Scheme for mapping scientific research data from EPrints to CERIF format

Valentin Penca*, Siniša Nikolić *, Dragan Ivanović *

* University of Novi Sad/Faculty of Technical Sciences/Department of Computing and Automatics, Novi Sad, Serbia
{valentin_penca,sinisa_nikolic, chenejac}@uns.ac.rs

***Abstract*— This paper describes basics of the EPrints institutional repository and CRIS systems and their data models. The result of this research is mapping scheme of the data from EPrints to the CERIF standard.**

I. INTRODUCTION

Rapid development of science and technologies resulted with huge amount of various data. One of the most important tasks will be how to store and make data accessible. Institutional Repository (IR) can resolve the mentioned issue. In [1], an IR is described as an electronic system that captures, preserves and provides access to the digital work products of a community.

The three main objectives for having an institutional repository are:

- To create global visibility and open access for an institution's research output and scholarly materials.
- To collect and archive content in a "logically centralized" manner even for physically distributed repositories.
- To store and preserve other institutional digital assets, including unpublished or otherwise easily lost ("grey") literature (for example, theses or technical reports).

The availability of open-source technologies affect on the rapid development of IRs worldwide, particularly among academic and research institutions. Therefore, it is not surprising the existence of several open source software platforms available for developing IRs like *EPrints* [2], *DSpace* [3], *Greenstone* [4], *Fedora* [5] and *Invenio* [6]. *IR EPrints* was the first open-source repository software to be developed. In [7] is stated that commonly used institutional repository systems are DSpace and Eprints. Paper [8] advocates that EPrints is a powerful and inexpensive solution for sharing scholarly works with the world. Drawback of all mentioned IR is that they have their specific metadata models causing difficulties in data exchange between diverse systems. One possible solution is using the Common European Research Information Format (CERIF) standard [9], which is the basis of Current Research Information Systems (CRISs), for exchanging data from scientific-research domain. In this paper the scheme for mapping data about research from EPrints IR to CERIF format is proposed. That scheme can be used as a guideline, supporting the exchange between *EPrints* repositories and CRIS systems. Motivation for this work was also to extend and improve research from [10].

II. EPRINTS IR

EPrints is an open-source software package for building open access Institutional repositories. Software was developed at the University of Southampton School of Electronics and Computer Science and released under a GPL license. This presents clear advantages for institutions with smaller budgets and that have programmers on staff. EPrints was the first IR software packages to appear and has been available for 14 years. EPrints is a Web and also command-line application based on the LAMP [11] architecture (but is written in Perl rather than PHP). It successfully runs under multiple OS platforms, like Linux, Solaris, Mac OS X and Microsoft Windows. Version 3 of the software introduced a (Perl-based) plug-in architecture for importing and exporting data. Current stable release is 3.3.11/31 from January 2013. EPrints is fairly interoperable [12], supporting OAI-PMH [13] and SWORD [14].Comparison and advantages of EPrints to other IR repositories is described in [15]. According to the official data presented in [16] there are over 535 EPrints registered instances running all over the world which are part of ROARMAP (Registry of Open Access Repositories Mandatory Archiving Policies).

Word "eprints" in particular means “electronic publications” (papers, lectures, videos, etc). Therefore, EPrints is a piece of software for managing "eprints". Basic entity in EPrints is the *Data Object (DataObj)*, which is a record containing metadata [17] and has unique identifier. In EPrints exists three core objects (Figure 1) *EPrint*, *Document*, *User*. All core objects extend *DataObj*. Between Core Objects are defined some relations like: one *User* owns (deposit) many *EPrints*, or one *EPrint* has many documents attached to it. In EPrints one or more documents (files) can be linked with the data object (*Document*). File objects are an interface to file-system files (or cloud buckets) which can be downloaded to allow a person to read a publication (eprint).

A *DataObj* is a collection of "metadata fields". There are many different types of metadata fields. Type affects how a field is rendered, indexed, searched and so forth. Every field has a *type*, *name* property and indicator that states if the field is multiple or not. *Data Object*s are usually stored in the database. They are generally spread over a number of tables containing with the same prefix. Every *Data Object* has system fields (which are set by the system, and not alterable), but the *User* object and *EPrint* object have additional fields which are configured on a per-repository basis. These non system fields can be customized in the Perl script files user_fields.pl and eprint_fields.pl. Detailed information of Metadata Field Types and their configuration can be found in [18].

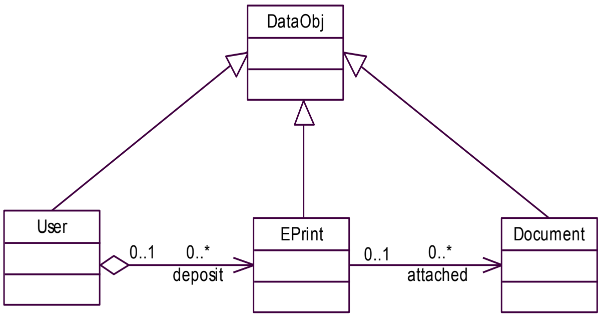

Figure 1 - EPrints data model

We've described the individual *Data Object* but a repository holds many eprints, documents and has many registered users. For that reason a concept of *dataset* for collection of *DataObj* is included. A dataset represents all data objects of a certain type in a single repository. All data objects in the repository are part of the three core datasets "eprint" (all eprints), "user" (all registered users) datasets, "document" (all document). Multiple fields (lists of values) are stored in their own table named after the dataset, then an underscore followed by the fieldname, e.g. "eprint_subjects". These tables also contain a "pos" value to indicate the order of the list.

III. CERIF MODEL

CERIF is the standard which describes data model that can be used as a basis for an exchange of data from scientific-research domain. CERIF Standard describes the physical data model [19] and the exchange of XML messages between the CRIS systems [20]. The best feature of CERIF is that it can be expanded and adapted to different needs. In practice, CERIF is often mapped to other standards that also represent the data of scientific-research domain, for example CERIF/MARC21 mapping described in [21]. Authors of [22] recommend an extension of CERIF that incorporates a set of metadata required for storing theses and dissertations. Another example is [23] where authors argue how CERIF can be used as a basis for storage of bibliometric indicators.

Hereinafter we will present main entities of the CERIF data model version 1.5

- Base Entities - represent the core (basic) model entities. There are only three basic entities cfPerson, cfOrganizationUnit and cfProject.
- Result entities - A group of entities which includes results from scientific research like publications, products and patents. Representatives of this group are: cfResultPublication, cfResultProduct and cfResultPatent.
- Infrastructure Entities - represent a set of infrastructure entities that are relevant for scientific research. The entities which belong in this group are: cfFacility, cfEquipment and cfService.
- 2nd Level Entities - Entities that further describe the Base Entities and Result Entities. E.g. cfEvent is one of those entities, stating the event.
- Link Entities - are used to link entities from different groups. Typical entities of this group are: cfOrganizationUnit_OrganizationUnit, cfOrganizationUnit_ResultPublication and cfResultPublication_DublinCore. Link Entities allow for a generic classification mechanism to define their meaning, indicating the role for each entity instance in a relationship. Every Link entity is described with a role (*cfClass*, *cfClassScheme*), timeframe of relation (*cfStartDate*, *cfEndDate*), value(*cfFraction*) and identifiers of elements creating relation (e.g. *cfOrgUnit*, *cfResPublId*). The 'role' in link entities is not stored directly as attribute value, but as reference to Semantic layer.
- Multiple Language Entities - These entities provide multilingualism in CERIF for some entities.
- Semantic Layer Entities - Provide different kinds of semantics in CERIF model. The entities in this group are cfClassificationScheme and cfClassification. Those entities are used to describe classes and classification schemes for link and other entities. CERIF prescribes a controlled vocabulary to describe some of the classifications.
- Additional Entities - Currently in this group are classified entities that represent DC record.

Figure 2 [Figure 2] shows some of Base, Result, Link and Multiple Language Entities which are relevant for the mapping proposed in this paper.

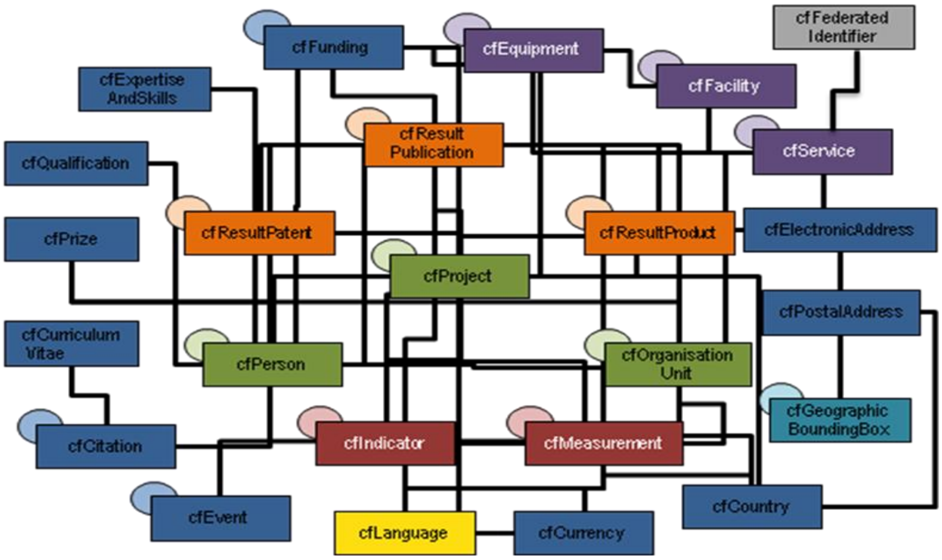

Figure 2 – CERIF model

IV. EPRINTS WITH OTHER SYSTEMS

Storage of data from another system in EPrints was performed in [24]. This paper presents mapping of data from Artext system whose primary users are a national and international community of artists, art historians and curators. In order to obtain the data from Artext, it was necessary to customize default EPrints distribution, adding the new fields and/or modify existing ones. Mappings to EPrints data model was carried out also in [25]. The ReCollect plug-in for EPrints was released at the end of the Research Data @Essex project in 2013. Based on the customization work undertaken while developing a pilot research data repository at the University of Essex, ReCollect offers a low barrier route to deploy such a service, allowing a deposit, review, publication and presentation of research data collections.

The necessity of mapping data between EPrints and systems based on CERIF model is completely natural when we know that these two systems are keeping similar data from scientific research domain [26]. EPrints IR has a large number of users in the UK since it was created at the University of Southampton. In the UK, universities and other research organizations are obligated to prepare themselves for the next national research assessment exercise, known as REF [27]. For that reason a REF2014 EPrints plug-in has been specifically designed to help institutions. REF plug-in [28] is special XML based application profile called CERIF4REF. The mentioned XML is foreseen as a document that is based on CERIF model. EPrints repositories have a potentially important role to play in institutions' responses to the requirements of REF [29].

EPrints contains the plug-ins for the export and import in various formats such as CSV, HTML, DC, ATOM, JSON, MODS, METS. One of the existing plug-ins allows exporting the data in CERIF 2008 1.0 XML format. That CERIF plug-in allows exporting only of some kind of data like books and articles publications to CERIF XML format.

V. MAPPING SCHEME FOR EPRINTS TO CERIF

The motivation of authors for mapping EPrints data to CERIF data model is found in the fact that they are part of the development team of CRIS UNS system [30], which currently does not have the ability to acquire data from the EPrints system. In the paper [31], a CERIF compatible research management system CRIS UNS is presented, which can be accessed at http://www.cris.uns.ac.rs. Currently, the system stores over 10,500 records of scientific publications (papers published in journals, conference proceedings, monographs, technical solutions and patents etc.). CRIS UNS system is under development since 2008 at the University of Novi Sad in the Republic of Serbia. Former development of that system covered implementation of the system for entering metadata about scientific research results [32]. Later phases in the development of CRIS UNS system included integration of various extensions that relay on CERIF model.

Also, the motivation was to include all the *Eprint* data object entities that are not covered with existing CERIF Export plug-in, in accordance with the newer CERIF 1.5 version. Proposed mapping scheme did not include any customization of EPrints distribution since the most users use just the default installation.

Every *EPrints* object has a property called *type* that determines the *EPrints* object type in system (eg. article, book, exhibition, video, etc). The value of that property sets which metadata fields are actually in use from *that* object. So, it represents one of the essential information for mapping. Table 1 defines the mapping to CERIF entities according to the types of *EPrints* object. For easier understanding, all *EPrints* types of objects are grouped and classified into general types of scientific research data (Scientific research data type). For every general type there is an adequate CERIF entity on which it can be mapped. Initial step of mapping is done by resolving the type of *EPrints* object and creating the instance of appropriate CERIF entity.

TABLE I.
EPRINTS OBJECTS CLASSIFICATION

Scientific researh data type	EPrints type	CERIF entity
Publications types	article book_section monograph conference_item book thesis	cfResPubl
Patent types	patent	cfResPat
Event types	exhibition performance	cfEvent
Product types	artefact composition image video audio dataset experiment	cfResProd
Other	teaching_resource other	Non Mapped

All the scientific research data in EPrints is stored as the values of metadata fields. Proposed mapping of all available metadata fields to CERIF format is available at [33]. Due to space limitations, only a segment of mapping is presented in Table 2 and Table 3. For the purpose of explaining the mapping concept, several EPrints metadata fields are selected. Table 2 defines a list of EPrints fields for which the mapping is possible. Also, within that table the EPrints fields' properties are explained. In addition to the field name (*EPrints metadata field*), there is a column *options* where a list of field possible values is noted. For example, field *event_type* states the supported classification of an event as: *conference*, *workshop* and *other*. Fields where the *options* column is empty may have an arbitrary value. Column *multiple* indicates that the field can be found several times in EPrints record (eg. record has more than one author). Column *EPrints type* defines which types of *EPrints* Objects (Table 1, column 2) use the corresponding field (e.g. EPrints field *series* is used only by EPrints type *book*). For some fields that can be applied to more than one *EPrints type* a general group name (Table 1, column 1) or keyword *all* is used. Thus, for the *creator* filed (used in context of for any kind of *EPrints* object) the value of the column *EPrints type* is *all.* The last field in the Table 2 *CERIF entity* defines for which of CERIF entities (single entity or a list) listed in Table 1 the appropriate EPrints field can be utilized. Value *All* indicates that the EPrints field can be used for any CERIF entity that is identified in Table 1 column 3, e.g. each of the specified entities have a *creator*. Value *NONE* is used when it is not possible to perform the mapping for EPrints metadata field to adequate CERIF entity (e.g. *learning_level*).

CERIF model relies on *Link* and *Semantic Layer Entities* to provide additional semantic between entities and for some particular entities. So, it is to be assumed that a large portion of metadata fields from EPrints object will be stored as instances of those CERIF entities. Table 3 presents CERIF entities and their attributes (enclosed in brackets) that are used in mapping. There are three distinct cases of mapping. First, in which the EPrints metadata field is directly stored as an attribute of CERIF entity (e.g. metadata field *series*). The value of that field is stored within attribute *cfSeries* of *cfResPubl* entity. Second case is the situation where the EPrints metadata field can have one of possible values defined in column options. For every possible value, an appropriate CERIF classification is necessary. Every created CERIF entity instance and its classification are connected with the identifier (e.g. *cfResPublId*, *cfPersId* etc). In this scenario column *CERIF core, result and 2nd level entities* specifies instance of CERIF entities for which link entities are defined in column *CERIF link entities. Classification* for those link entities are stated in column *Used CERIF classification*. For example, EPrints metadata field *ispublished* with value *inpress* is classified with CERIF scheme *cfResPubl_Class* and class *In Press*. Entity *cfResPubl* and *cfResPubl_Class* are connected with identifier *cfResPublId*. The most complex is the third scenario, where one value of EPrints metadata field is usually mapped to more than one CERIF entity. This scenario requires the creation of entities from columns *CERIF core, result and 2nd level entities* and *CERIF link entities.* Also, the adequate classification for *CERIF link entities* is defined in column *Used CERIF classification*. Third scenario will be explained with metadata field *creator*. At first, for *creator* an instance of core entity *cfPers* needs to be created. The value of *creator* field will be stored within attributes *cfFamilyNames* and *cfFirstNames* of entity *cfPersName*. *cfPersName* is

TABLE II.
EPRINTS METADATA FIELDS PROPERTIES

Eprints metadata field	options	multiple	EPrints type	CERIF enity
creators		X	all	All
event_type	conference workshop other		Conference item	cfEvent
subjects		X	all	cfResPubl cfResPat cfResProd
learning_level			Teaching resource	NONE
ispublished	pub inpress submitted unpub		publications types	cfResPubl

TABLE III.
EPRINTS METADATA FIELDS MAPPING

Eprints metadata field	options	CERIF core, result and 2nd level entities	CERIF link entities	Used CERIF classification
creators		cfPers (cfPersId) cfPersName_Pers (cfPersId,cfPersNameId) cfPersName (cfPersNameId,cfFamilyNames, cfFirstNames)	Publications types: cfPers_ResPubl (cfPersId,cfResPublId) Patent types: cfPers_Res (cfPersId,cfResPatId) Event types: cfPers_Event (cfPersId,cfEventId) Product types: cfPers_ResProd (cfPersId,cfRedProdId)	scheme:cfPers_ResPubl, class:Author scheme:cfPers_ResPat, class:Patentee scheme:cfPers_Event, class:Performer scheme:cfPers_ResProd, class:Constructor
ispublished	pub inpress submitted unpub	cfResPubl(cfResPublId)	Publications types:cfResPubl_Class (cfResPublId)	scheme:cfResPubl_Class, class:Published scheme:cfResPubl_Class, class:In Press scheme:cfResPubl_Class, class:Submitted for Consideration scheme:cfResPubl_Class, class:Unpublished
series		cfResPubl(cfSeries)		

connected to *cfPers* by link entity *cfPersName_Pers*. The *creator* field is integral part of information about publications, patents, events and products. Thus in CERIF, *cfPers* can be linked with *cfResPubl*, *cfResPat*, *cfEvent* and *cfResProd*. In case when *creator* is an author of publication, linking is done within the entity *cfPers_ResPubl*. For stating the role "author of publication", a CERIF semantic scheme *cfPers_ResPubl* and class *Author* are utilized.

VI. CONCLUSION

The importance of institutional repositories and CRIS systems for scientific research data is enormous. Making data accessible between these systems is unavoidable. Therefore, this paper presents mapping scheme for EPrints data to CERIF model for CRIS systems.

The main contribution of this research is:

- Proposal for mapping data from EPrints repository to the current 1.5 CERIF model
- Potential possibility for creation of new or expansion of existing CERIF-XML Export plug-in

Future work will be directed towards mapping the data from other IRs like Fedora, Greenstone and Invenio to CERIF format.

ACKNOWLEDGMENT

Results presented in this paper are part of the research conducted within the Grant No. III-47003, Ministry of Science and Technological Development of the Republic of Serbia.

REFERENCES

[1] N. F. Foster and S. Gibbons, “Understanding faculty to improve content recruitment for institutional repositories,” -Lib Mag., vol. 11, no. 1, pp. 1–12, 2005.

[2] “EPrints - Digital Repository Software.” [Online]. Available: http://www.eprints.org/. [Accessed: 20-Dec-2014].

[3] “DSpace | DSpace is a turnkey institutional repository application.” [Online]. Available: http://www.dspace.org/. [Accessed: 20-Dec-2014].

[4] “Welcome :: Greenstone Digital Library Software.” [Online]. Available: http://www.greenstone.org/. [Accessed: 20-Dec-2014].

[5] “Fedora Repository | Fedora is a general-purpose, open-source digital object repository system.” [Online]. Available: http://fedora-commons.org/. [Accessed: 20-Dec-2014].

[6] “Invenio.” [Online]. Available: http://invenio-software.org/. [Accessed: 20-Dec-2014].

[7] J. Kim, “Finding documents in a digital institutional repository: DSpace and Eprints,” Proc. Am. Soc. Inf. Sci. Technol., vol. 42, no. 1, 2005.

[8] E. Sponsler and E. F. Van de Velde, “Eprints.org Software: a Review,” Jul. 2001.

[9] “Common European Research Information Format | CERIF,” 2000. [Online]. Available: http://www.eurocris.org/. [Accessed: 18-Jan-2014].

[10] V. Penca and S. Nikolić, “Scheme for mapping Published Research Results from Dspace to Cerif Format,” in 2. International Conference on Information Society Technology and Management, 2012, pp. 170–175.

[11] “ONLamp.com.” [Online]. Available: http://www.onlamp.com/. [Accessed: 20-Dec-2014].

[12] M. Castagné, “Institutional repository software comparison: DSpace, EPrints, Digital Commons, Islandora and Hydra,” Library, Archival and Information Studies (SLAIS), School of, Aug. 2013.

[13] “Open Archives Initiative Protocol for Metadata Harvesting.” [Online]. Available: http://www.openarchives.org/pmh/. [Accessed: 20-Dec-2014].

[14] “SWORD.” [Online]. Available: http://swordapp.org/. [Accessed: 20-Dec-2014].

[15] J.-G. Bankier, Institutional Repository Software Comparison. UNESCO, 2014.

[16] “Registry of Open Access Repositories.” [Online]. Available: http://roar.eprints.org/view/software/eprints.html. [Accessed: 20-Dec-2014].

[17] L. YOGESH and P. NEELIMA, “OPEN SOURCE DIGITAL LIBRARY SOFTWARE (OSS-DL): ASSESSMENT AND EVALUATION,” Int. J. Libr. Sci. Res. IJLSR, vol. 3, no. 3, pp. 21–30.

[18] “EPrints Metadata Fields Documentation.” [Online]. Available: http://wiki.eprints.org/w/Category:EPrints_Metadata_Fields. [Accessed: 20-Dec-2014].

[19] B. Jörg, K. Jeffery, J. Dvorak, N. Houssos, A. Asserson, G. van Grootel, R. Gartner, M. Cox, H. Rasmussen, T. Vestdam, L. Strijbosch, V. Brasse, D. Zendulkova, T. Höllrigl, L. Valkovic, A. Engfer, M. Jägerhorn, M. Mahey, N. Brennan, M. A. Sicilia, I. Ruiz-Rube, D. Baker, K. Evans, A. Price, and M. Zielinski, CERIF 1.3 Full Data Model (FDM) Introduction and Specification. 2012.

[20] J. Dvořák and B. Jörg, “CERIF 1.5 XML - Data Exchange Format Specification,” 2013, p. 16.

[21] D. Ivanović, D. Surla, and Z. Konjović, “CERIF compatible data model based on MARC 21 format,” Electron. Libr., vol. 29, pp. 52–70, 2011.

[22] L. Ivanovic, D. Ivanovic, and D. Surla, “A data model of theses and dissertations compatible with CERIF, Dublin Core and EDT-MS,” Online Inf. Rev., vol. 36, no. 4, pp. 548–567, 2012.

[23] S. Nikolić, V. Penca, and D. Ivanović, “STORING OF BIBLIOMETRIC INDICATORS IN CERIF DATA MODEL,” Kopaonik mountain resort, Republic of Serbia, 2013.

[24] T. Neugebauer, C. MacDonald, and F. Tayler, “Artexte metadata conversion to EPrints: adaptation of digital repository software to visual and media arts documentation,” Int. J. Digit. Libr., vol. 11, no. 4, pp. 263–277, Dec. 2010.

[25] T. ENSOM, “RECOLLECT: TECHNICAL OUTPUTS FROM THE RESEARCH DATA @ESSEX PROJECT,” presented at the JISC MANAGING RESEARCH DATA PROGRAMME WORKSHOP, BIRMINGHAM, 2013.

[26] L. Carr, “EPrints: A Hybrid CRIS/Repository?,” in CERIF and Institutional Repositories, Rome, 2010.

[27] “REF 2014.” [Online]. Available: http://www.ref.ac.uk/. [Accessed: 20-Dec-2014].

[28] “EPrints - Are you ready for REF 2014?” [Online]. Available: http://www.eprints.org/ref2014/. [Accessed: 20-Dec-2014].

[29] A. Clements and V. McCutcheon, “Research Data Meets Research Information Management: Two Case Studies Using (a) Pure CERIF-CRIS and (b) EPrints Repository Platform with CERIF Extensions,” Procedia Comput. Sci., vol. 33, pp. 199–206, 2014.

[30] “Current Research Information System of University of Novi Sad.” [Online]. Available: http://www.cris.uns.ac.rs/. [Accessed: 18-Jan-2014].

[31] D. Surla, D. Ivanovic, and Z. Konjovic, “Development of the software system CRIS UNS,” in Proceedings of the 11th International Symposium on Intelligent Systems and Informatics (SISY), Subotica, 2013, pp. 111–116.

[32] D. Ivanović, G. Milosavljević, B. Milosavljević, and D. Surla, “A CERIF-compatible research management system based on the MARC 21 format,” Program Electron. Libr. Inf. Syst., vol. 44, no. 3, pp. 229–251, 2010.

[33] “Mapping EPrints to CRIF.” [Online]. Available: http://s000.tinyupload.com/?file_id=35285195853574771781. [Accessed: 29-Dec-2014].