Le format CERIF du projet euroCRIS. Un cadre de référence pour l'identification des chercheurs et les archives institutionnelles ?

Joachim Schöpfel

Le format CERIF du projet euroCRIS

Un cadre de référence pour l'identification des chercheurs et les archives institutionnelles ?

Joachim Schöpfel, université de Lille 3, laboratoire GERIICO

Le séminaire OASIE (Open access, Services, Interdisciplinarité, Expertise) du CNAM à Paris du 28 mars 2012 a interrogé la visibilité des chercheurs et de leurs domaines d'expertise par le biais des archives ouvertes. Entre autre, OASIE a posé la question de l'intérêt des systèmes d'information de recherche (CRIS) pour le mouvement du libre accès à l'information scientifique. Plus concrètement, il a étudié l'intérêt du format européen CERIF, développé pour les CRIS européen mais encore peu connu en France. Notre communication décrit le format CERIF et présente l'organisme qui le maintient et le développe (euroCRIS), puis fait le lien avec les archives institutionnelles et la question des identifiants uniques, en particulier pour les auteurs-chercheurs.

Mots-clés : Système d'information de recherche, archive institutionnelle, identifiant auteur, format, normalisation, interopérabilité, CERIF, euroCRIS, DAI

Keywords: Current research information system, institutional repository, author identifier, format, standardization, interoperability, CERIF, euroCRIS, DAI

A propos de CERIF

L'acronyme CERIF désigne un format informatique utilisé pour représenter les données d'un système d'information de recherche. CERIF signifie *Common European Research Information Format* ou, en français, format européen commun pour l'information sur la recherche.

Il s'agit d'un format ouvert (non propriétaire) et standardisé. Son objectif principal est de faciliter l'échange d'information entre les systèmes d'information de recherche[1] des différents pays membres de l'Union Européenne et/ou leurs établissements de recherche. Ces systèmes d'information peuvent avoir un caractère national (Norvège, République Tchèque, Serbie…) ou régional (Région Flandres en Belgique), et ils peuvent se limiter à une université, un site (campus), ou une organisation de recherche, comme par exemple un établissement public à caractère scientifique et technologique (CNRS, INRA …).

Le format CERIF n'est pas un format récent. Son histoire remonte aux années 1980, avec l'apparition des premiers systèmes d'information de recherche (CRIS), dont celui du CNRS en France. A partir de 1987, un groupe d'experts a travaillé sur la conception d'un format européen. Soutenu par la Commission Européenne, ce format devait répondre à plusieurs besoins :

1. proposer un format normalisé pour l'information sur la recherche et le développement (R&D) en Europe. La normalisation du format devait faciliter l'interopérabilité des différents systèmes, l'échange d'information, et la connexion de plusieurs réservoirs (*silos*) de données.
2. Le format devait permettre une représentation détaillée, formalisée et structurée de l'information liée à la recherche.

[1] En anglais, *Current Research Information System* ou abrégé, CRIS.

3. Il devait être flexible, évolutif et extensible, sans enfermer cette description dans un carcan qui serait inadapté aux spécificités des besoins locaux et vite rendu caduc par l'évolution des contextes.
4. En particulier, le format devait permettre la synchronisation des systèmes locaux.

Il ne s'agissait pas de développer un nouveau système. Au contraire, les travaux reposaient sur une analyse précise des données, structures, formats et fonctions des systèmes existants et, également, sur une étude prospective des développements à venir. L'idée n'était donc pas d'imposer un système meilleur que les autres mais de proposer une description de l'information pour l'évolution des différents systèmes déjà en place. D'une manière réaliste, les experts ont cherché un moyen qui tiendrait compte de la diversité des solutions logicielles.
Le format cible devait être un format ouvert mais compatible avec des systèmes propriétaires. Il devait correspondre aux besoins d'un système de gestion. De même, il fallait tenir compte des particularités de la recherche, comme par exemple du rôle des publications pour la diffusion des résultats scientifiques. En outre, il devait être capable de répondre aux besoins d'un CRIS indépendant (*stand alone*) tout en jouant le rôle d'un format d'échange entre plusieurs systèmes avec des formats de données différents.
A terme, le format devait contribuer à un futur portail d'information et d'accès aux projets scientifiques européens, du moins ceux financés par l'Union Européenne.
Le résultat fut la première version du CERIF publié en 1991, en anglais et avec un point d'entrée unique pour l'information dans l'ensemble du système (*single-entry format*).
Par la suite, CERIF a continuellement évolué. Ainsi, depuis 2004, CERIF a été publié en plusieurs versions, comme modèle de données (*full data model*), format d'échange, comme scripts SQL et en XML. Voici une synthèse de l'évolution du format, telle que documentée sur le site officiel d'euroCRIS[2] (tableau 1).

Année	Appellation	Commentaire
2004	CERIF2004 1.1	Ontologie, modèle de données
2006/2007	CERIF2006 1.1	Format d'échange XML, couche sémantique
2008	CERIF2008 1.2	Scripts SQL, XML
2012	1.3	Développement des institutions et indicateurs et de la sémantique
	1.4	Reprise du modèle de données 1.3 en XML
	1.5	Terminologie, identifiant fédéré, noms

Tableau 1 : Evolution du format CERIF

Voici à titre d'illustration un extrait du format CERIF1.5 XML, la description du nouvel identifiant fédéré d'un organisme de recherche (figure 1).

```
  --> 
- <cfOrgUnit>
  - <!--
    the usual CERIF attributes of the entity
    --> 
  <cfOrgUnitId>internal-orgunit-identifier1</cfOrgUnitId>
  - <!--
    a couple of multiple-language attributes
    --> 
  <cfName cfLangCode="cs" cfTrans="o">InfoScience Praha s.r.o.</cfName>
  <cfName cfLangCode="en_GB" cfTrans="h">InfoScience Praha Ltd.</cfName>
  <cfName cfLangCode="en_US" cfTrans="h">InfoScience Praha LLC</cfName>
```

[2] http://www.eurocris.org/Index.php?page=CERIFreleases&t=1

```
<cfName cfLangCode="de" cfTrans="h">InfoScience Praha GmbH</cfName>
<cfResAct cfLangCode="en" cfTrans="o">The company provides information systems and services in the
    domain of research information processing.</cfResAct>
- <cfOrgUnit_Class>
  - <!--
    this is the CERIF SME classification uuid
    --> 
  <cfClassId>eda2b2f2-34c5-11e1-b86c-0800200c9a66</cfClassId>
  - <!--
    this is the CERIF Organisation Types Scheme uuid
    --> 
  <cfClassSchemeId>759af939-34ae-11e1-b86c-0800200c9a66</cfClassSchemeId>
  <cfStartDate>2004-03-27T00:00:00</cfStartDate>
    </cfOrgUnit_Class>
  - <!--
    this is the VAT as a Federated Identifier: the currently valid value
    --> 
- <cfFedId>
  <cfFedIdId>internal-fed-id-identifier1</cfFedIdId>
  - <!--
    this is the InfoScience VAT ID No.
    --> 
  <cfFedId>CZ27131157</cfFedId>
  <cfStartDate>2004-05-01T00:00:00</cfStartDate>
- <cfFedId_Class>
  - <!--
    this is the CERIF VAT classification uuid
    --> 
  <cfClassId>634197d9-3a2e-4df7-b982-18b41a7e6f24</cfClassId>
  - <!--
    this is the CERIF Identifier Types Scheme uuid
    --> 
  <cfClassSchemeId>bccb3266-689d-4740-a039-c96594b4d916</cfClassSchemeId>
    </cfFedId_Class>
- <cfFedId_Srv>
  - <!--
    the issuing of the VAT Identification Number service
    --> 
  <cfSrvId>internal-srv-identifier1</cfSrvId>
  - <!--
    this is the CERIF Issuer classification uuid
    --> 
  <cfClassId>d52af160-1d58-49b7-8314-65c7b0aaffc2</cfClassId>
  <cfClassSchemeId>e571769d-0058-478d-8c41-c8c0a11b24ba</cfClassSchemeId>
    </cfFedId_Srv>
    </cfFedId>
  - <!--
    the former VAT as a Federated Identifier for historical records
    --> 
- <cfFedId>
  <cfFedIdId>internal-fed-id-identifier2</cfFedIdId>
  - <!--
    this is the former InfoScience VAT Identification No.
    --> 
  <cfFedId>005-27131157</cfFedId>
  <cfStartDate>2004-04-07T00:00:00</cfStartDate>
  <cfEndDate>2004-04-30T24:00:00</cfEndDate>
- <cfFedId_Class>
  - <!--
    this is the CERIF VAT classification uuid
    --> 
  <cfClassId>634197d9-3a2e-4df7-b982-18b41a7e6f24</cfClassId>
  - <!--
    this is the CERIF Identifier Types Scheme uuid
    --> 
```

<cfClassSchemeId>**bccb3266-689d-4740-a039-c96594b4d916**</cfClassSchemeId>
</cfFedId_Class>

Figure 1: Extrait du format CERIF1.5 XML (exemple d'un identifiant fédéré)3

Le format CERIF est composé d'entités et de relations. Il y a trois types d'entités : les entités cœur (*core entities*), les résultats (*result entities*) et les entités de deuxième niveau, aussi appelées entités périphériques (*second level entities*). Le schéma suivant contient les entités les plus importantes (figure 2).
En vert, les entités cœur, au nombre de trois : l'information sur le projet scientifique (nom), l'information sur une personne (chercheur, administratif…), et l'information sur une structure ou unité (organisme, institution…). En orange, les résultats, également au nombre de trois : les publications, les brevets et les produits issus du projet en question (services, biens…). Autour de ces entités primaires, CERIF propose un certain nombre d'entités de deuxième niveau (en bleu), comme un événement, un indicateur, un CV, un équipement, une subvention, une adresse électronique etc.
Toutes ces entités sont liées entre elles, avec une ou plusieurs autres entités du même type ou d'un autre type. Ainsi, le pays est lié à l'adresse postale et à l'unité de recherche. Entités et liens sont définis mais évolutifs et modulables.[4]

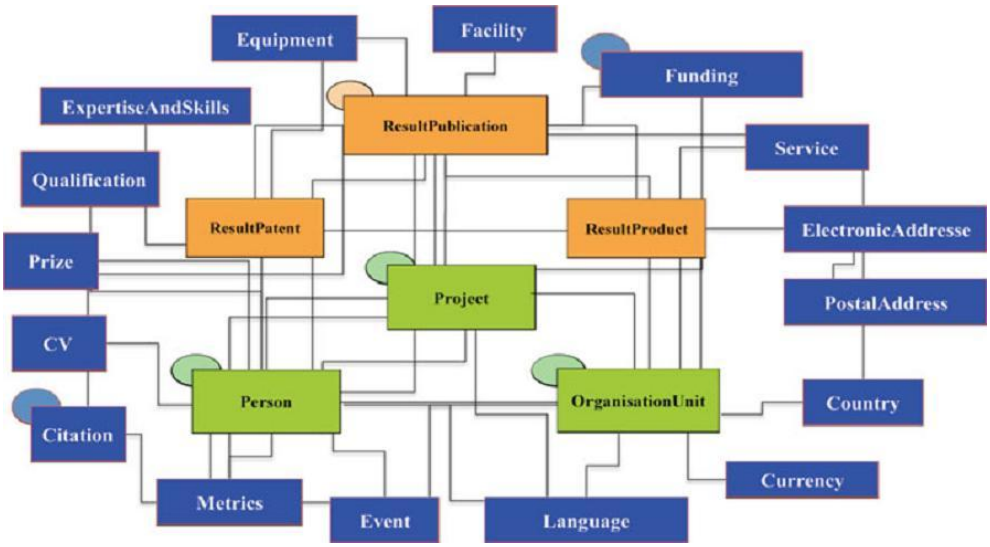

Figure 2 : L'architecture CERIF avec les entités les plus importantes (© Brigitte Joerg)

Une autre particularité de CERIF est que le format utilise de la sémantique pour définir les attributs et rôles des entités. Cette couche sémantique (*semantic layer*) permet d'indiquer avec précision la signification des différentes entités sans multiplier leur nombre. Par exemple, une personne peut être auteur, chef de projet, responsable d'une journée d'étude etc. Une publication peut être un article, un rapport, une communication ou une thèse. Une unité peut être un laboratoire, une équipe de recherche, un département ou un institut.
Cette approche rend CERIF très flexible là où d'autres formats sont obligés, pour tenir compte de changements sur le terrain, d'introduire continuellement de nouvelles entités et liens. Cette

[3] Source : http://www.eurocris.org/Uploads/Web%20pages/CERIF-1.5/orgUnit-with-fedId-sample.xml
[4] L'objectif de la communication n'est pas de présenter le format en détail. Une information plus complète est accessible sur le site d'euroCRIS http://www.eurocris.org/Index.php?page=CERIFreleases&t=1 et dans les différentes versions du tutoriel CERIF (Joerg, 2011).

flexibilité rend le format CERIF particulièrement utile pour un système d'échange (*middleware* ou *middle-layer*) entre des systèmes d'information isolés, tels que des logiciels de gestion de projet, des finances ou ressources humaines, des bases de données bibliographiques ou catalogues, d'archives institutionnelles etc.
Le format CERIF est utilisé aussi bien pour des développements dans le secteur public que par des sociétés de service en ingénierie informatique. Des sociétés comme AVEDAS Innovation Management (Allemagne) ou Atira A/S[5] (Danemark) s'appuient sur CERIF pour leurs systèmes d'information commercialisés dans les secteurs publics et privés de la recherche et du développement[6].
Plusieurs organismes publics ont développé des logiciels libres sans caractère commercial, avec des spécifications locales mais a priori adaptables à d'autres environnements. Voici un tableau non exhaustif de plusieurs logiciels CRIS (tableau 2).

Nom	Caractère	Fournisseur	Site
PURE	Commercial	ATIRA A/S	http://www.atira.dk/en/pure/
CONVERIS	Commercial	AVEDAS	http://www.avedas.com/en/converis.html
Metis	Libre	Nijmegen University	http://aptest.uci.kun.nl/metis/service/Metisguide/index.htm
ORBIT	Libre	Technical University of Denmark	http://orbit.dtu.dk
Lund University Publications	Libre	Lund University	http://lup.lub.lu.se

Tableau 2 : Logiciels CRIS utilisant le format CERIF

En ceci, CERIF reste fidèle à ses origines et à son ambition de ne pas imposer un système, mais de proposer une solution capable de faire communiquer un maximum de systèmes différents.
Le format CERIF est diffusé avec une licence *Creative Commons* CC-BY-ND qui permet une large diffusion, y compris à caractère commercial, sans modification.

Plus qu'un projet : euroCRIS

EuroCRIS est une association européenne à but non lucratif. Son siège (secrétariat) se trouve aux Pays Bas. Sa raison d'être : promouvoir et améliorer les systèmes d'information de recherche, et maintenir et diffuser le format CERIF. C'est dans cette perspective que la Commission Européenne a formellement mandaté euroCRIS en 2002 comme « gardien » du format CERIF.[7]
En fait, euroCRIS est né du développement de CERIF, des réunions d'experts de CRIS et des événements pour faire connaître CERIF et partager l'expérience des systèmes d'information existants. Après un financement initial par la Commission Européenne, euroCRIS est aujourd'hui capable de s'autofinancer, via les cotisations, l'organisation d'événements et la vente de prestations.
L'association euroCRIS comptait en août 2012 107 membres institutionnels, 41 membres à titre personnel et 17 membres affiliés, en tout 301 délégués de 42 pays, y compris des pays non européens (Canada, Etats-Unis…). Ses membres sont issus des universités et organismes de recherche, des milieux informatiques, des sociétés, gouvernements régionaux et nationaux et des éditeurs. EuroCRIS entretient des liens étroits avec d'autres organismes dans l'environnement de l'information scientifique et de la gestion de la recherche, comme par

[5] Récemment acheté par Elsevier pour développer la gamme de produits *SciVal* http://www.atira.dk/en/articles/acquisition.html
[6] Atira A/S vend le système PURE ; AVEDAS a développé le système CONVERIS.
[7] Voir aussi comme introduction Jeffery, 2011

exemple ERCIM, ESF, ICSU/CODATA ou encore le JISC. Il s'intéresse en particulier aux évolutions dans les domaines du libre accès (OAI) et des publications non commerciales (littérature grise).
L'activité principale d'euroCRIS se divise en deux axes : les événements et le développement. Les événements organisés par euroCRIS servent à fédérer les experts et communautés des systèmes d'information de recherche. Il s'agit de plusieurs types d'événements :

- des conférences internationales biennales. La dernière conférence CRIS a eu lieu en juin 2012 à Prague[8], la prochaine conférence aura probablement lieu en 2014 à Rome.
- des séminaires CRIS annuels. Ces réunions ont lieu en septembre à Bruxelles, sont consacrées à des thématiques stratégiques et produisent des rapports d'orientation stratégiques pour le développement des CRIS.
- des réunions bisannuelles des membres d'euroCRIS. Ces réunions ont lieu en mai et novembre, dans différents lieux et pays. En 2011, la réunion de novembre a eu lieu à Lille. A l'ordre du jour de ces réunions : les affaires courantes d'euroCRIS, le développement de CERIF, une thématique particulière et des réalisations et projets du pays hôte de la réunion.
- des ateliers annuels sur les CRIS, le format CERIF et les archives institutionnelles. Pour l'instant, il y a eu deux ateliers, tous les deux organisés par le CNR italien à Rome[9]. Le deuxième atelier a produit une déclaration sur l'intérêt de coordonner le développement des archives institutionnelles et des CRIS[10].

Le développement – deuxième axe d'euroCRIS – se fait essentiellement dans différents groupes de travail mis en place, pilotés et coordonnés par le comité de direction (*board*) de l'association. Actuellement, euroCRIS compte six groupes de travail[11] :

- *CERIF* : maintien et développement du format CERIF.
- *Institutional Repositories (CRIS-IR)* : lien avec les archives institutionnelles.
- *Best Practice (including also the Directory of Research Information Systems DRIS)* : recommandations, meilleures pratiques, mise en place d'un répertoire des CRIS.
- *Projects* : coordination de l'activité de conseil et d'assistance en partenariat avec d'autres organismes (comme le JISC anglais).
- *CRIS Architecture and Development* : développement des CRIS, également en partenariat avec des sociétés d'ingénierie informatique.
- *Linked Open Data* : développement du format par rapport à l'évolution du web sémantique et du web des données.

D'une manière transversale, plusieurs thématiques sont à l'ordre du jour des différents débats, travaux et événements d'euroCRIS. Il s'agit de la promotion et du développement de l'association et du format CERIF, de la mise en place d'un portail unique pour les différents CRIS, l'application du format CERIF aux archives institutionnelles (en remplacement du Dublin Core), l'intégration des données de recherche, et le lien vers les systèmes d'information liés à l'enseignement supérieur (Jeffery & Dvorak, 2012).
Par ailleurs, euroCRIS maintient un site web public (www.eurocris.org) avec un espace réservé aux membres de l'association comportant des forums de discussion, les scripts du format CERIF et une liste de diffusion.

CRIS et archives ouvertes

Comme déjà indiqué, euroCRIS s'intéresse de près aux archives ouvertes, en particulier aux archives institutionnelles (*institutional repositories*, IR). Plusieurs réunions et ateliers ont été

[8] http://www.cris2012.org/
[9] http://www.irpps.cnr.it/en/events/2nd-workshop-on-cris-cerif-and-institutional-repositories-integrating-research-information-crisoar
[10] http://www.eurocris.org/Documents/RomeDeclaration.pdf
[11] http://www.eurocris.org/Index.php?page=taskgroups&t=1

consacrés à cette thématique[12], plusieurs communications et posters ont traité des liens entre CRIS et IR lors des conférences internationales CRIS, et un groupe de travail a été mis en place en 2011. Ce dernier a pour objectif de « travailler sur une solution optimale pour l'interopérabilité des systèmes d'information de recherche d'une part et les archives institutionnelles, d'autre part, à l'échelle européenne, en tenant compte de tous les aspects pertinents ». Ce programme doit favoriser la coopération entre les communautés CRIS et OA (*open access*) et se décline en plusieurs axes :

- mener des études sur des cas concrets, et développer un cadre général pour les liens entre CRIS et IR.
- définir des métadonnées CRIS/IR, et travailler sur la conversion vers des formats habituels (*mapping* CERIF/Dublin Core, étude des sémantiques…).
- adapter le schéma CERIF-XML aux besoins de la gestion des archives institutionnelles.
- élaborer un modèle technique pour l'interopérabilité entre CRIS et IR (procédures, technologies, interface d'échange, services web).
- explorer le potentiel d'un fichier d'autorité avec identifiant pérenne (ID) pour les personnes (auteurs, chercheurs…).

La déclaration de Rome de 2011 évoquée plus haut souligne l'importance de la qualité et disponibilité (libre accès) de l'information sur la recherche scientifique, pour les chercheurs, les tutelles, les agences de moyen et la société civile. La déclaration exprime ensuite l'engagement des experts CRIS et IR pour « développer, soutenir et promouvoir une architecture (y compris un modèle de données et services) adaptée à la poursuite de ces principes, pour adopter, développer et promouvoir des standards ouverts partagés, et pour promouvoir ces principes auprès de toutes les parties prenantes ».

Pourquoi cet engagement ? L'intérêt est double. Les CRIS contiennent souvent des métadonnées plus riches et normalisées que les archives ouvertes, en particulier pour les auteurs (affiliations) car ces données sont souvent importées de grands catalogues et/ou bases de données (Vernooy-Gerritsen, 2009). Les données d'un CRIS pourraient donc enrichir la ou les archive(s) institutionnelle(s) d'un organisme, d'un établissement ou d'un site, par exemple avec des informations sur les projets de recherche ou programmes scientifiques à l'origine des publications. De l'autre côté, les archives institutionnelles peuvent ajouter les notices (métadonnées) des publications des chercheurs d'une institution aux CRIS et aussi du texte intégral, notamment pour les publications qui ne figurent pas dans les catalogues ou bases de données (y compris de la littérature grise). Cela enrichit le CRIS, évite une double saisie et ouvre la voie à des services web tels que la création de pages personnelles ou de curriculum vitae à partir des données des deux systèmes.

L'université de Glasgow par exemple a connecté son archive institutionnelle à son nouveau CRIS et a par ce biais augmenté le nombre de publications signalées par le CRIS de 4500 à 23500, un gain considérable pour la visibilité des publications mais surtout pour l'évaluation de la recherche via le CRIS (Nixon, 2010).

Synergie et valeur ajoutée sont les deux éléments cruciaux. DRIVER, le projet européen pour la création d'une infrastructure d'archives ouvertes, a clairement exprimé ce point : « The synergy between the two information domains is interesting for the DRIVER community because evidence show that well populated repositories are backed by CRIS's (…). With two systems that are traditionally managed and implemented by two different organisational units, but covering similar information and concerning the same people, the risk of building information silos and duplicated work is evident. One of the biggest motivations of

[12] Notamment le meeting des membres d'euroCRIS à Bath, en février 2012
http://www.eurocris.org/Uploads/Web pages/members_meetings/201202 - Bath, United Kingdom/

discovering the correlation between CRIS and repositories is the synergies that are obtainable and eliminate redundant work[13] » (Vernooy-Gerritsen, 2009).
Le projet DRIVER a décrit cette synergie et le potentiel d'une connexion entre CRIS et IR à partir de l'infrastructure mise en place aux Pays Bas, autour de plusieurs CRIS (dont Metis pour les universités) et archives institutionnelles (dont le réseau DARENet avec le portail NARCIS) (figure 3).

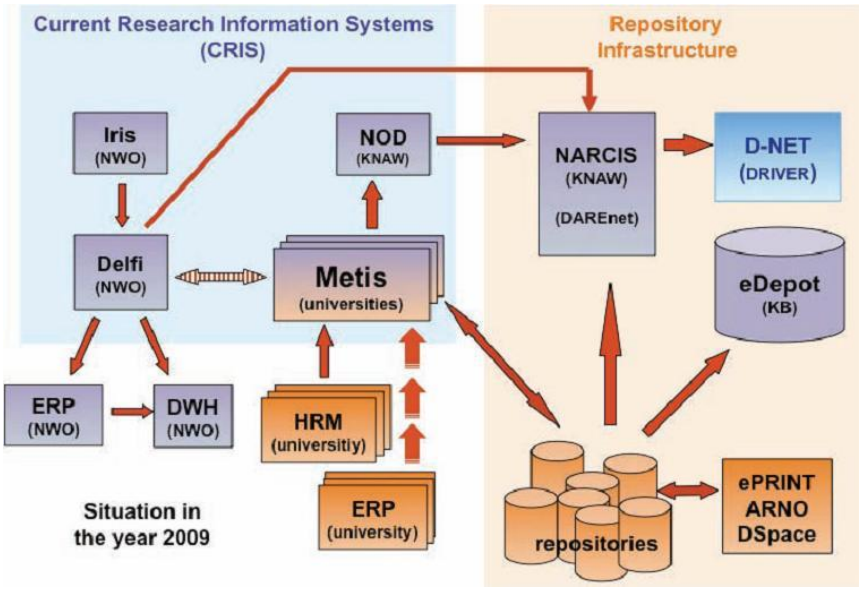

Figure 3 : Exemple d'une connexion CRIS/IR (© Vernooy-Gerritsen)

L'essentiel à retenir n'est pas cette solution particulière, adaptée à l'environnement néerlandais. L'essentiel est l'intérêt d'une telle connexion. Il n'y pas une solution unique. Par exemple, les universités de Glasgow, Southampton et Kingston utilisent leurs archives institutionnelles comme outil d'évaluation, probablement à la manière de l'université de Liège, tandis que le CRIS de l'université de St Andrew gère ses propres publications, avec toutefois l'archive institutionnelle comme point d'accès (Joint, 2008). La question se pose toujours de la même façon : où se trouvent les données dont l'institution a besoin pour gérer et évaluer la recherche ? Comment y accéder ? Comment les agréger dans un format exploitable ? Comment fournir le résultat dans un format utile ? (Rumsey, 2010)
Cette connexion entre CRIS et IR peut se faire de différentes façons. Le plus souvent elle se fera à partir des chercheurs/auteurs, les institutions et/ou les publications. Elle demandera un travail sur les métadonnées (Elbæk et al., 2010), en particulier la consolidation de la description des auteurs (Freire et al., 2012), voire la création de nouvelles métadonnées normalisées pour naviguer entre chercheurs (CV), institutions et publications, dans un environnement CRIS/ CERIF (Poltronieri et al., 2010). L'utilisation d'identifiants uniques et pérennes pour les institutions, auteurs et publications devient une nécessité.

[13] « La synergie entre les deux domaines de l'information est intéressante pour la communauté DRIVER parce qu'il s'avère que les archives riches liées à un CRIS (...). Avec deux systèmes traditionnellement gérés et mis en œuvre par différents services mais couvrant les mêmes informations et concernant les mêmes personnes, le risque du cloisonnement de l'information et du double travail est évident. L'une des plus grandes motivations pour lier CRIS et archives ouvertes sont les synergies qui peuvent être obtenues et qui éliminent le travail redondant. » (traduction JS)

L'identification des chercheurs

Pour fonctionner correctement et produire les résultats attendus, un CRIS doit être capable d'identifier d'une manière efficace et fiable les différents éléments essentiels, c'est-à-dire en particulier les projets, établissements et chercheurs, mais aussi les événements, publications, brevets et autres résultats produits par la recherche. Le format CERIF est capable de définir plusieurs rôles d'un chercheur, via la sémantique des liens et la classification des résultats (typologie des publications etc.). Quelques exemples (Ivanovic et al., 2011) :

- *author of publication or some other intellectual responsibility (editor, translator etc.),*
- *author of product or some other role (manager etc.),*
- *author of registered patent,*
- *organizer of event or some other role (performer, passive participant–observer etc.),*
- *prize winner,*
- *creator of artwork or some other role (performer, processor etc.),*
- *some of following role in sport result creation: coach, player, referee etc.*
- *author of paper published in proceedings from conference of national importance,*
- *author of paper published in journal of international importance,*
- *author of paper published in leading journal of international importance,*
- *editor of proceedings from conference of national importance,*
- *author of monograph of international importance,*
- *lector of monograph of national importance,*
- *editor of journal of international importance,*
- *winner of international prize, etc.*

Réduire l'ambiguïté et éviter des résultats équivoques deviennent des objectifs prioritaires à mesure que la demande d'une évaluation de la production scientifique augmente. Ce n'est pas par hasard qu'un certain nombre des communications de la dernière conférence sur les CRIS (CRIS 2012 à Prague) étaient consacrées à cette thématique. De même, ce n'est pas un hasard non plus si le MESR travaille sur un répertoire pour référencer de façon unique (par un seul identifiant national) l'ensemble des structures de recherche, avec un objectif d'exhaustivité.[14]

Un CRIS a plusieurs options qui par ailleurs peuvent coexister au sein du même système :

- attribution automatique d'un identifiant système à chaque nouvelle entrée, après contrôle et recherche d'éventuels doublons. Exemple : l'attribution d'un identifiant *handle* (numéro) à chaque nouveau dépôt par certaines archives ouvertes.
- utilisation d'identifiants existants. Ces identifiants peuvent être normalisés ou pas. Exemple : le DOI d'un article scientifique, l'ISBN d'un ouvrage, le numéro institutionnel d'un chercheur, l'acronyme institutionnel ou national d'un projet.
- mise en place d'une fonctionnalité (service) de contrôle, mise à jour et validation des entrées et identifiants.

Les identifiants jouent donc un rôle crucial pour le fonctionnement et la qualité d'un CRIS. Le format CERIF supporte différents types d'identifiants. L'idée est qu'un CRIS peut fonctionner avec des identifiants internes mais que pour un travail en réseau, pour récupérer des données en provenance d'autres systèmes et services (catalogues, bases de données, plates-formes de publication etc.) et dans un souci d'interopérabilité, il est crucial d'utiliser et/ou de créer des identifiants communs, normalisés, uniques, pérennes. Par ailleurs, comme l'exemple plus haut le montre (figure 1), il peut s'agir d'identifiants fédérés qui font le lien entre plusieurs identifiants pour le même élément (dans l'exemple, une institution scientifique).

[14] Répertoire national des structures de recherche (RNSR), http://appliweb.dgri.education.fr/appli_web/repStruct/index.jsp

Le même principe s'applique à l'identification du chercheur comme élément-clé d'un projet de recherche et par conséquent, d'un CRIS. Les problèmes liés à l'identification d'une personne sont connus en scientométrie[15] comme dans d'autres domaines des sciences sociales, et nous n'allons pas revenir sur cet aspect. Voici juste quelques éléments pour illustrer les propos et pour montrer quelques applications et réalisations.
Tout d'abord, l'infrastructure aux Pays Bas qui sert de modèle au projet européen DRIVER (Jippes et al., 2010). Le lien entre le réseau des archives institutionnelles et les CRIS se fait via un identifiant unique et pérenne des chercheurs, appelé Digital Author Identifier ou DAI (figure 3). En fait, comme l'indique Vernooy-Gerritsen (2009), le portail NARCIS créé une donnée « personne » qui reprend « les informations uniques sur un chercheur (…), y compris : la date de naissance, le sexe, l'identifiant unique METIS, le numéro DAI (identification de l'auteur numérique) et le numéro issu du système de gestion des ressources humaines, principalement SAP ou Oracle HR ».

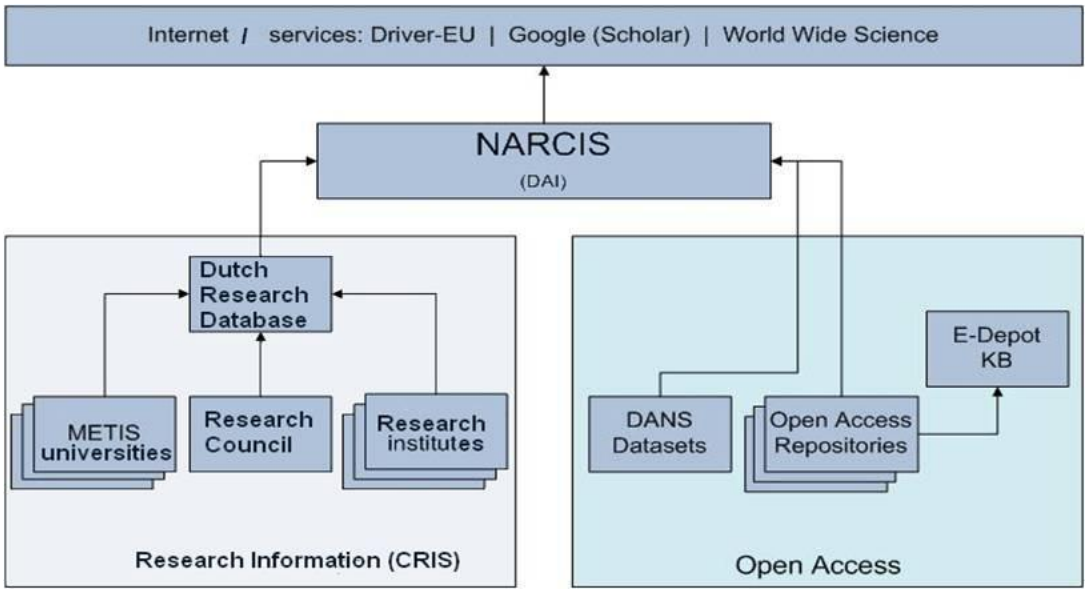

Figure 4 : Exemple d'un lien CRIS/IR via un identifiant unique de chercheur (© Jippes et al.)

Le DAI est un identifiant unique attribué d'office à chaque chercheur aux Pays Bas. L'avantage est évident, mais il est certain aussi qu'un tel choix n'est pas sans poser de problèmes et qu'il ne peut pas être transposé tel quel dans d'autres contextes. Citons parmi ces problèmes la protection de la vie privée et des libertés individuelles dans un monde numérique, d'éventuelles erreurs introduites en dehors du CRIS et a priori, difficiles à mettre à jour, l'équation postulée chercheur=auteur (en France, « chercheur-publiant »). Néanmoins, dans l'environnement scientifique des Pays Bas, ce choix semble fonctionner de façon fructueuse.
Une analyse de différents CRIS révèle l'utilisation d'autres identifiants pour les chercheurs. Voici quelques exemples (tableau 3).

Appellation	**Editeur, organisme**	**Commentaire**
DAI[16]	Enseignement supérieur et recherche, Pays Bas	
ResearcherID[17]	Thomson Reuters	Web of Science

[15] Voir par exemple Warner et al., 2009
[16] http://www.surf.nl/en/themas/openonderzoek/infrastructuur/pages/digitalauthoridentifierdai.aspx

Virtual International Authority File[18]	OCLC	WorldCat
AuthorID[19]	arXiv	
ORCID[20]	Organisation internationale	Initiative par Thomson Reuters et Nature Publishing Group ; cf. *First Name Identifier Summit* à Londres, en 2009
OpenID[21]	OIDF, fondation internationale	Initiative Google, Facebook etc.
ISNI[22]	ISO TC 46/SC9	Norme internationale
OKKAM[23]	Projet UE FP7	Entités web

Tableau 3 : Identifiants des chercheurs utilisés par différents CRIS

Certains identifiants ont une vocation universelle et globale, d'autres sont liés à un outil, organisme ou contexte particulier. De même, les informations liées aux identifiants sont assez variées : certains identifiants s'appuient sur des informations déclarées par les personnes concernées, d'autres exploitent d'informations administratives ou font un mélange des deux.
Dans le modèle néerlandais par exemple, la personne du chercheur est liée à l'institution et aux projets et résultats de la recherche via son statut professionnel (*employee*) et son contrat (*appointment*) (figure 5 ; Vernooy-Gerritsen, 2009).

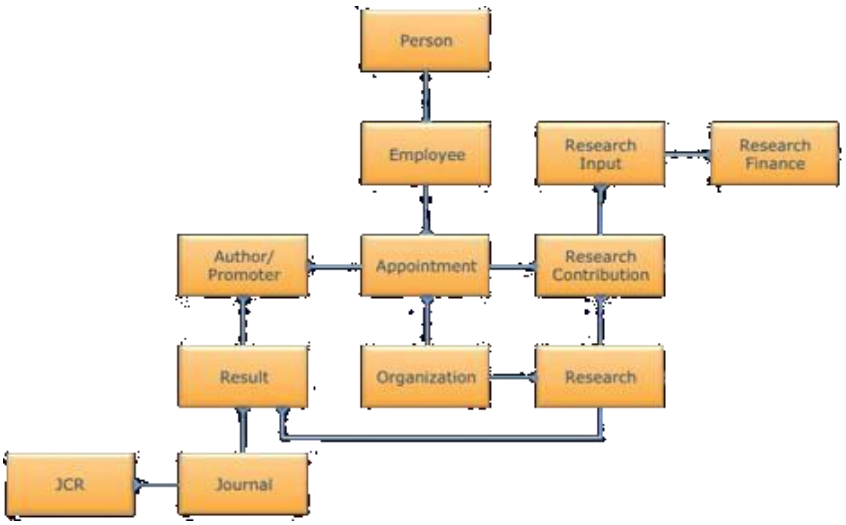

Figure 5 : Informations liées à la personne, dans le modèle néerlandais (© Vernooy-Gerritsen)

Dans un tout autre contexte – création d'un CRIS à l'université de Novi Sad en Serbie – le format CERIF a été adapté de manière à pouvoir faire la distinction entre plusieurs rôles d'auteur qui seront liés à un identifiant unique (PersonID). Un exemple pour l'auteur d'un article de revue dont les catégories plus détaillées correspondraient aux exigences de l'évaluation scientifique en Serbie (Ivanovic et al., 2009) :

- *author of paper published in leading journal of international importance,*

[17] http://www.researcherid.com/
[18] http://viaf.org/
[19] http://arxiv.org/help/author_identifiers
[20] http://about.orcid.org/
[21] http://openid.net/
[22] http://www.isni.org/
[23] http://www.okkam.org/

- *author of paper published in outstanding journal of international importance,*
- *author of paper published in journal of international importance,*
- *author of paper published in leading journal of national importance or*
- *author of paper published in journal of national importance.*

Pour finir, ajoutons donc que le choix d'un identifiant pour chaque chercheur est nécessaire pour le fonctionnement d'un CRIS, mais ne suffit pas pour résoudre tous les problèmes liés à la personne et ses activités.

Perspectives

Les systèmes d'information de recherche (CRIS) prennent et prendront de l'importance sur le plan de l'infrastructure scientifique, aussi bien pour des raisons techniques (intégration des systèmes, échange de données…) que politiques et administratives (rationalisation de la gestion, évaluation de la recherche, justification d'un retour sur investissement scientifique…). Dans ce contexte, les fonctionnalités d'un CRIS sont essentielles. Or, une partie des données des CRIS se trouvent aussi dans les archives institutionnelles. En parallèle, les archives institutionnelles peuvent exploiter plusieurs données cœur des CRIS et développer certaines fonctions de gestion et d'évaluation. Dans l'infrastructure d'un établissement ou organisme, la synergie entre CRIS et archive institutionnelle permet d'améliorer la qualité des deux outils et de réduire la redondance. Cette synergie n'a pas qu'un seul visage, il peut s'agir d'un échange de données, d'une interconnexion, voire d'une intégration ou fusion.

Dans tous les cas de figure, le format est au centre de la question – comme format d'échange entre plusieurs silos de données, comme support d'un référentiel, ou bien comme format commun et partagé de plusieurs applications. Or, le seul format européen suffisamment avancé, normalisé, soutenu par l'administration scientifique européenne, libre, accepté par un nombre croissant d'organismes, d'institutions et d'administrations, que ce soit au plan local, régional ou national, est le format CERIF. Tout comme l'initiative OAI avec une normalisation minimale (OAI-PMH, Dublin Core) a largement contribué à l'essor et à l'interopérabilité des archives ouvertes, CERIF est à ce jour le seul format capable de fournir une solution pour le développement des systèmes d'information de recherche en Europe, dans la perspective d'une infrastructure connectée, interopérable, partagée.

D'une certaine manière, on peut donc dire que ces systèmes présentent l'un des multiples avenirs de l'information scientifique, comme vecteur de communication et comme outil d'analyse et d'évaluation. Plus avancés dans la modélisation et la normalisation des données et procédures, y compris par rapport à l'utilisation des identifiants uniques pour les auteurs-chercheurs, institutions, projets etc., les CRIS peuvent contribuer au développement des archives institutionnelles et d'une façon plus générale, au déploiement du libre accès à l'information scientifique, même si cela n'est pas leur principal objectif.

A une échelle plus large et à moyen terme, deux autres technologies peuvent intervenir et modifier le modèle décrit : d'une part les archives et dépôts des données de recherche (*data sets repositories*) et d'autre part, les réseaux sociaux avec certaines fonctionnalités compatibles avec les archives ouvertes et/ou les CRIS. Par ailleurs, l'avenir dira aussi si les solutions et outils seront sous contrôle public ou pas, quel métier s'en chargera, s'il s'agira d'applications locales, en réseau ou *in the cloud*, de systèmes propriétaires ou libres, au service des communautés scientifiques ou de l'administration etc. Et l'avenir dira aussi comment la question de la protection des données personnelles trouvera une réponse satisfaisante et en conformité avec l'environnement légal.

Bibliographie

M. K. Elbæk, et al. (2010). `CRIS/OAR interoperability workshop'. In *CRIS 2010. 10th International Conference on Current Research Information Systems, June 2-5, 2010, Aalborg, Denmark.*
N. Freire, et al. (2012). `Author Consolidation across European National Bibliographies and Academic Digital Repositories'. In *CRIS 2012. 12th International Conference on Current Research Information Systems, June 6-8, 2012, Prague, Czech Republic*.
D. Ivanović, et al. (2011). `A CERIF data model extension for evaluation and quantitative expression of scientific research results'. *Scientometrics* **86**(1):155-172.
K. G. Jeffery (2011). `CERIF - CRIS Overview'. In *euroCRIS membership meeting, 2-3 November 2011, Lille, France*.
K. G. Jeffery & J. Dvorak (eds.) (2012). *e-Infrastructures for Research and Innovation. Linking Information Systems to Improve Scientific Knowledge Production.*, Prague.
Zeithamlova Milena Ing Agentura Action M.
A. Jippes, et al. (2010). `NARCIS: research information services on a national scale'. In *The 5th International Conference on Open Repositories (OR2010) Madrid, Spain, 6-9 July 2010*.
B. Joerg (2011). `CERIF Tutorial'. In *euroCRIS membership meeting, 2-3 November 2011, Lille, France*.
N. Joint (2008). `Current research information systems, open access repositories and libraries'. *Library Review* **57**(8):570-575.
W. J. Nixon (2010). `Enrich: Improving integration between an Institutional Repository and a CRIS at the University of Glasgow'. In *CRIS 2010. 10th International Conference on Current Research Information Systems. June 2-5, 2010, Aalborg, Denmark*.
E. Poltronieri, et al. (2010). `Science, institutional archives and open access: an overview and a pilot survey on the Italian cancer research institutions'. *Journal of Experimental & Clinical Cancer Research* **29**(1):168+.
S. Rumsey (2010). `A case analysis of registering research activity for institutional benefit'. *International Journal of Information Management* **30**(2):174-179.
M. Vernooy-Gerritsen (ed.) (2009). *Emerging standards for enhanced publications and repository technology : survey on technology*. Amsterdam University Press, Amsterdam.
S. Warner, et al. (2009). `Author Identifiers in Scholarly Repositories'. In *4th International Conference on Open Repositories, Atlanta, GA, 18 May 2009*.

D'autres ressources bibliographiques : http://www.citeulike.org/user/Schopfel/tag/cris

Sites web

Site officiel d'euroCRIS avec un tutoriel sur le format CERIF :
http://www.eurocris.org/
http://www.eurocris.org/Index.php?page=CERIFtutorial&t=1

Sites en lien avec des identifiants uniques :
http://www.isni.org/
http://about.orcid.org/
http://www.researcherid.com
http://viaf.org/
https://repinf.pbworks.com/w/page/13779410/Author%20identification

Tous les sites ont été consultés en août 2012.