



14th International Conference on Current Research Information Systems, CRIS2018

OpenAIRE compatibility for CRIS systems: recent progress

Pablo de Castro^{a,d*}, Jan Dvořák^{b,d}, Ed Simons^{c,d}

^aUniversity of Strathclyde, Glasgow, United Kingdom

^bCharles University Prague, Prague, Czech Republic

^cRadboud University Nijmegen, Nijmegen, The Netherlands

^dVereniging euroCRIS, Nijmegen, The Netherlands

Abstract

This paper reports on the progress that the OpenAIRE-funded, euroCRIS-led METIS2OpenAIRE project means for the implementation of the OpenAIRE Guidelines for CRIS Managers. An update of these guidelines has taken place in the past two years to enable a smoother and more effective information exchange process between CRIS systems and the OpenAIRE aggregation. METIS2OpenAIRE aimed to ensure OpenAIRE compatibility based on these version 1.1 guidelines for a first institutional research information management system, namely METIS at Radboud University Nijmegen in the Netherlands. This three-and-a-half-month project also set out to support a number of parallel implementations by external, budget-neutral stakeholders that would ensure a widespread adoption of the newly released CRIS interoperability standard. These were PURE, the single most widely adopted CRIS platform worldwide, and OMEGA-PSIR, a widespread CRIS solution in Poland.

The work carried out under the METIS2OpenAIRE initiative and its results as of mid-2018 are described in this piece. These results include the development of a minimally sufficient OpenAIRE validator for CRIS systems. This software, specifically designed to support the interoperability requirements for the METIS institutional CRIS at Radboud, should gradually evolve into the default mechanism for other CRIS systems to use in order to test their own level of compliance with the OpenAIRE Guidelines. As for the external stakeholders, by mid-2018 there is already significant progress in the route towards OpenAIRE compatibility for both PURE and OMEGA-PSIR.

© 2019 The Authors. Published by Elsevier B.V.

This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/3.0/>)

Peer-review under responsibility of the scientific committee of the 14th International Conference on Current Research Information Systems, CRIS2018.

* Corresponding author. Tel.: +44(0) 141 548 4666

E-mail address: pablo.de-castro@strath.ac.uk

Keywords: CERIF; research information management; current research information systems; system interoperability; information exchange; OpenAIRE; METIS; PURE; OMEGA-PSIR

1. Introduction: the benefits of CRIS interoperability with OpenAIRE

In the past few years there has been steady progress towards turning Current Research Information Systems (CRIS) into suitable data providers for OpenAIRE sitting alongside Literature and Data Repositories. A first version (1.0) of the OpenAIRE Guidelines for CRIS Managers based on CERIF-XML was officially presented at the April 2014 CRIS2014 Conference in Rome [1] and a subsequent presentation ‘Progress in the Implementation of the OpenAIRE Guidelines for CRIS Managers’ was delivered at the CRIS2016 Conference in St Andrews [2]. Following recent progress in the development of an updated version 1.1 that will result in a more effective and agile information exchange mechanism [3], the stage has been reached where an implementation of such guidelines on a specific CRIS can be attempted.

The reasons behind the update in the OpenAIRE Guidelines for CRIS Managers were summarized in a contribution by Thomas Vestdam, Brian Plauborg and Leen Van Campe to the CRIS2016 Conference in St Andrews [4]. This explored the challenges involved in the research information exchange via CERIF-XML between the Flemish Research Information Space (FRIS) R3 portal and its data providers – the research institutions taking part in the project. A number of shortcomings were identified in the design of the CERIF-XML information exchange process which resulted in a set of recommendations for the standard “to stay relevant in commercial and large-scale contexts”. These recommendations have been addressed in the updated version 1.1 of the OpenAIRE Guidelines.

Enabling system interoperability between CRIS systems and OpenAIRE will mean significant progress towards a more comprehensive collection of research information for the European Research Area in a number of different aspects:

- The gradual addition of CRIS systems to the list of OpenAIRE data providers [5] will mean a *qualitative leap forward in the volume of contextual metadata that gets delivered for research output records*, including research data, publications and other text-based outputs. CRIS records typically contain detailed information on areas like organisational affiliation and research funding that may not be so thoroughly covered by other platforms. This additional input will potentially allow OpenAIRE to step up one gear in the quality of the information offered from the aggregator.
- The interoperability between CRIS systems and OpenAIRE will also mean progress for the purpose of *complying with the European Commission's policies on Open Access and Open Research Data*. While most institutions in Europe currently rely on their institutional repositories for delivering the relevant information on EU-funded project outputs to OpenAIRE, there are numerous cases where this is not technically feasible – mainly due to the lack of system interoperability. This is for instance a frequent case in the UK, very few of whose institutional repositories are listed as OpenAIRE-compliant at the moment. This means that FP7 and H2020 project outputs produced at UK institutions often fail to reach the OpenAIRE aggregation and do subsequently not appear on the publications section of the project pages on the EU CORDIS database.
- National-level CRIS systems are often playing a key role in the wide-scoped aggregation of research information in countries where the repository infrastructure is yet insufficiently consolidated. These national-level platforms not being connected to OpenAIRE due to the lack of interoperability mechanisms, it means that all these outputs are again missing in the OpenAIRE aggregation, with the subsequent negative impact on their visibility and exposure. At a European level, this issue is mainly arising in specific Central and Eastern European countries these days, but recent surveys of the research

information management landscape [6] show that it may well be a frequent situation in other geographies too.

While an increasing alignment on international repository infrastructures is already ensuring a much higher level of interoperability across continents [7], the expansion in the amount of contextual metadata resulting from the integration of CRIS systems into the wider range of OpenAIRE data providers will mean a much more comprehensive approach to a fully interconnected database of all research outputs. Moreover, depending on the source of the research information that gets delivered into the aggregation, it may also be feasible to introduce the concept of authoritative records at a much larger scale than currently, covering for instance funded project information as provided by research funders' systems.

2. The role of the METIS2OpenAIRE project in enabling CRIS interoperability

2.1. The project concept

The euroCRIS-led METIS2OpenAIRE project kicked-off at the beginning of February 2018 after being selected for funding among the proposals received to the late-2017 OpenAIRE call for project proposals [8]. The 3.5-month project set out to make the METIS institutional CRIS at Radboud University the first OpenAIRE-compatible CRIS system. This compatibility was to be based on the updated CERIF-XML Guidelines for CRIS Managers v1.1 also released early in 2018. These Guidelines are to play a similar harmonising role with regard to CRISs to the one the DRIVER/OpenAIRE Guidelines for Repositories [9] have played for a long time in the area of institutional Open Access repositories. In order to operate as a data provider, the METIS records for a number of entities needed to be exposed in the appropriate format via an OAI-PMH endpoint [10]. That feed would subsequently be tested for compliance with the Guidelines via a minimally sufficient validator to be developed by the project.

As an in-house-developed CRIS at Radboud University, the METIS platform development and upgrade is fully controlled by the internal team of IT experts at the institution. The flexibility offered by this arrangement makes these in-house-built systems the most suitable candidates to attempt OpenAIRE compatibility in a first instance. However, because these platforms tend to be case-specific solutions – the METIS platform currently features two implementations in the Netherlands – it is hard to ensure a large uptake of the CERIF-XML Guidelines on this basis. The way METIS2OpenAIRE addressed this low uptake issue at project design stage was to involve external, budget-neutral partners with a wider implementation basis: these were Pure, the Elsevier CRIS with over 200 implementations worldwide, and Omega-PSIR, a Polish research information management solution with more than 20 implementations in the country. Neither of these widely implemented platforms was understandably able to commit to the project's very short development timeschedule, so the project goals just included the provision of technical support for these two large platforms to achieve OpenAIRE compatibility later in 2018. DSpace-CRIS, another widely implemented CRIS solution, is working towards its own OpenAIRE compatibility within a different OpenAIRE-funded project under the same call.

2.2. Technical work

METIS2OpenAIRE was mainly conceived as a technical exercise to validate the METIS feed against the OpenAIRE Guidelines for CRIS systems, with an additional work package on the dissemination of the project progress and outputs. This technical work was structured in three different areas:

- enabling an OAI-PMH endpoint on METIS in order to allow the OAI-PMH CERIF feed to be exposed,
- developing a minimally sufficient validator to test the Guidelines on the METIS feed, and

- performing a loop of tests for such feed against the Guidelines, including the delivery of feedback to the METIS developers on any issues found so that further testing could take place once the required modifications had been implemented.

2.2.1. Exposing METIS records via an OAI-PMH CERIF endpoint

Radboud University already runs a DSpace-based institutional repository that is delivering the information on institutional publications to the OpenAIRE aggregation via the NARCIS repository aggregator in the Netherlands. On top of this, the opportunity to expose records from the institutional CRIS provides an opportunity to add a number of contextual metadata elements that are not covered by the information exchange with the repository. These include interlinked CERIF entities like Persons, Organisational Units, Projects, that make it possible to cover areas like researchers' affiliations or funding information associated to research outputs.

2.2.2. A minimally sufficient validator

The implementation of OpenAIRE compatibility for CRIS systems aims to follow the same rationale that is applied for publication repositories: the metadata feed gets validated against the OpenAIRE Guidelines via an external validator that checks the structural soundness and the completeness of the metadata set. Due to the complexity of the underlying CERIF structure, the metadata feed from CRIS systems requires a specific tool to perform these checks. The METIS2OpenAIRE project set out to build a first approach for such a CRIS validator, specifically aiming to develop a minimally sufficient tool that would carry out the tests on the METIS metadata feed against the updated version of the CERIF-XML Guidelines for CRIS Managers.

This prototype OpenAIRE CRIS validator, available in GitHub, tests the CRIS metadata for all nine sets (i.e. object types) defined in the OpenAIRE Guidelines. For every record the tool checks if the XML metadata is valid with respect to the XML Schema that accompanies the standard specification. It also checks if the OAI identifiers for the objects are well-formed and correspond to the type of the object and its internal identifier, while watching for possible non-unique (i.e. duplicate) identifiers and enforcing a few other basic expectations.

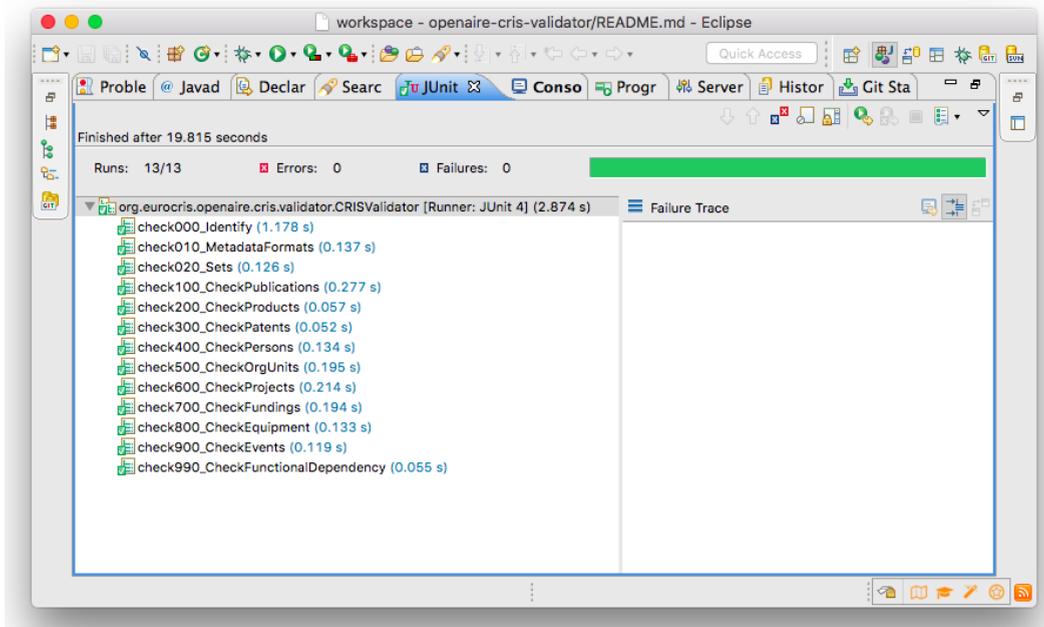


Fig. 1. Application of the validator tests to a sample for the METIS metadata feed

2.2.3. Loop testing of METIS feed

Again due to the complexity of the underlying data model, CRIS systems are bound to be far less homogeneous than repositories in terms of their metadata feeds. The degree of CERIF implementation in these platforms is uneven and some of them cannot be considered to be CERIF-compliant at all. In this sense it was convenient for the first CRIS to be tested against the CERIF-XML Guidelines (i.e. METIS) to be fully CERIF-compliant. This way the validation process has meant an opportunity for a two-way improvement: on the one hand, the OpenAIRE Guidelines themselves have been tested for a first real-life CRIS feed, which has resulted in a degree of fine-tuning. On the other hand, the METIS data model has also undergone a number of adjustments to meet the structural requirements stated in the Guidelines. This exercise has taken place in a number of loops, and should result in both the consolidation of the CERIF-XML Guidelines and especially in a deeper standardisation of the CRIS data models. In the words of the METIS lead Ed Simons:

“the implementation of the Guidelines revealed some shortcomings and gaps in the current data model for the system that will need to be addressed in the near future. It turned out that specific information on 'funding' is missing in the CRIS right now and also some of the information elements on 'projects'. This finding is a direct consequence of the 'old tradition' in which CRIS emerged as reporting systems mainly focused on outputs or research results and not so much as fully-fledged research information management and profiling systems of use for the research managers and researchers themselves. This view is however rapidly changing and a project like METIS2OpenAIRE certainly helps to raise awareness of the changing position and role of CRISs and hence of the need to review and update the "traditional" CRIS data model and functionality”

2.2.4. Supporting external stakeholders to enable a wider implementation for the Guidelines

A fourth and to some extent independent technical work line was the identification of the appropriate mechanisms to support the work towards OpenAIRE compatibility to be carried out by the external, budget-neutral project partners.

The work for these platforms (Pure and Omega-PSIR) mainly involves the mapping of their internal data model to the CERIF-XML structure embedded in the Guidelines. The support provided by euroCRIS ranges from holding regular discussions on specific issues around such a mapping to putting together and organising a working group of institutions to jointly produce the mapping together with the vendor [11]. More information on these collaborations with both external stakeholders is regularly being provided at conferences and membership meetings held by euroCRIS.

3. Next steps towards a generalised CRIS interoperability via CERIF-XML

There is still some way to go until the first CRIS systems are listed as OpenAIRE data providers, but the METIS2OpenAIRE project has meant a significant progress towards that objective. Moreover, with parallel initiatives under way to make DSpace-CRIS and other platforms OpenAIRE-compatible, it should be possible to ensure that the first metadata feeds from specific Research Information Management Systems are soon harvested and ingested into the OpenAIRE aggregation so that towards the end of the year or at the beginning of 2019 there is already a well-established mechanism for such platforms to join the list. Some of the steps that need to be taken along the way are described and analysed below.

3.1. Consolidation of the CERIF-XML Guidelines

The updated version of the CERIF-XML Guidelines for CRIS Managers has only been released at the beginning of 2018 under RFC status. Comments have indeed arisen from the actual attempt to test them against real-life CRIS feeds like METIS or others, and the openly available specification has been evolving as a result of these tests. Once the Guidelines get tested against a sufficient number of different platforms, they will reach a 'temporarily frozen' status [12]. This status will mean that while they will still be open for minor updates, these will not happen on a rolling basis but will be kept as enhancements to be jointly implemented into a new version whenever an enough – or sufficiently significant – number of them has been collected. The existence of such 'frozen' version of the Guidelines is clearly a prerequisite for their widespread implementation.

3.2. The handshake: ingestion of METIS feed into the OpenAIRE aggregation

The METIS2OpenAIRE project has dealt with one side of the handshake that needs to happen in order for the CRIS metadata feed to be part of the OpenAIRE aggregation. By making available a CRIS validator that ensures that the metadata feed meets the requirements enshrined in the CERIF-XML Guidelines, the project has made it possible for the CRIS records – over 300,000 of them in the case of METIS – to be delivered at the front door of the OpenAIRE aggregation. A further step needs to take place in order for such records to actually be ingested into an aggregation whose data model is in fact CERIF-compatible [13]. This has a technical side – the provenance of specific metadata feeds needs to be uniquely identified and the appropriate de-duplication checks need to be in place – and a cultural side: the incoming feed from CRIS systems are bound to be much wider than the average repository feed in a way that may enable the completion of the repository records. Moreover, most CRIS records for publications will typically lack a full-text version of the output, which might in turn be available in a repository elsewhere. It remains to be specified how much information should be displayed while this merging of records takes place, and whether it might be possible to add the workflow for the ingestion of CRIS feeds to other concurrent processes for automatically enriching repository records such as the text-mining of acknowledgement sections in published papers in order to retrieve the funding information from them into the metadata sets. The emergence of multiple, mutually validating parallel workflows for metadata completion could be one of the most valuable outcomes of this initiative to make CRIS systems OpenAIRE-compatible.

In this sense, while the 3.5-month METIS2OpenAIRE project has officially been completed as of mid-May 2018, the project team remains in place and there is ongoing work to make sure the handshake is fully implemented for METIS and eventually for the additional stakeholders involved in the project.

3.3. Dissemination: initiatives for awareness-raising

The original METIS2OpenAIRE project plan contained a third Work Package devoted to the dissemination of the progress achieved “in order to raise awareness of such advances among other potentially interested stakeholders such as CRIS vendors, national-level CRIS platforms and in-house-built institutional CRISs that may wish to start taking their own steps towards OpenAIRE compatibility”.

This has been done while the project work was taking place via a dedicated project webpage in the euroCRIS website where posts were occasionally added by members of the project team [14]. The ongoing work has been mentioned in presentations delivered by team members and there has been a fair amount of dissemination via social media too. However, the real dissemination stage begins once the project has achieved its objectives, because it's only then when a path can be shown that other can follow. The CRIS2018 conference in Umeå meant a first opportunity to discuss this workline with the research information management community at large, as it brought together most of the many stakeholders working in this specific area. Moreover, there are plans for several of these stakeholders to team up in order to carry out joint dissemination actions along the year where the general progress of this distributed initiative towards the integration of CRIS systems into the collection of OpenAIRE data providers is presented and discussed with the wider community.

3.4. Completion of work by the large external stakeholders

As mentioned above, the commitment to OpenAIRE compatibility made by the external platforms involved in the METIS2OpenAIRE project did deliberately not match the 3.5-month project timeline: for an initiative that needs to look deeply into the platform data model it is one thing to work with a fully committed single instance of an in-house-built CRIS, and a very different one to engage a much larger commercial platform that needs to serve its customers' pressing needs in the meantime. Nevertheless, both Pure and Omega-PSIR stated their plans to become OpenAIRE-compatible sometime in 2018 in the letters of intent that were included in the project proposal: by November 2018 in the case of the latter, to coincide with the euroCRIS membership meeting that will be held in Warsaw, and by the release planned for October 2018 in the case of Pure.

It's only when these larger-scale platforms are able to implement their own handshake with OpenAIRE that the mission of the METIS2OpenAIRE project can be considered to be accomplished. In the meantime, while having completed its funded project stage, the initiative is still running within euroCRIS and making sure that the appropriate stakeholders are brought together for the gradual completion of the pending steps. In the meantime, both Pure and Omega-PSIR are progressing with their own implementations and contributing presentations at euroCRIS events.

3.5. Using the euroCRIS DRIS as a basis for a widespread implementation of the Guidelines

Early discussions with OpenAIRE with regard to the procedure for the ingestion of CRIS feeds into the aggregation have highlighted the need for CRIS systems like METIS to feature some kind of unique, persistent identifier that allows its feed to be distinct. This is again a parallel mechanism to the one that OpenAIRE uses for repositories, which is based on the OpenDOAR identifier. One of the first steps for any new Open Access repository to join the worldwide community is to register with the OpenDOAR directory maintained by the SHERPA/Jisc team in Nottingham. Because of the neatly different nature of the research information management community, there is however no equivalent directory for CRIS systems that could be used for the purpose. The closest asset we have to such a database

is the euroCRIS Directory of Research Information Systems (DRIS) [15], but this is far from comprehensive at the moment. Nevertheless, the possibility to revamp the DRIS is presently under consideration at euroCRIS so that a simplified and much more comprehensive version of the current DRIS could become the default registry for CRIS systems that supported the progress towards OpenAIRE compatibility. Not all systems registered in the updated DRIS would need to be OpenAIRE-compatible, but every OpenAIRE-compatible CRIS system would need to be registered in the DRIS in order to get the persistent, unique identifier that would allow their metadata feed to be ingested into the OpenAIRE aggregation.

Given that the worldwide survey on research information management practices recently carried out by OCLC and euroCRIS has received over 350 responses from 44 different countries, the circumstances might seem auspicious for a significant step forward towards a better mapping of the available CRIS system infrastructure.

Acknowledgements

The authors acknowledge the support received from the OpenAIRE2020 project funded under the Horizon 2020 framework programme (grant agreement no 643410) in the funding of the euroCRIS-led METIS2OpenAIRE project.

References

1. Houssos N et al. OpenAIRE Guidelines for CRIS Managers: Supporting Interoperability of Open Research Information through Established Standards. *Procedia Comput Sci* 2014;**33**:33-38, DOI 10.1016/j.procs.2014.06.006
2. De Castro P et al. Progress in the Implementation of the OpenAIRE Guidelines for CRIS Managers. *Procedia Comput Sci* 2017;**106**:104-111, DOI 10.1016/j.procs.2017.03.042
3. Dvořák J (2017), "OpenAIRE Guidelines for CRIS Managers 1.1". Presentation delivered at the euroCRIS Strategic Membership Meeting Autumn 2017 (CVTI SR, Bratislava, Slovakia, Nov 20-22, 2017), <http://hdl.handle.net/11366/604>
4. Vestdam T, Plauborg B, Van Campe L. FRIS R3 - CERIF XML in large scale exchange of research information. *Procedia Comput Sci* 2017;**106**:74-81, DOI 10.1016/j.procs.2017.03.037
5. OpenAIRE data providers: literature repositories, journals, data repositories and CRIS systems, <https://www.openaire.eu/search/data-providers> (accessed 15 July 2018)
6. Bryant R et al (2018), "Preliminary Findings from the euroCRIS/OCLC Research Survey of Research Information Management Practices" (In press)
7. Confederation of Open Access Repositories (COAR, 2016), "Landmark agreement between OpenAIRE and LA Referencia on repository aligning", <http://bit.ly/2HfqDMB>
8. "OpenAIRE tender calls results!", <https://www.openaire.eu/openaire-tender-calls-results>
9. Feijen M et al. DRIVER: Building the Network for Accessing Digital Repositories across Europe. *Ariadne* 2007:53, <http://www.ariadne.ac.uk/issue53/feijen-et-al>
10. CRIS information elements relevant to OpenAIRE, http://guidelines.readthedocs.io/en/latest/cris/cris_elements_openaire.html
11. Clements A et al. Implementation of the OpenAIRE CERIF-XML guidelines in Pure: a marriage made in heaven. Contribution to CRIS2018: 14th International Conference on Current Research Information Systems (Umeå, June 13-16, 2018). <http://hdl.handle.net/11366/667>
12. Dvořák J et al (2018), "OpenAIRE Guidelines for CRIS Managers 1.1". Zenodo, <http://dx.doi.org/10.5281/zenodo.1298650>
13. CRIS information elements relevant to OpenAIRE, http://guidelines.readthedocs.io/en/latest/cris/cris_elements_openaire.html
14. METIS2OpenAIRE project webpage at euroCRIS site, <https://www.eurocris.org/projects/metis2openaire>
15. euroCRIS Directory of Research Information Systems (DRIS), <https://dspacecris.eurocris.org/cris/explore/drisc>