tel ou tel document sans outrepasser les droits afférents. Par exemple, la gestion numérique des droits (GND), ou gestion des droits numériques (GDN), en anglais digital rights management (DRM), s'attache à la traçabilité des usages des objets et permet ainsi de protéger les œuvres contre le piratage.

Par ailleurs, sur Internet, l'usager se retrouve confronté à des systèmes qui s'attachent à construire une connaissance collective, mais en ignorant le plus souvent quelles données ont été utilisées et sans toujours comprendre comment les résultats de ses requêtes ont été obtenus. Des métadonnées de provenance sont donc nécessaires pour attribuer de la confiance aux données rencontrées lors de la recherche, favoriser ainsi leur exploitation et offrir ainsi de meilleurs services aux ré-utilisateurs potentiels. La gouvernance des données numériques est donc un enjeu stratégique pour les autorités publiques, comme l'atteste la création, par décret du 16 septembre 2014, de la fonction d'administrateur général des données :
l'exploitation de données de confiance permet d'objectiver dans certains cas les décisions politiques ;
des données de confiance sont un gage de transparence d'un État démocratique à travers leur ouverture ;
elles sont créatrices d'innovations et de richesses à travers leur réutilisation.

Enfin, des métadonnées de provenance riches et bien structurées permettront aux institutions culturelles de se positionner comme tiers de confiance :

- en reprenant l'initiative dans la communication du savoir sur le Web grâce au partage de données fiables ;
- en vérifiant que les données qu'elles produisent et publient sont bien citées, ce qui peut contribuer à ce qu'elles deviennent des données de référence, enjeu primordial pour les services et établissements sous tutelle du MCC ;
- en se liant à d'autres données de confiance.

4. RISQUES À NE PAS FAIRE

L'absence de traçabilité des données numériques est un facteur de risques pour le secteur culturel :

- elle peut constituer un frein important au développement des industries culturelles à l'ère numérique ;
- une mauvaise gestion des informations de provenance peut entraîner une dégradation de la qualité des données et un manque de confiance des utilisateurs qui ne sont plus en mesure de hiérarchiser les différentes sources d'information en leur attribuant des niveaux de fiabilité ;
- les institutions culturelles risquent à terme de perdre de leur visibilité si les utilisateurs n'ont plus la possibilité de distinguer les sources produites par les services et établissements sous tutelle du MCC et celles enrichies par d'autres.

5. RECOMMANDATIONS DU GROUPE DE TRAVAIL

A l'issue de l'état de l'art dressé par le Gt8 en matière de description de la provenance (voir les annexes 2 « Références des modèles étudiés », 3 « Présentation détaillée des modèles » et 4 « Tableau comparatif des métadonnées de provenance dans les différents modèles étudiés »), quelques recommandations peuvent déjà être faites en direction des producteurs de données culturelles.

Il convient en premier lieu que les producteurs de données mettent en place des dispositifs pour instaurer la confiance. Outre l'adoption d'identifiants pérennes ou le développement d'un système d'identification pour les auteurs de ressources culturelles, par exemple, il serait très souhaitable que les institutions culturelles fassent auditer leurs systèmes d'information, pour évaluer les modalités d'implémentation des standards, règles et processus concernant la gestion et l'utilisation des métadonnées de provenance, ainsi que la complétude des informations de traçabilité. Cette démarche est analogue à celle définie par la norme ISO 16363:2012 – Systèmes de transfert des informations et données spatiales - Audit et certification des référentiels numériques dignes de confiance : obtenir une forme de certification mais aussi évaluer les limites et les risques en matière de traçabilité de l'information en toute transparence, pour établir une confiance.

Un deuxième ensemble de recommandations porte sur le choix du modèle de provenance à implémenter.

- Le consortium W3C a élaboré un modèle, PROV-DM décrivant les informations minimales de provenance qui doivent exister dans des contextes métier différents. Générique, extensible et interopérable, ce modèle semble destiné à faire l'objet d'une transposition au niveau français, à l'exemple des recommandations d'accessibilité du W3C qui avaient été reprises dans le Référentiel Général d'Accessibilité pour les Administrations (RGAA ; ***http://references.modernisation.gouv.fr/rgaa-accessibilite***) ;
- Les « métiers » ont de leur côté élaboré des modèles spécialisés décrivant la provenance pour leur contexte spécifique. Ces modèles « métier » sont sémantiquement plus riches que les modèles abstraits proposés par le W3C. Par contre, ils ne sont pas forcément interopérables ni réutilisables hors du métier pour lequel ils ont été définis, car ils contiennent des concepts et des propriétés qui leur sont spécifiques. Par ailleurs, pour la plupart, ces modèles ne sont pas exprimés sous forme d'ontologies. Ils fournissent des modèles pour la structuration des métadonnées de provenance des données numériques, mais ils ne permettent pas à des moteurs d'inférence de travailler avec ces métadonnées, de vérifier qu'elles sont cohérentes et de déduire de manière automatique d'autres métadonnées de provenance.

Enfin, certaines des recommandations du référentiel de bonnes pratiques élaboré pour les producteurs de données publiques en Région Aquitaine (***http://checklists.opquast.com/fr/opendata***) pourraient être reprises au sein du Référentiel Général d'Interopérabilité (RGI), notamment :

- les métadonnées associées à chaque jeu de données doivent être proposées dans un format standard ;
- les métadonnées décrivant le jeu de données doivent être structurées de manière normative ;
- le catalogue des métadonnées doit être disponible sous la forme d'un jeu de données ;
- le catalogue des métadonnées doit être récupérable selon un protocole normalisé ;
- chaque jeu de données doit être livré avec un « changelog » ou « journal des modifications » ;
- le ou les jeux de données doivent être accompagnés d'au moins un moyen de contact avec le producteur (ou le mainteneur) ;
- les jeux de données doivent être accompagnés d'un résumé et d'un lien vers la version complète de la licence ;
- la licence doit mentionner les conditions de paternité, de réutilisation, de redistribution et de commercialisation.

COMPOSITION DU GROUPE DE TRAVAIL

Roselyne Aliacar, Ministère de la Culture et de la communication / Secrétariat général / Département des programmes numériques

Francisca Cabrera, Ministère de la Culture et de la communication / Secrétariat général / Département des programmes numériques

Maya Khelifi, Ministère de la Culture et de la communication / Secrétariat général / Département de la stratégie et de la modernisation

Alain Mille, Laboratoire d'informatique en image et systèmes d'information (LIRIS)

Florent Palluault, Ministère de la Culture et de la Communication / Direction générale des médias et des industries culturelles / Service du livre et de la lecture

Stéphane Reecht, Bibliothèque nationale de France / Direction des services et des réseaux / Département de la conservation

Pascal Romain, Conseil général de la Gironde, Direction des systèmes d'information

Claire Sibille – de Grimoüard, Ministère de la Culture et de la communication / Direction générale des patrimoines / Service interministériel des archives de France, animatrice du groupe

Gwendoline Stab, Ministère des Affaires étrangères et du développement international et équipe projet du programme VITAM (Valeurs Immatérielles Transmises aux Archives pour Mémoire)

Ana Teixeira, Ministère de la Culture et de la communication / Secrétariat général / Département des programmes numériques

Édouard Vasseur, Ministère de la Défense et équipe projet du programme VITAM (Valeurs Immatérielles Transmises aux Archives pour Mémoire)

CONTENU DU DOCUMENT

RÉFÉRENCES DES MODÈLES ÉTUDIÉS PAR LE GROUPE DE TRAVAIL

Toutes les références des ressources citées ont été vérifiées et étaient correctes au 1er septembre 2014.

1. Modèles et ontologies du W3C

1.1. Le modèle PROV

PROV-DM est le modèle conceptuel de données qui sert de base pour la famille de spécifications du W3C sur la provenance. PROV-DM distingue les structures de base, formant l'essence des informations de provenance, des structures étendues pour des usages plus spécifiques de provenance. PROV-DM est organisé en six éléments, qui portent respectivement sur : (1) les entités et activités, et le moment auquel ils ont été créés, utilisés ou achevés ; (2) les dérivations d'entités à partir d'entités ; (3) les agents qui exercent des responsabilités pour les entités qui ont été générées et les activités qui ont eu lieu ; (4) la notion d'ensemble, un mécanisme nécessaire pour exprimer la provenance de la provenance ; (5) les propriétés pour relier les entités qui font référence à la même chose ; et, (6) les collections formant une structure logique pour leurs composantes.

L'ontologie PROV (PROV-O) exprime le modèle de données PROV-DM au moyen du langage OWL. Elle fournit un ensemble de classes, de propriétés et de restrictions qui peuvent servir à représenter et à échanger des informations de provenance générées dans différents systèmes et dans différents contextes. Elle peut également être spécialisée pour créer de nouvelles classes et propriétés pour modéliser les informations de provenance pour différents applications et domaines.

Modèle :
PROV Overview (*http://www.w3.org/TR/prov-overview/*)
PROV Primer (*http://www.w3.org/TR/prov-primer/*)
PROV Data Model (*http://www.w3.org/TR/prov-dm/*)
PROV Constraints (*http://www.w3.org/TR/prov-constraints/*)

Sérialisation :
PROV Ontology (*http://www.w3.org/TR/prov-o/*)
PROV XML Serialization (*http://www.w3.org/TR/prov-xml/*)
PROV Notation (*http://www.w3.org/TR/prov-n/*)

Extensions :
PROV Semantics (*http://www.w3.org/TR/prov-sem/*)
PROV Access and Query (*http://www.w3.org/TR/prov-aq/*)
PROV Links (*http://www.w3.org/TR/prov-links/*)
PROV Dictionary (*http://www.w3.org/TR/prov-dictionary*/)

Implémentations :
PROV Implementations (*http://www.w3.org/TR/prov-implementations/*)

Alignement avec Dublin Core :
PROV DC Mapping (*http://www.w3.org/TR/prov-dc/*)

1.2. Data Catalog Vocabulary (DCAT)

http://www.w3.org/TR/vocab-dcat/

Profil d'implémentation proposé par la Commission européenne
https://joinup.ec.europa.eu/asset/dcat_application_profile/asset_release/dcat-application-profile-data-portals-europe-final

Référentiel de bonnes pratiques élaboré par la communauté open data
http://checklists.opquast.com/fr/opendata

2. Protocoles d'authentification

WebID (W3C) : *http://www.w3.org/2005/Incubator/webid/spec/*

OpenID (OpenID Foundation) : *http://openid.net/*

3. Normes et standards « métier »

3.1. Normes et standards pour la pérennisation de l'information numérique

ISO 14721:2003 - Systèmes de transfert des informations et données spatiales -- Système ouvert d'archivage d'information -- Modèle de référence

ISO 16363:2012 – Systèmes de transfert des informations et données spatiales - Audit et certification des référentiels numériques dignes de confiance

PREMIS (Dictionnaire de données pour les métadonnées de pérennisation)
http://www.loc.gov/standards/premis/

Groupe Interpares : projet international de recherche sur la préservation à long terme de l'authenticité des documents d'archives numériques
http://www.interpares.org/

3.2. Normes et standards d'archivage et de records management

Voir la note d'information DGP/SIAF/2012/005 du 15 février 2012 relative à l'actualité de la normalisation en matière de records management
http://www.archivesdefrance.culture.gouv.fr/static/5570

3.2.1. NF ISO 15489

La norme NF ISO 15489 est la norme matricielle, basée sur les fondamentaux de l'archivage au sens managérial du terme : quelles traces d'une activité conserver et pour quelle raison ? Une fois qu'elles sont identifiées, organiser leur conservation pendant la durée requise ? ISO 15489 a inspiré d'une part MoReq, d'autre part ICA-Req. La série de normes ISO 3030X va vers une certification de la démarche d'archivage managérial.

NF ISO 15489-1:2001 – Information et documentation – « Records management » - Partie 1 : principes directeurs

FD ISO/TR 15489-2 - Information et documentation – « Records management » - Partie 2 : guide pratique

ISO 23081-1: 2006 – Information et documentation -- Processus de gestion des enregistrements -- Métadonnées pour les enregistrements -- Partie 1: Principes

ISO 23081-2: 2009 – Information et documentation -- Gestion des métadonnées pour l'information et les documents -- Partie 2: Concepts et mise en œuvre

ISO 23081-3: 2011 – Information et documentation -- Gestion des métadonnées pour l'information et les documents -- Partie 3: Méthode d'auto-évaluation

ISO/TR 26122:2008 – Information et documentation -- Analyse du processus des « records »

NF ISO 30300:2011 – Information et documentation – Systèmes de gestion des documents d'activité – Principes essentiels et vocabulaire

NF ISO 30301:2012 – Information et documentation – Systèmes de gestion des documents d'activité – Exigences

3.2.2. ICA-Req

Le standard ICA-Req a été publié en juillet 2008 par le Conseil International des Archives (ICA/CIA). Il a été porté à l'ISO au début de 2011.

ISO 16175-1:2010 – Information et documentation -- Principes et exigences fonctionnelles pour l'archivage dans un environnement électronique -- Partie 1: Aperçu et déclaration de principes

ISO 16175-2:2011 – Information et documentation -- Principes et exigences fonctionnelles pour l'archivage dans un environnement électronique --- Partie 2: Lignes directrices et exigences fonctionnelles pour les systèmes de management des enregistrements numériques

ISO 16175-3:2010 – Information et documentation -- Principes et exigences fonctionnelles pour l'archivage dans un environnement électronique -- Partie 3: Lignes directrices et exigences fonctionnelles pour les enregistrements dans les systèmes d'entreprise

3.2.3. MoReq

MoReq signifie Model Requirements for the Management of Electronic Records / Exigences types pour la maîtrise de l'archivage électronique. La première version, décidée par la Commission européenne lors du DLM Forum de Bruxelles de 1996, dans l'esprit du Records management et d'ISO 15489 alors en chantier, a été publiée en 2001, mise à jour et enrichie en 2008. Les spécifications ont été refondues en 2010 (MoReq2010 a été publié en juin 2011) et le nom MoReq signifie aujourd'hui « Modular Requirements for Records Systems ».

MoReq2 – Model Requirements for the Management of Electronic Records / Exigences types pour la maîtrise de l'archivage électronique
http://moreq2.eu/ et
http://www.archivesdefrance.culture.gouv.fr/gerer/archives-electroniques/standard/moreq2/

MoReq 2010 – Modular Requirements for Records Systems, vol. 1 : core services & plug-in modules
http://moreq2010.eu

3.2.4. Normes et standards nationaux

France
NF Z44-022: 2014 – Modélisation des Échanges de Données pour l'Archivage (MEDONA) et Standard d'échange de données pour l'archivage (SEDA), version 1.0
http://www.archivesdefrance.culture.gouv.fr/seda/

Australie
Australian Government Recordkeeping Metadata Standard Version 2.0 (AGRkMS)
http://www.naa.gov.au/records-management/agency/create-capture-describe/describe/recordkeeping-metadata.aspx

Canada
Norme de métadonnées de la gestion des documents du gouvernement du Canada (NMGD GC) ou Government of Canada Records Management Metadata Standard (GC RMMS) et
Profil d'application de la gestion des documents du gouvernement du Canada (PAGD GC) ou Government of Canada Records Management Application Profile (GC RMAP)
http://www.collectionscanada.gc.ca/gouvernement/002/007002-5002.2-f.html

3.3. Normes et standards de description archivistique

3.3.1. Normes du Conseil international des archives :

ISAD(G) : Norme générale et internationale de description archivistique, 2ème édition, 1994

ISAAR(CPF) : Norme Internationale sur les notices d'autorité utilisées pour les Archives relatives aux collectivités, aux personnes ou aux familles, 2ème édition, 2004

ISDF : Norme internationale pour la description des fonctions, 1re édition, 2008

ISDIAH : Norme internationale pour la description des institutions de conservation des archives, 1re édition, 2008

Les quatre normes internationales de description sont téléchargeables depuis le centre de ressources du Conseil international des archives : ***http://www.ica.org/10241/normes/liste-des-normes.html***

3.3.2. Formats d'encodage de la Société des archivistes américains

DTD EAD (Description archivistique encodée) 2002
Site officiel : ***http://www.loc.gov/ead/***

Traduction française de la documentation et autres informations utiles sur le site du service interministériel des archives de France
Guide des bonnes pratiques de l'EAD dans les bibliothèques :
http://www.archivesdefrance.culture.gouv.fr/gerer/classement/normes-outils/ead/

Schéma EAC-CPF (Contexte archivistique encodé – Collectivités, personnes, familles)
Site officiel : ***http://eac.staatsbibliothek-berlin.de/***

Traduction française de la documentation technique :
http://www.archivesdefrance.culture.gouv.fr/gerer/classement/normes-outils/eac/

3.3.3. Normes et formats bibliographiques

IFLA universal bibliographic control and international MARC programme Manuel UNIMARC : format bibliographique. Trad. par Marc Chauveinc. - 4e éd. version française. - München : K.G. Saur, 2002.
http://catalogue.bnf.fr/ark:/12148/cb38974314z/ISBD

International Federation of Library Associations and Institutions. UNIMARC manual : bibliographic format. - 3rd ed. / edited by Alan Hopkinson. - München : K. G. Saur, 2008. - 760 p. - (IFLA Series on Bibliographic Control ; vol. 36).
ISBN 978-3-598-24284-7
http://www.worldcat.org/oclc/804227080

Manuel UNIMARC : format bibliographique. Édition française en ligne
http://www.bnf.fr/fr/professionnels/anx_formats/a.unimarc_manuel_format_bibliographique.html

Rapport de Pierre Lescure : « Mission « Acte II de l'exception culturelle » - Contribution aux politiques culturelles à l'ère numérique », mai 2013
http://www.culturecommunication.gouv.fr/var/culture/storage/culture_mag/rapport_lescure/index.htm

3.4. Programme HADOC

Piloté par le Département des systèmes d'information patrimoniaux du MCC, le programme HADOC (Harmonisation de la production des données culturelles) a pour objectif de fournir un cadre normatif pour la production des données culturelles et d'en outiller la mise en œuvre. Pour en savoir plus et notamment télécharger le modèle harmonisé pour la production des données culturelles, voir : ***http://www.culturecommunication.gouv.fr/Ressources/HADOC***

LES MODÈLES DE PROVENANCE GÉNÉRIQUES ET LES PROTOCOLES D'AUTHENTIFICATION

1. Le modèle PROV

PROV-DM est le modèle conceptuel de données qui sert de base pour la famille de spécifications du W3C sur la provenance (***http://www.w3.org/TR/prov-dm/***). L'ontologie PROV-O (***http://www.w3.org/TR/prov-o/***) exprime le modèle de données PROV (PROV-DM) au moyen du langage OWL, permettant ainsi à des machines de faire des raisonnements ou des inférences sur les informations. Le processus pour pouvoir gérer des objets tracés et inférer de manière automatique leur provenance est aussi important que la structuration des métadonnées de provenance de ces objets. C'est ce dispositif d'inférence qui va permettre d'établir la confiance.

Le modèle établit une distinction entre une provenance simple, facilement réutilisable et une provenance complexe permettant de fournir des informations détaillées sur les origines, les versions des données. Les livrables du groupe de travail du W3C incluent une proposition d'alignement avec le Dublin Core. Les agents à l'origine des jeux de données (personnes, organisations, agents logiciels) doivent eux aussi être identifiés et authentifiés, d'où le développement, parallèlement à l'ontologie PROV-O, de protocoles Web d'authentification comme OpenIDConnect (développé par l'OpenID Foundation ; ***http://openid.net/connect/***) et WebID (également développé par le W3C ; ***http://www.w3.org/wiki/WebID***).

1.1. Origines

Les travaux du W3C sont issus d'une réflexion autour de la notion de provenance, considérée comme un élément-clé pour décrire les évolutions d'une ressource, l'entité responsable de ces évolutions et les conséquences de ces changements sur la version finale de la ressource. Les informations sur la provenance permettent de répondre aux interrogations suivantes :

- qui est responsable de la création des données ?
- qui en est propriétaire ?
- qui a contribué à leur création ?
- comment ont-elles été modifiées depuis leur première version ?
- sont-elles affectées par d'autres données ?
- quels outils ont été utilisés pour générer chaque version ?

Pour décrire la provenance, il faut pouvoir disposer d'un modèle :

- décrivant les différents constituants (acteurs, révisions, etc.) ;
- compatible avec le vocabulaire RDF pour être compatible avec le web sémantique ;
- permettant plusieurs niveaux de granularité ;
- simple et facile à utiliser ;

complexe (« complet ») : avec des informations sur l'origine, les versions, etc.

« La provenance est définie comme l'enregistrement des personnes, des institutions, des entités et des activités qui jouent un rôle dans la production, la modification et la mise à disposition de données ou d'autres choses. [...] Les informations de provenance font partie des métadonnées contextuelles qui peuvent elles-mêmes devenir importantes en raison de leur propre provenance. » (Groupe d'incubation sur la provenance du W3C).

La provenance est fournie par les métadonnées, mais toutes les métadonnées ne concernent pas de provenance. Par exemple, le titre ou le format d'un livre constituent des métadonnées mais ne donnent pas d'informations sur sa provenance, tandis que la date de création, l'auteur, l'éditeur et les droits sur le livre donnent des informations sur sa provenance.

Avant 2009, plusieurs modèles et vocabulaires sur la provenance avaient déjà été élaborés :

- Open Provenance Model (OPM) (***http://openprovenance.org/***) : modèle développé en 2007 et décrivant la provenance en termes de processus (qui génèrent des artefacts), d'artefacts (impactés par les processus), et d'agents (qui contrôlent les processus). Un processus a utilisé un artefact ; un artefact a été généré par un processus ; un artefact a été dérivé à partir d'un autre artefact ; un processus a été déclenché par un autre processus ; un processus a été contrôlé par un agent ;

- Ontologie Provenir (***http://wiki.knoesis.org/index.php/Provenir_Ontology***) : approche modulaire pour la gestion de la provenance dans le domaine de la cyberscience ; trois classes de base dans l'ontologie Provenir servent à représenter les principales composantes de la provenance, à savoir les « données » (entités représentant le matériau de départ, le matériau intermédiaire et les produits finaux d'une expérimentation scientifique), les « agents » (qui affectent les processus) et les « processus » (qui affectent les données) ;

- Provenance vocabulary (***http://sourceforge.net/projects/trdf/***) : modèle développé pour décrire la provenance des Données liées (Linked Data) sur le Web ;

- Dublin Core (***http://dublincore.org/***) : le Dublin Core est un schéma de métadonnées génériques permettent de décrire des ressources numériques ou physiques et d'établir des relations avec d'autres ressources. Plusieurs des éléments de description ont trait à la provenance de la ressource décrite : qui l'a créée, quand elle a été modifiée, etc. ;

- PREMIS (***http://www.loc.gov/standards/premis/)*** : PREMIS est l'acronyme de Preservation metadata : implementation strategies. Ce format, mis au point par un groupe de travail soutenu par OCLC et RLG, propose un « framework » des éléments principaux (« core ») pour la conservation du document numérique. PREMIS se concentre sur la provenance des objets numériques archivés (fichiers, flux binaires, agrégations), mais pas sur la provenance des métadonnées descriptives ; un mapping de PREMIS avec le modèle OPM, considéré par le groupe d'incubation comme le plus générique, a été réalisé ;

- Ontologie SWAN (***http://www.w3.org/TR/hcls-swan/***) : applications du web sémantique en médecine neurologique ; ontologie modélisant le discours scientifique, développée dans le contexte de la construction d'une série d'applications pour la recherche biomédicale ;

- SIOC (Semantically-Interlinked Online Communities ; ***http://semanticweb.org/wiki/SIOC***) : vocabulaire permettant de décrire des objets couramment utilisés dans les réseaux sociaux et leurs relations ;

- VOID (Vocabulary of Interlinked Datasets ; ***http://semanticweb.org/wiki/VoID***) : schéma basé sur RDF pour décrire des jeux de données liées.

Mais ces modèles abordaient la notion de provenance à des niveaux de granularité différents. La problématique était double :

- comment rendre les informations décrivant la provenance échangeables ?
- comment intégrer ces données de provenance hétérogènes ?

En 2009, un groupe d'incubation présidé par Yolanda Gil (Director of Knowledge Technologies and Associate Division Director at the Information Sciences Institute of the University of Southern California, and Research Professor in the Computer Science Department) a donc été créé au sein du W3C. L'année suivante, ce groupe a publié un rapport sur la provenance (***http://www.w3.org/2005/Incubator/prov/XGR-prov/***) décrivant les approches et vocabulaires existants et proposant la création d'un groupe de travail dédié au W3C. Le groupe d'incubation a également défini les exigences relatives à la provenance sur le web (***http://www.w3.org/2005/Incubator/prov/wiki/User_Requirements***) :

Catégorie	Dimension	Description
Contenu	Objet	L'artefact auquel se rapporte l'assertion de provenance
	Attribution	Les sources ou entités qui ont contribué à créer l'artefact en question.
	Processus	Les activités (ou étapes) qui ont été réalisées pour générer ou mettre à portée de main l'artefact.
	Version	Enregistrements des changements qu'a connus un artefact avec le temps et des entités et processus associés à ces changements.
	Justification	Documentation enregistrant pourquoi et comment une décision donnée a été prise.
	Lien	Explications montrant comment des faits ont été dérivés d'autres faits.
Gestion	Publication	Rendre la provenance disponible sur le Web.
	Accès	La capacité à trouver la provenance d'un artefact donné.
	Diffusion	Définir comment la provenance doit être partagée et son accès contrôlé.
	Échelle	Faire face à un très grand nombre d'informations de provenance.
Utilisation	Compréhension	Comment permettre à l'utilisateur final d'utiliser des informations de provenance.
	Interopérabilité	Combiner les informations de provenance produites par plusieurs systèmes différents .
	Comparaison	Comparer des artefacts à travers leur provenance.
	Responsabilité	Utiliser la provenance pour accorder du crédit ou blâmer.
	Confiance	Utiliser la provenance pour faire des jugements de confiance.
	Imperfections	Faire face aux imperfections des enregistrements sur la provenance
	Débogage	Utiliser la provenance pour détecter les bugs ou défaillances des processus.

Par ailleurs, le Groupe d'incubation, estimant que le modèle OPM était le plus générique, a fait un alignement d'OPM avec d'autres modèles, PREMIS notamment. Ce travail est disponible à : ***http://www.w3.org/2005/Incubator/prov/wiki/Provenance_Vocabulary_Mappings#Mappings***

Enfin, trois cas d'usage de la provenance orientés sur les usages Web ont été proposés :

- un cas d'usage sur l'agrégation d'actualités : un site agrégeant des informations provenant de diverses sources (sites d'actualités, blogs et tweets), où des enregistrements sur la provenance peuvent aider à vérifier la véracité des informations et à identifier les droits afférents ;
- un cas d'usage relatif à la propagation d'une maladie : intégration et analyse de données pour étudier la propagation d'une maladie, en impliquant les pouvoirs publics et la recherche scientifique ;
- un cas d'usage relatif à un contrat commercial : vérification de la conformité d'un livrable avec le contrat original et l'analyse du processus de conception à partir d'enregistrements de métadonnées de provenance.

Suite à ces travaux, un groupe de travail *ad hoc* co-présidé par Paul Groth et Luc Moreau a été constitué en avril 2011, avec comme objectif de définir un modèle pour échanger des informations sur la provenance sur le web, en se concentrant sur le web sémantique. Le groupe a produit 13 livrables (voir Références des modèles étudiés par le Gt8, p. 7 sq.) dont quatre (modèle, ontologie, contraintes, notations) ont valeur de recommandation du W3C.

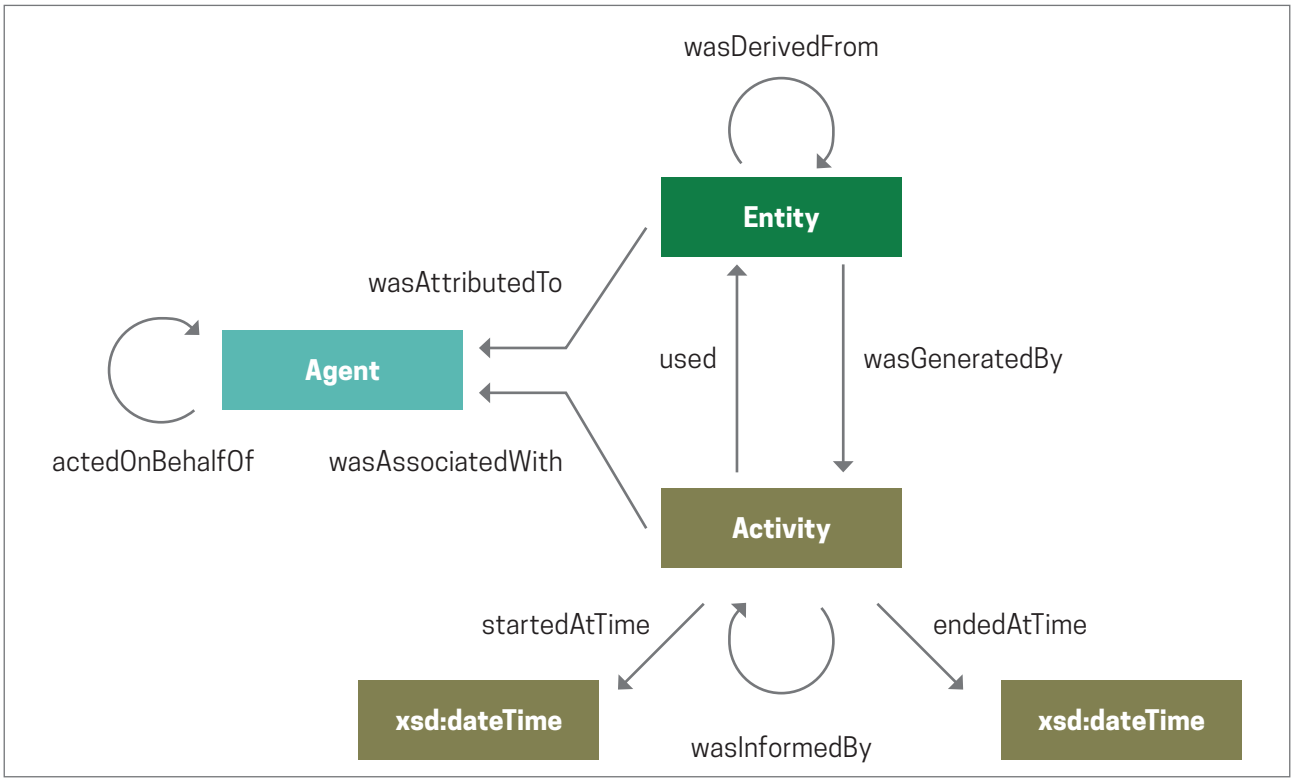
1.2. Présentation du modèle

Le modèle PROV-O comprend trois catégories de concepts :

- 3 classes et 7 relations principales qui constituent le cœur du modèle ;
- 7 classes et 18 relations étendues (concepts complémentaires pour une description plus riche) ;
- des classes et des relations qualifiées (pour des descriptions plus complexes).

1.2.1. Modèle générique

Le schéma ci-dessous représente les trois principales entités du modèle PROV-O ainsi que les relations qu'elles peuvent avoir les unes avec les autres.



Le modèle générique sur la provenance décrit l'utilisation et la production d'entités par des activités, qui peuvent être influencées de différentes manières par des agents.

Entity (Entité)
Dans le modèle PROV, une entité est une ressource dont on veut décrire la provenance. « Une entité est un objet physique, numérique, conceptuel ou tout autre type d'objet avec des aspects déterminés ; les entités peuvent être réelles ou imaginaires. » Par exemple : un document, une partie d'un document, une idée, un article de nouvelles, un contrat, un résultat, etc.

Activity (Activité)
Les activités sont les processus qui ont utilisé ou généré des entités, comme par exemple : calculer un résultat, écrire un livre, faire une présentation. Les activités ne sont pas des entités. « Une activité est quelque chose qui se produit pendant une période déterminée et qui agit sur ou avec des entités ; elle peut inclure l'utilisation, la transformation, la modification, la délocalisation, ou la génération d'entités. »

Agent (Agent)
Les agents sont responsables des activités affectant les entités. Un agent est quelque chose qui porte une forme de responsabilité dans le déroulement d'une activité, dans l'existence d'une entité ou dans l'activité d'un autre agent. Ce peut être une personne, une composante de logiciel, un objet inanimé, une organisation, ou une autre entité.

Les entités, les activités et les agents peuvent interagir les unes par rapport aux autres :

- d'un point de vue temporel :
  - une Entité a été générée (wasGeneratedBy) par une Activité ;
  - une Activité a utilisé (used) une Entité ;
  - une Activité a été fondée (wasInformedBy) sur une autre Activité ;
  - une Entité est dérivée (wasDerivedFrom) d'une autre Entité ;
  - une Activité a commencé (startedAtTime) à une Date/heure donnée ;
  - une Activité s'est terminée (endedAtTime) à une Date/heure donnée.

- du point de vue de la responsabilité :
  - une Activité a été associée (wasAssociatedWith) à un Agent ;
  - une Entité a été attribuée (wasAttributedTo) à un Agent ;
  - un Agent a agi pour le compte (actedOnBehalfOf) d'un autre Agent.

Le tableau ci-dessous récapitule les principaux concepts ainsi que leur dénomination dans PROV-O.

Concepts	Types ou relations	Dénomination
Entité	Types du modèle PROV	Entity
Activité		Activity
Agent		Agent
Génération	Relations du modèle PROV	WasGeneratedBy
Utilisation		Used
Communication		WasInformedBy
Dérivation		WasDerivedFrom
Attribution		WasAttributedTo
Association		WasAssociatedWith
Délégation		ActedOnBehalfOf

On trouvera ci-après une définition des relations entre entités, activités et agents.

Generation (Génération)
Les activités génèrent de nouvelles entités. La génération permet de décrire l'origine des entités et de répondre à des questions, comme par exemple : comment un document a-t-il été généré ? Comment un résultat de calcul a-t-il été obtenu ? Comment une entité a-t-elle été modifiée ? Comment un résultat a-t-il été validé ?

Usage (Utilisation)
Les activités utilisent également des entités. L'utilisation permet de préciser quelles sont les entités qui ont participé à une activité, par exemple : les références utilisées pour créer un document, la requête faite pour obtenir un résultat ou encore les flux entrants d'un processus informatique.

Communication (Communication)
La communication sert à décrire l'interdépendance entre deux activités. Quelles sont les activités qui ont précédé l'activité actuelle ? Quelles sont les étapes nécessaires pour exécuter une requête ?

Derivation (Dérivation)
Les activités utilisent et génèrent des entités. Dans certains cas, l'utilisation d'une entité a une influence sur la création d'une autre entité. Cette « influence » ou dérivation est la transformation d'une entité en une autre entité, le résultat de la mise à jour d'une entité est la génération d'une nouvelle entité à partir de l'entité préexistante. La dérivation permet de décrire l'interdépendance de différentes entités entre elles. Par exemple : les contenus d'un document s'appuient-ils sur d'autres entités ? Comment un résultat de calcul dépend-t-il de bases de données externes ? Quelles ressources ont influencé cette entité et dans quelle mesure ?

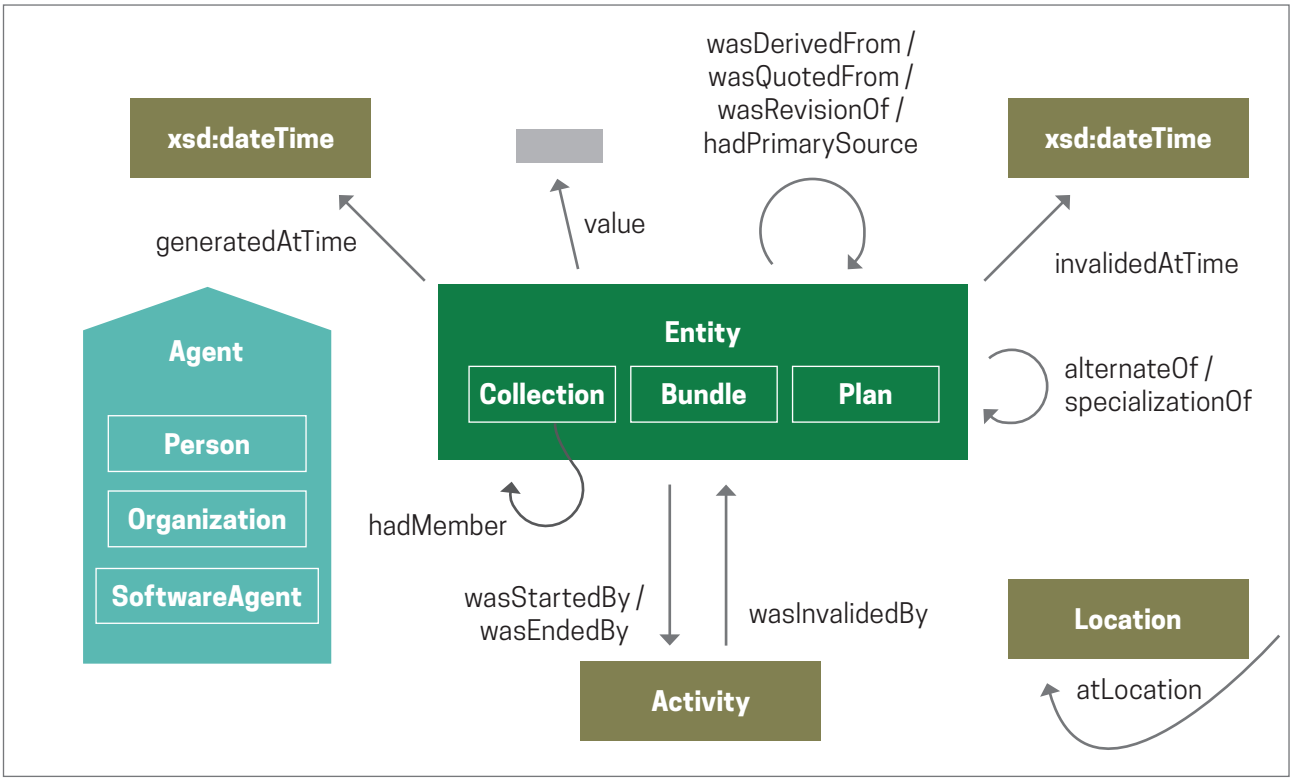
Attribution (Attribution)
L'attribution est l'assignation d'une activité à un agent. Elle permet de répondre à des questions comme par exemple : qui est l'auteur d'un document particulier ? Quel logiciel a été utilisé pour générer tel résultat ? Qui a créé tel jeu de données ?

Association (Association)
Un agent peut se voir attribuer une certaine responsabilité dans le déroulement d'une activité. D'après le modèle PROV, l'Activité est associée à l'Agent. La relation d'association permet de répondre à des questions comme par exemple : qui a la responsabilité d'un document ? Qui a la responsabilité de la production d'un résultat de calcul expérimental ? Qui a la responsabilité de l'élaboration d'un produit/contrat ?

Delegation (Délégation)
La délégation est l'assignation d'une autorité et d'une responsabilité à un agent (par lui-même ou par un autre agent) pour exercer une activité spécifique comme délégué ou représentant, tandis que l'agent pour le compte duquel il agit détient une certaine responsabilité dans le résultat du travail qui a fait l'objet de la délégation. La délégation sert à préciser les responsabilités de plusieurs agents les uns par rapport aux autres. Par exemple : qui est responsable de la génération du résultat d'une l'expérience de calcul (agissant pour le compte de l'Université Polytechnique de Madrid) ?Quel utilisateur a activé tel outil pour générer tel rapport ?

1.2.2. Modèle étendu

Le modèle étendu est une extension des principaux concepts de PROV-O. Le schéma ci-dessous présente les concepts supplémentaires permettant de décrire plus finement la provenance.



Un premier type d'extension du modèle générique consiste à créer des super-types ou des sous-types pour les principales classes et les relations.

Dans le modèle étendu, les classes peuvent être spécialisées au moyen de sous-classes. C'est ainsi que la classe Agent (Agent) comprend trois sous-classes : Person (Personne), Organization (Organisme) et SoftwareAgent (Agent logiciel).

De même, la classe Entity (Entité) est spécialisée via trois sous-classes : Collection (Collection), Bundle (Ensemble) et Plan (Plan).

- Une Collection est une Entité qui fournit une structure (par exemple : une série, une liste, etc.) à des composants (qui sont eux-mêmes des entités). La sous-classe Collection peut servir à exprimer la provenance de la collection elle-même : par exemple qui a constitué la collection, quels en étaient les composants alors que cette collection a connu des évolutions, et comment ces composants ont été assemblés. La relation hadMember (avait pour membre) sert à exprimer l'appartenance à une collection.
- Un Ensemble (bundle) est un ensemble nommé de descriptions de provenance (par exemple un graphe nommé, un fichier contenant des descriptions de provenance). Dans le modèle étendu, la provenance est elle-même considérée comme une entité, il est donc possible d'exprimer la provenance de la provenance.

Un Plan est une entité représentant un ensemble d'actions ou d'étapes envisagées par un ou plusieurs agents pour remplir des objectifs.

Les relations définies dans le modèle générique peuvent, elles aussi, être complétées ou précisées. La relation wasInfluencedBy (a été influencé par) est une super-relation décrivant l'influence qu'une Entité, une Activité ou un Agent ont pu avoir sur une autre Entité, une autre Activité ou un autre Agent. Quant à la relation de « dérivation », elle peut être décrite plus finement au moyen de trois sous-types :

- la Citation (wasQuotedFrom; a été cité de) est un type de dérivation indiquant qu'une Entité (une citation) a été citée à partir d'une autre Entité (un document) ;
- la Révision (wasRevisionOf ; était la révision de) indique que l'Entité dérivée contient du contenu substantiel provenant de l'Entité originelle ;
- la relation hadPrimarySource (avait pour source primaire) mentionne une Entité précédente produite par un agent ayant une expérience et une connaissance directes du sujet (par exemple : quelle est la source primaire d'un billet sur un blog , d'un article de nouvelles, d'un résultat de recherche … ?).

Un deuxième type d'extension du modèle porte sur le niveau d'abstraction des entités. Certaines entités peuvent présenter des aspects plus spécifiques que d'autres entités plus générales. La relation specializationOf (est la spécialisation de) relie une Entité spécifique à une Entité plus générale (par exemple : la page d'accueil des nouvelles de la BBC aujourd'hui et la page d'accueil des nouvelles de la BBC de n'importe quel jour ; deux versions différentes d'un document qui peuvent être les spécialisations d'une entité générale représentant le document). La relation alternateOf (est la variante de) relie des entités qui présentent différents aspects d'un même objet mais pas au même moment (par exemple : la sérialisation d'un document dans différents formats ou une copie de sauvegarde d'un fichier informatique).

Une troisième catégorie d'extension de PROV-O vise à permettre une description plus détaillée des entités. La propriété prov:value (valeur) fournit une valeur littérale qui est la représentation directe d'une entité. Par exemple, la valeur d'une citation peut être une série des phrases énoncées. La propriété atLocation (a pour localisation) peut servir à décrire la Localisation (Location) de toute Entité, de toute Activité ou de tout Agent ou l'Événement instantané (InstantaneousEvent) (c'est-à-dire le début et la fin d'une activité ou la génération, l'utilisation ou l'invalidation d'une entité).

Une quatrième catégorie d'extension porte sur la description de la durée de vie d'une Entité après sa génération par une Activité et son Utilisation par des Activités. Par exemple, une peinture ne peut pas avoir été exposée avant d'avoir été peinte et ne peut pas avoir été vendue après sa destruction par le feu. De même que les Activités ont un début et une fin, une Entité peut être délimitée par le moment où elle a été générée et celui où elle ne peut plus être utilisée. Les propriétés generatedAtTime (a été générée à tel moment) et invalidatedAtTime (a été invalidée à tel moment) peuvent servir à mettre en relation le début et la fin de l'existence dune Entité.

Une cinquième et dernière catégorie d'extension vise à décrire le début et la fin d'une Activité et les Activités qui l'ont précédée. Les Activités peuvent aussi être initiées et achevées par des Entités ; les relations entre Activités et Entités seront décrites au moyen des propriétés wasStartedBy (a été commencée par) et wasEndedBy (a été terminée par). Les Agents peuvent être des Entités et peuvent donc commencer et terminer des Activités. Décrire le début et la fin d'une activité peut être intéressant pour déterminer par exemple les causes de l'échec de l'exécution d'une opération.

1.2.3. Relations de qualification

Les concepts de base et les concepts étendus restent binaires : une Entité a été générée (wasGeneratedBy) par une Activité, une Activité a utilisé (used) une Entité, une Activité a été associée (wasAssociatedWith) à un Agent, etc. Mais on peut vouloir décrire davantage les relations entre Entités, Activités et Agents. Où a été générée telle Entité ? Quel a été le rôle des parties prenantes dans telle Activité ? Quand telle Entité a-t-elle été utilisée ? etc. Il est possible de qualifier les différentes relations des Entités, Activités et Agents. Il est possible par exemple de qualifier comment une Activité a utilisé une Entité. Les relations suivantes peuvent ainsi être qualifiées :

- Utilisation (used)
- Génération (wasGeneratedBy)
- Association (wasAssociatedWith)
- Dérivation (wasDerivedFrom)
- Citation (wasQuotedFrom)
- Révision (wasRevisionOf)
- Source primaire (hadPrimarySource)
- Influence (wasInfluencedBy)
- Début (wasStartedBy)

- Fin (wasEndedBy)
- Communication (wasInformedBy)
- Annulation (wasInvalidatedBy)
- Attribution (wasAttributedTo)
- Délégation (actedOnBehalfOf).

2. DCAT (Data CATalog)

DCAT-AP est un guide d'implémentation d'un modèle de données développé dans le cadre du programme européen Solutions d'interopérabilité pour les administrations publiques européennes (Interoperability Solutions for European Public Administrations (ISA)), basé et compatible avec le vocabulaire de catalogue de données du W3C (DCAT) – actuellement le vocabulaire du Web sémantique le plus largement utilisé pour décrire les catalogues de données et les jeux de données.

Le but de DCAT-AP est de définir un format d'échange commun pour les portails de données d'union européenne et des pays membres de l'Union européenne. En vue d'y parvenir, DCAT-AP définit les classes et les propriétés au sein des niveaux de conformité obligatoire (mandatory), recommandé (recommended) et optionnel (optional). Ces classes et ces propriétés correspondent aux informations concernant les jeux de données et les catalogues qui sont partagés par de nombreux portails de données européens, contribuant à leur interopérabilité. Bien que DCAT-AP soit conçu pour être indépendant d'une quelconque implémentation, RDF [RDF] et les données liées (Linked Data) [LDBOOK] sont les technologies de référence couramment utilisées.

Le schéma DCAT vise à normaliser la description des catalogues d'informations publiques. Il est actuellement maintenu par l'organisme de normalisation du web W3C. Il est consultable librement à l'adresse suivante : ***http://www.w3.org/TR/vocab-dcat/***

Ce schéma est organisé autour de l'utilisation des concepts de catalogue (dcat:catalogue), de jeux de données (dcat:dataset), de vocabulaire contrôlé (skos:ConceptScheme), de catégorie (skos:Concept), d'agents (foaf:Agent) et de ressources (dcat:resource).

Pour chacune de ces classes, des champs obligatoires, recommandés et optionnels ont été définis par un groupe de travail coordonné par la Commission européenne de manière à standardiser l'implémentation de ce schéma de métadonnées.
Les résultats du travail de ce groupe de travail sont disponibles à cette adresse :
https://joinup.ec.europa.eu/asset/dcat_application_profile/asset_release/dcat-application-profile-data-portals-europe-final

2.1. Objectifs

Ce schéma de métadonnées peut donc être utilisé pour décrire à la fois un catalogue (dcat :catalogue) de jeu de données publié et maintenu par une organisation, mais également pour décrire finement chacun des éléments constitutifs de ce catalogue, à savoir les jeux de données (dcat :dataset) décrivant les fichiers (dcat :resource) qu'il rend accessible ainsi que les différents acteurs (foaf :agent) impliqués dans cette mise à disposition.

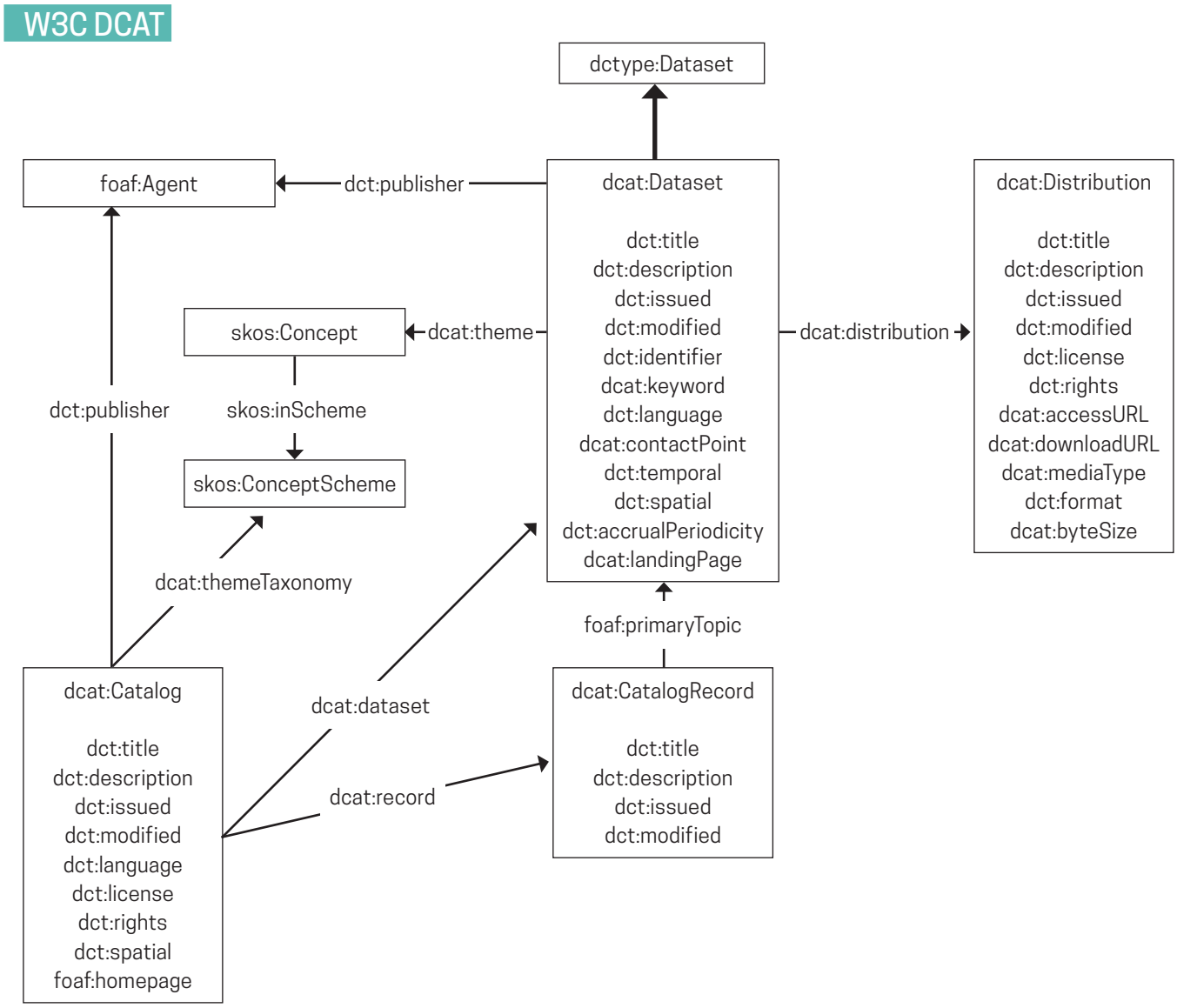
Il recommande également l'utilisation de vocabulaires contrôlés (skos :ConceptScheme) , c'est-à-dire des liste fermées de termes permettant de catégoriser les jeux de données, publiés sur Internet, de manière à faciliter la mise en relation des jeux de données avec d'autres jeux de données décrits au sein d'autres catalogues.
L'utilisation de ce standard permet enfin de fournir aux utilisateurs des fiches descriptives structurées de manière consistante sur laquelle ils peuvent baser leurs appréciations des contenus mis à disposition.
Il prend également en compte les standards de description en vigueur dans le monde du patrimoine culturel ou de la description géographique et fournit un cadre d'interopérabilité avec les langages de description utilisées par ces communautés par son caractère extensible.

Un travail mené par le Joint Research Centre of the European Commission (Unit H.6 - Digital Earth and Reference Data) concernant l'alignement des métadonnées INSPIRE avec DCAT [DCAT-AP] est notamment actuellement en cours.[1]

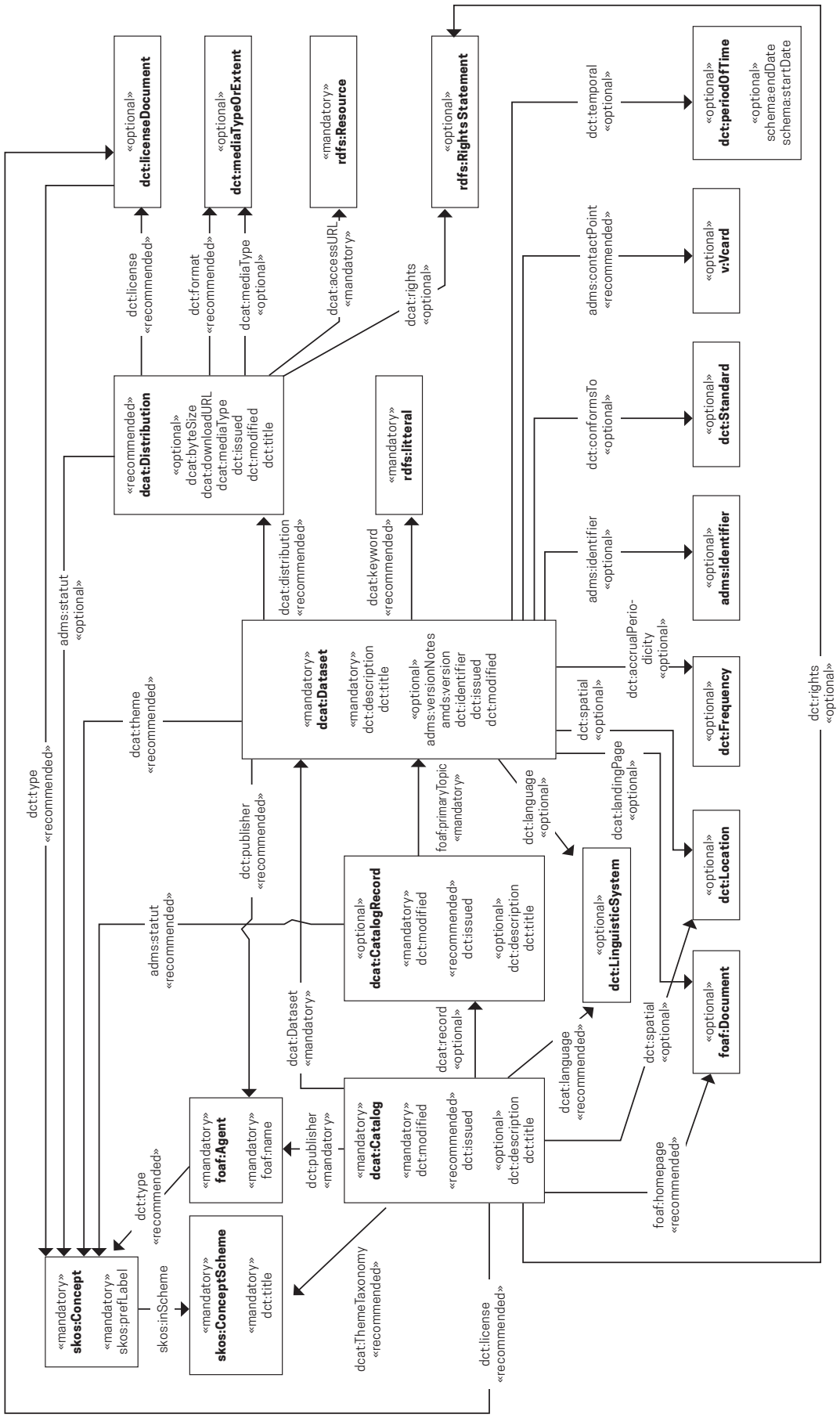
1 - https://ies-svn.jrc.ec.europa.eu/projects/metadata/wiki/Alignment_of_INSPIRE_metadata_with_DCAT-AP

2.2. Modèle de données

Le profil d'implémentation proposé par la commission européenne permet de définir le niveau minimal d'interopérabilité en précisant pour chaque catégorie d'objet décrit, les champs obligatoires, recommandés et optionnels à partir du schéma de métadonnées défini par le W3C. Le schéma ci-dessous illustre le schéma DCAT.



Le schéma ci-dessous présente la proposition d'implémentation du groupe de travail ISA pour la Commission européenne.



2.3. Prise en compte de la traçabilité

La notion existante dans le schéma INSPIRE d'origine (lineage) n'a pas d'équivalent dans le schéma DCAT, la proposition d'alignement entre ces 2 vocabulaires s'appuie sur l'utilisation du descripteur dct:provenance. Etant donné que le type d'objet de dct:provenance n'est pas un litéral mais une classe dct:ProvenanceStatement, le contenu en plain teste de l'élément « origine »(lineage)" peut être retranscrit en utilisant rdfs:label

```
# Resource metadata
[]  a dcat:Dataset ;
   dct:provenance [ a dct:ProvenanceStatement ;
    rdfs:label «Forest Map 2006 is derived from the IMAGE2006 (SPOT/LISS scenes) and CORINE2006 landcover
dataset. In
            addition, MODIS composites are used for the Forest type classification.»@en ] .
```

L'exemple ci-dessous établit dans le contexte d'une correspondance des descripteurs du modèle INSPIRE dans le profil d'implémentation DCAT fournit une proposition de structuration

```
# Resource metadata

[] dct:rightsHolder [ a foaf:Organization ;
    foaf:mbox <mailto:efdac@jrc.ec.europa.eu> ;
    foaf:name «European Union»@en ] ;
  prov:qualifiedAttribution [ a prov:Attribution ;
    dct:type <http://inspire.ec.europa.eu/codelist/ResponsiblePartyRole/resourceProvider> ;
     prov:agent [ a vcard:Kind ;
       vcard:hasEmail <mailto:efdac@jrc.ec.europa.eu> ;
       vcard:organization-name «European Commission, Joint Research Centre»@en ] ],
    [ a prov:Attribution ;
     dct:type <http://inspire.ec.europa.eu/codelist/ResponsiblePartyRole/author> ;
     prov:agent [ a vcard:Kind ;
       vcard:hasEmail <mailto:efdac@jrc.ec.europa.eu> ;
       vcard:organization-name « European Commission, Joint Research Centre «@en ] ],
[ a prov:Attribution ;
     dct:type <http://inspire.ec.europa.eu/codelist/ResponsiblePartyRole/owner> ;
     prov:agent [ a vcard:Kind ;
       vcard:hasEmail <mailto:efdac@jrc.ec.europa.eu> ;
       vcard:organization-name «European Union»@en ] ] ;

   foaf:isPrimaryTopicOf

# Metadata on metadata

[  dcat:contactPoint [ a vcard:Kind ;
    vcard:hasEmail <mailto:efdac@jrc.ec.europa.eu> ;
    vcard:organization-name «European Commission, Joint Research Centre»@en ] ;
  prov:qualifiedAttribution [ a prov:Attribution ;
    dct:type <http://inspire.ec.europa.eu/codelist/ResponsiblePartyRole/pointOfContact> ;
    prov:agent [ a vcard:Kind ;
     vcard:hasEmail <mailto:efdac@jrc.ec.europa.eu> ;
     vcard:organization-name «European Commission, Joint Research Centre»@en ] ] ] .
```

3. Les protocoles d'authentification (WebID, OpenID)

Le protocole WebID (Web Identity and discovery) a été développé par le W3C pour permettre l'identification unique sur le Web d'une personne, d'une société, d'une organisation ou de tout autre agent. WebID permet de publier son identité numérique sur n'importe quel serveur Web (indépendamment d'un quelconque fournisseur de réseau social). Il permet également d'identifier les amis des amis d'une personne et de leur accorder (ou non) l'accès à une ressource donnée.

Le protocole WebID utilise les fonctionnalités apportées par l'ontologie FOAF et le protocole de sécurité SSL/TLS :

- l'ontologie FOAF décrit les personnes et organisations ainsi que leurs activités, mais aussi leurs relations. Ses données sont utilisées dans WebID pour apporter la connaissance de la personne.
- Le protocole SSL/TLS permet la sécurisation des transmissions sur internet en utilisant un système de cryptographie par clé publique/très privée, géré par une architecture client/serveur.

Développé par l'OpenID Foundation, OpenID est un système d'authentification décentralisé qui permet à un utilisateur de s'authentifier auprès de plusieurs sites (devant prendre en charge cette technologie) sans avoir à retenir un identifiant pour chacun d'eux mais en utilisant à chaque fois un unique identifiant OpenID. Le modèle se base sur des liens de confiance préalablement établis entre les fournisseurs de services et les fournisseurs d'identité (OpenID providers). Il permet aussi d'éviter de remplir à chaque fois un nouveau formulaire en réutilisant les informations déjà disponibles.

LES MÉTADONNÉES DE PROVENANCE DANS LES NORMES ET STANDARDS « MÉTIER »

1.1. Les dictionnaires de données de métadonnées de pérennisation

Le standard PREMIS (Dictionnaire de données de métadonnées de pérennisation), qui fournit les lignes directrices pour assurer la pérennité des objets numériques et garantir leur compréhension sur le long terme, identifie lui aussi des métadonnées permettant de renseigner l'historique des opérations effectuées sur tout ou partie des paquets d'information et les agents (humains ou logiciels) qui les ont réalisés. Toutefois, PREMIS s'avère insuffisant pour exprimer seul les informations permettant d'assurer la traçabilité. L'ontologie développée à partir de PREMIS a toutefois été alignée sur le modèle OPM (cf. supra : ***http://www.w3.org/2005/Incubator/prov/wiki/Provenance_Vocabulary_Mappings#Mappings***).

1.2. Les normes et standards pour le records management et l'archivage électronique

Les différents outils analysés par le Gt 8 (normes ISO 15489 et 23081, normes nationales australienne, canadienne et coréenne, standard MoReq, travaux du groupe international InterPARES) se focalisent sur l'authenticité, la fiabilité et la traçabilité des documents en tant que tels dans un système de records management, mais sans ouverture Web. La norme ISO 15489 définit la gestion de la provenance comme étant « le fait de créer, enregistrer et préserver les données relatives aux mouvements et à l'utilisation des documents ». Le projet anglais InSPECT s'est quant à lui interrogé sur la notion de métadonnées essentielles appliquées aux messages électroniques, pour garantir la préservation des informations et leur authenticité. Ces expérimentations font écho aux travaux du programme VITAM (Valeurs Immatérielles Transmises aux Archives pour Mémoire), qui vise à développer un logiciel d'archivage électronique interministériel pour l'archivage intermédiaire et définitif. La problématique de la traçabilité est au cœur de la sélection qu'ont à faire les archivistes.

1.3. Les normes et formats de description archivistique

Depuis le début des années 1990, le Conseil international des archives a élaboré quatre normes descriptives internationales :
la Norme générale et internationale de description archivistique ISAD(G), 1ère édition en 1994, 2e édition en 2000 ;
- la Norme internationale pour les notices d'autorité utilisées par les archives – Collectivités, personnes ou familles ou ISAAR(CPF), 1ère édition en 1996, 2e édition en 2004 ;
- la Norme internationale pour la description des fonctions ou ISDF, 2008 ;
- la Norme internationale pour la description des institutions de conservation ou ISDIAH, 2008.

Ces normes sont conceptuelles et le modèle qu'elles proposent pour tracer la provenance reste très insuffisant. Un modèle conceptuel pour les archives est d'ailleurs en cours d'élaboration afin de revisiter les concepts sous-tendant les quatre normes dans une approche plus orientée « web sémantique ». Les métadonnées permettant d'identifier la provenance sont très réduites dans la norme ISAD(G) : la zone du contrôle de la description comprend en effet : les notes de l'archiviste, les règles ou conventions suivies, et la ou les dates de la description. La zone du contrôle de la description est plus détaillée dans les trois autres normes :
- Code d'identification de la notice d'autorité ;
- Code(s) d'identification du ou des services ;
- Règles ou conventions ;
- Niveau d'élaboration ;
- Niveau de détail ;
- Dates de création, de révision ou de destruction ;
- Langue(s) et écriture(s) ;
- Sources ;
- Notes relatives à la mise à jour de la notice.

Les formats d'encodage développés par la Société des archivistes américains pour publier sur Internet des descriptions d'archives (EAD ou Description archivistique encodée) et de leur contexte de production, de gestion

et d'utilisation (EAC-CPF ou Contexte archivistique encodé - Collectivités, personnes, familles) proposent un modèle plus riche comprenant des métadonnées sur la provenance du jeu de données et le nom du producteur, le statut du traitement des données, etc.. Par contre, le modèle est insuffisant en ce qui concerne la transition des données par différents acteurs, leur portée géographique et temporelle et les critères de leur sélection.

Les métadonnées de provenance contenues dans l'élément <control> (équivalent de la zone du Contrôle de la description de la norme ISAAR) peuvent être réparties en sept grandes catégories.

1) Provenance du jeu de données et le nom du producteur :

- <recordId> - Identifiant de la notice. Contient un ou plusieurs identifiant(s) unique(s) pour l'instance EAC-CPF. Obligatoire.
- <maintenanceAgency> - Agence de maintenance. Nom et information codée sur l'institution ou le service responsable de la création, de la gestion et/ou de la diffusion de l'instance EAC-CPF. Obligatoire.

2) Statut du traitement :

<maintenanceStatus> - État de la notice. Contient le statut du traitement de l'instance EAC-CPF dans le processus de travail courant. Les valeurs sont : «new» (nouveau), «revised» (révisée), «deleted» (supprimée), «cancelled» (annulée), «deletedSplit» (supprimée et scindée), «deletedReplaced» (supprimée et remplacée) ou «derived» (dérivée). Obligatoire.

3) Statut éditorial

<publicationStatus> - État de publication. Contient des informations sur le statut éditorial de l'instance EAC-CPF. Facultatif.

4) Processus de création de mise à jour et historicisation des événements :

<maintenanceHistory> - Historique des interventions. Contient des informations sur la date, le type d'agent et les évènements relatifs au cycle de vie d'une instance EAC-CPF. Contient un ou plusieurs élément(s) Intervention <maintenanceEvent> qui permettent de documenter la création, l'import, la mise à jour et la suppression de la description. Chaque intervention est précisée par le nom d'un agent, le type de l'agent (humain ou machine), le type de l'évènement, une description de l'évènement et la date de l'évènement. Obligatoire.

5) Complétude des données

- <sources> - Sources. Contient des informations sur les sources consultées pour la création de la description de l'entité ou des entités dans l'instance EAC-CPF. Contient un ou plusieurs élément(s) Source <source>. Obligatoire.

- <otherRecordId> - Identifiant d'une autre notice. Élément permettant l'enregistrement d'autres identifiants additionnels pouvant être associés à l'instance EAC-CPF. Facultatif.

- <localControl> - Contrôle local. Élément permettant d'enregistrer toute information de contrôle nécessaire pour les pratiques locales et qui n'est pas représenté par d'autres éléments dans <control>. Facultatif.

6) Structure des données

- <conventionDeclaration> - Déclaration de règle ou de convention. Contient des informations sur les règles utilisées pour élaborer l'instance EAC-CPF, en particulier les noms structurés dans l'élément Identité <identity>, ainsi que des informations sur les vocabulaires contrôlés et les thésaurus utilisés dans l'instance EAC-CPF. Facultatif.

- <localTypeDeclaration> - Déclaration d'une convention locale. Élément utilisé pour déclarer des conventions locales utilisées dans l'attribut @localType.

7) Langue et script :

<languageDeclaration> - Déclaration de la langue. Contient des informations sous forme codée et en langage naturel sur la langue de rédaction de l'instance EAC-CPF. Obligatoire.

1.4. Les métadonnées de provenance dans les bibliothèques

Les notices bibliographiques sont échangées entre bibliothèques dans un format respectant la norme ISO 2709, en particulier le format MARC (en MARC natif ou sous une forme XML). Certaines données présentent des informations sur la traçabilité de la notice.

Nota bene :
- *il n'est question ici que des données bibliographiques, le Gt8 n'a pas conduit le travail parallèle sur les données d'autorité qui reste donc à faire ;*
- *les bibliothèques publiques font beaucoup de récupération de données (dérivation) mais sans en garder nécessairement trace de la provenance. La manière dont les informations récupérées sont tracées est très liée au paramétrage des systèmes intégrés de gestion de bibliothèques (SIGB), qui peut être très variable.*

Label
Les notices bibliographiques en MARC comportent un label de notice (appelé aussi guide), situé au début de chaque notice en format UNIMARC et contenant des données permettant le traitement informatique. Il s'agit d'un ensemble de 24 caractères (positions 0-23), dont certains peuvent être générés automatiquement à la création ou modification de la notice.

Ex : positions 0-4 : Longueur de la notice (généralement calculé automatiquement) [peut servir à vérifier la complétude du transfert]
position 5 : statut de la notice (des codes précisent s'il s'agit d'une notice corrigée, à détruire, nouvelle, «fille» d'une autre notice «mère», ou complétée (mise à jour d'une notice provisoire avant publication).

001 : Numéro d'identification de la notice
Ex : FRBNF346517900000005 (notice du Catalogue général de la BnF)

003 : Identifiant pérenne de la notice
Identifiant pérenne attribué à la notice par l'agence qui a créé cette notice, qui l'utilise ou la diffuse. L'identifiant pérenne figurant dans cette zone s'applique à la notice bibliographique, non à la ressource décrite.
Ex : 001 FRBNF401336220000001
003 http://catalogue.bnf.fr/ark:/12148/cb40133622z/PUBLIC

005 : Identifiant de la version
Cette zone contient la date et l'heure de la dernière mise à jour de la notice. Les systèmes peuvent ainsi déterminer si la version de la notice en cours de traitement est antérieure, postérieure ou identique à celle qui a déjà fait l'objet d'un traitement.
Ex : 005 19850901141236.0
La date de dernière mise à jour est le 1er septembre 1985 à 14 h 12 min 36 s.

035 : Identifiant de la notice dans un autre système
Numéro d'identification des notices provenant d'autres sources.
Sous-zones
$a Identifiant de la notice dans un autre système
L'identifiant est constitué d'un code entre parenthèses désignant l''établissement, suivi du numéro d'identification ou « numéro système » attribué à cette notice dans la base de données de cet établissement. Le code doit être établi conformément à la norme ISO 15511 – Identifiant international normalisé pour les bibliothèques et organismes apparentés (ISIL). À défaut, le nom complet de la bibliothèque ou un code national peut être utilisé.

Des notices existantes peuvent contenir des codes de la liste MARC Code list for Organizations, le nom complet de la bibliothèque ou un code national.
Obligatoire sauf si la sous- zone $z est présente. Non répétable.
$z Identifiant annulé ou erroné (Facultative. Répétable.)

Ex : 001 b9301298
035 ## $a(CiZaNSB)920701098

100 : Données générales de traitement
Ensemble de 36 caractères (positions 0-35)
Positions 0-7 (obligatoire) : date de création de la notice. Date sur huit caractères numériques, sous la forme internationale normalisée (norme ISO 8601-1988).
Ex : 100 $a/0-7 = 19671005
La notice a été créée directement en format MARC le 5 octobre 1967.

Positions 22-24 (obligatoire) : langue de catalogage (ISO 639-2b).

801 : Source de catalogage
Indications sur l'origine de la notice, notamment : l'agence de catalogage ayant établi les données, l'agence ayant transcrit les données sous une forme lisible en machine, toute autre agence ayant modifié
la notice ou les données d'origine, et l'agence diffusant la notice actuelle.
Obligatoire en cas d'échange de données bibliographiques. Dans de nombreux cas cette zone sera générée automatiquement lors de l'échange.

Ex : 801 #0aFRbFR-751072303$c20041026$gAFNOR
Notice créée par la Bibliothèque de la Fondation nationale des sciences politiques en suivant les normes AFNOR. Note : en $b, la deuxième partie de l'identifiant, «751072303», se réfère au code RCR utilisé en France. RCR est l'abréviation de Répertoire des Centres de Ressource (utilisé comme numéro ISIL pour les bibliothèques).

Autre ex :
801 #0 aUSbDLC$c19590000$gAACR1
801 #l aUSbMH$c19790506
801 #2 aUSbMH$c19790506$gAACR2
801 #3 aUSbDLC$c19790912
Le catalogage original a été effectué par la Bibliothèque du Congrès en 1959 (1e édition des Anglo-American Cataloguing Rules). En 1979, l'Université de Harvard a modifié les données et les a transcrites sous forme lisible en machine (selon les règles de la 2e éd. Des AACR). Cette notice a ensuite été distribuée par la Bibliothèque du Congrès. Les codes de Marc Code List for Organizations ont été utilisés en $b pour identifier la Bibliothèque du Congrès et l'Université de Harvard.

802 : Centre ISSN
Code représentant le centre ISSN responsable de l'attribution de l'ISSN et du titre clé.
Ex : 802 ## $a07 [Code du centre ISSN France]

830 : Note générale du catalogueur
Zone utilisée pour enregistrer des informations biographiques, historiques ou autres concernant la notice : notes de travail du catalogueur sur ses sources d'information, les données douteuses, des renvois à certaines règles suivies, des justificatifs sur certains choix, etc.

886 : Données du format source qui n'ont pas été converties
Données pour lesquelles il n'y a pas de zone UNIMARC spécifique. Elle est utilisée quand une agence de catalogage convertit des notices d'un autre format et tient à conserver des éléments de données de zones n'ayant pas d'équivalent.

Ex : 886 2# $2ukmarc$a083$b00$aRussia. Education$b- Biographies – Collections
Il n'y a pas d'équivalent dans UNIMARC pour une forme rédigée de la vedette d'UKMARC établie selon le système PRECIS : 083 00$aRussia. Education$b- Biographies – Collections.

1.5. Le modèle HADOC

Le Modèle pour la production des données culturelles a été élaboré dans le cadre du programme HADOC (Harmonisation de la production des données culturelles) initié par le MCC en février 2008. Les informations actuellement modélisées sont celles qui sont nécessaires à l'établissement de la « Carte d'identité » du Bien culturel. Toutefois, les données d'identification font appel à des notions telles que « Événement », « Localisation » ou « Datation ». Ces dernières ont donc été abordées dans leur globalité de manière à répondre à tous les besoins. Un travail d'alignement du modèle sur les normes métier existantes est engagé. Est également envisagée une version RDF du modèle facilitant l'exposition des données produites selon ce modèle sur le web sémantique. La problématique de la traçabilité est abordée notamment à travers les notions d' « Événement » (tout ce qui affecte le cycle de vie d'un Bien culturel) et de « Datation » (informations sur la méthode de datation et sur sa précision).

Pour en savoir plus sur les enjeux et objectifs du modèle HADOC : ***http://www.culturecommunication.gouv.fr/Ressources/Harmonisation-des-donnees-culturelles/Modele-de-donnees2/Enjeux-et-objectifs***

ANNEXE
LES MÉTADONNÉES DE PROVENANCE DANS LES MODÈLES ÉTUDIÉS

PROV-O (classes)	PREMIS	Dublin Core	ISO 23081	HADOC	Normes ICA	EAD3	EAC-CPF	MARC
prov:Entity	premis:Object	dct:BibliographicResource dct:LicenseDocument dct:PhysicalResource dct:RightsStatement						
prov:Activity	premis:Event	dcmiType:Event	event	Événement	Dates de création, de révision ou de destruction	maintenanceevent	maintenanceEvent	100 : Données générales de traitement
prov:Plan		dct:LinguisticSystem dct:MethodOfAccrual dct:MethofdOfInstruction						
prov:Agent	premis:Agent	dct:Agent				agent	agent	
prov:SoftwareAgent				Identifiant du ou des services		agenttype=»machine»	agentType=»machine»	
prov:Organization				Identifiant du ou des services		maintenanceagency	maintenanceagency	
prov:Person						agenttype=»human»	agentType=»human»	
prov:Location		dct:Location						

ANNEXE
LES MÉTADONNÉES DE PROVENANCE DANS LES MODÈLES ÉTUDIÉS

PROV-O (classes)	PREMIS	Dublin Core	ISO 23081	HADOC	Normes ICA	EAD3	EAC-CPF	MARC
prov:generatedAtTime		dct:created dct:dateAccepted dct:dateCopyrighted dct:dateSubmitted dct:issued dct:modified						
prov:wasAttributedTo	premis:linkingAgentIdentifier	dct:creator dct:contributor dct:rightsHolder						
prov:wasDerivedFrom	premis:relatedObjectIdentification	dct:isFormatOf dct:references dct:source						
prov:alternateOf		dct:hasFormat dct:isFormatOf						
prov:hadPrimarySource		dct:source			Sources	source	source	
prov:wasRevisionOf		dct:isVersionOf						
prov:wasAssociatedWith	premis:linkingObjectIdentifier							