Summary of European Publication Information Infrastructure Data Model

For interoperability reasons, the European Publication Information Infrastructure Data Model relies on the CERIF data model which allows representation of various research entities, their connection and outputs with their semantic relationships. For this first iteration of the data model, a certain subset of attributes from the CERIF data model, more specifically the ones related to Publication element, were chosen.

How this subset differs from the OpenAIRE's Guidelines for CRIS Managers is that the control over which attributes should be included in the publication metadata is greatly expanded. This is indicated by the list of attributes under "Mandatory" title. To some extent this core set of attributes was already discussed on the ENRESSH-VIRTA-POC, but is now also corresponding to the CERIF data model and its attributes. This is made evident by providing each attribute with a equivalent CERIF representation as well as what is mentioned on the OpenAIRE's Guidelines for CRIS Managers.

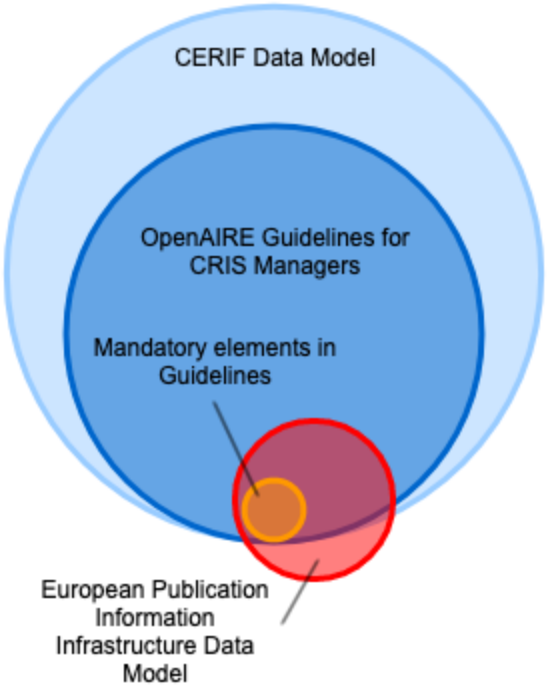

Figure 1: European Publication Information Infrastructure Data Model in relation to CERIF data model and OpenAIRE Guidelines for CRIS Managers

A set of these mandatory elements should be provided with every record that is to be aggregated to European Publication Information Infrastructure. For each attribute, a justification i.e. why this attribute is relevant for the interoperability and quality of metadata, is provided.

In addition, a disclaimer i.e. what kind of context related issues there might be for certain attribute, is discussed and analyzed. These are to support the validation phase and also add up to the bottom-up discussion on source systems on how to make metadata comparable and unified in relation to e.g. classifications of publications.

Also, a set of attributes is stated under "Conditional" title with less detailed information. These attributes would be required based on the record's publication type. For example a book chapter should be accompanied by ISBN number of the book and the source title of this book. Otherwise, these attributes would be optional.

Last, some attributes are listed under "Optional" title. These would not be required by the validation process for any records, but would be strongly recommended in metadata exchanges to support the quality of metadata and help with e.g. de-duplication processes of the infrastructure.

Proposed European Publication Information Infrastructure Data Model's attributes

Mandatory

```
Publication
 Internal identifier
 Publication type
 Publication title
 Publication date
 Author
  Organizational author and affiliation
 Discipline
```

Conditional

ISSN*
ISBN*
Source title*
Peer review*

Mandatory based on publication type*

Optional

Audience
DOI
Volume
Number
Start page
End page

Mandatory attributes

Publication	CERIF representation	OpenAIRE Guidelines for CRIS Managers	European Publication Information Infrastructure
One publication equals one record, to which other information is related to.	XML element `Publication`	Other attributes are children to this element, as such a mandatory (1)	Other attributes are children to this element, as such a mandatory (1)

Justification: As it's the base element for all the other elements it should be mandatory to provide.

Disclaimer: High variation on source systems in terms of publication inclusion criteria, e.g. are non-scholarly publications included as well (professional and general/popular books, articles, reports etc.) or are conference presentations or short abstracts included. To some extent this can be traced by the `type`-element and what kind of types are included. Scope of what is actually recorded as a publication could be provided by for example provenance information in `about` element or otherwise controlled/validated in metadata exchange. Limitations to inclusion criteria are not recommended though, as comparability and analysis of records are more dependent on e.g. publication types to be consistent.

Internal identifier	CERIF representation	OpenAIRE Guidelines for CRIS Managers	European Publication Information Infrastructure
The organization's own ID for the publication.	XML attribute `id`	mandatory (1)	mandatory (1)

Justification: For technical purposes e.g. harvesting and updating the record, a unique id is needed for the metadata to be traceable.

Disclaimer: None

Publication type	CERIF representation	OpenAIRE Guidelines for CRIS Managers	European Publication Information Infrastructure
Publication type according to the publication type classification.	XML element `Type` from namespace https://www.openaire.eu/cerif-profile/vocab/COAR_Publication_Types	mandatory (1)	mandatory (1)

Justification: Metadata should be provided on the publication type to have at least some structure and information on the publications. This allows the data to be comparable and used as a basis for analysis and research.

Disclaimer: Source systems almost never have the exact same semantics / classifications for publication types e.g. what is included in a review article type might differ based on the source system definition. Mapping publication types from data model to another always leads to loss of information. Three different approaches can be had on the mapping (1) try to achieve the most accurate possible mapping between data models (2) use the so called lowest denominator of publication types; a limited set of broad categories (3) use lowest denominator publication types, but include refinements where possible and use the broad categories externally.

Publication title	CERIF representation	OpenAIRE Guidelines for CRIS Managers	European Publication Information Infrastructure
Publication title as given in the article or the book. If necessary, the title of a foreign-language publication may be transliterated.	XML element `Title` as a multilingual string	optional, possibly multiple (0..*)	at least one mandatory, possibly multiple (1..*)

Justification: In terms of proper identification of the publication, a publication title should be provided. Although in theory, this could be enriched from elsewhere, but for accurate information it should be provided from the source systems to get title in the original publishing language. Data exchange can be done using multilingual strings to provide information on alternative languages as well.

Disclaimer: As publication title is provided in a free form text string, there is no guarantee that the title would match either the original publication title or title provided in other databases for the same publication.

Publication year	CERIF representation	OpenAIRE Guidelines for CRIS Managers	European Publication Information Infrastructure
The year in which the publication was published for the first time as a version with full bibliographic information.	XML element `PublicationDate`	optional (0..1)	mandatory (1)

Justification: For publishing date a year of the publication is needed for comparison and analyzing purposes. This allows the data to be e.g. analyzed between time periods.

Disclaimer: Publication year is not always definitive, but it should be based on the time when publication was published as a version with full bibliographic information available.

Authors	CERIF representation	OpenAIRE Guidelines for CRIS Managers	European Publication Information Infrastructure
Authors of the publication that were listed in the original publication or source database.	XML element `Authors` with ordered embedded XML elements `Author` XML element `Author` with embedded XML element `Person`	optional (0..1) optional, possibly multiple (0..*)	mandatory (1) at least one mandatory, possibly multiple (1..*)

Justification: Author information is highly relevant to be able to both identify and analyze the bibliographic metadata for research purposes. Without author data, there would be no ties to the persons and a research dimension, that are the people behind the publications, would be lost. To collect this information on the source systems, the metadata quality should be high as the different

Disclaimer: As the format and spelling of names might vary heavily between source systems, the use of identifiers e.g. ORCID is highly recommended. But as many researchers are still missing the unique identifiers or the systems are not supported by it, the author information should still be provided in text string format as well.

Organization authors / affiliation	CERIF representation	OpenAIRE Guidelines for CRIS Managers	European Publication Information Infrastructure
Authors affiliated in the reporting organisation.	XML element `Person` optionally followed by one or several `Affiliation` elements, or `OrgUnit`. A `DisplayName` may be specified, too.	optional, possibly multiple (0..*)	mandatory, possibly multiple (1..*)

Justification: Organizational author information is relevant as it ties the publication via affiliation information to a certain organization. Thus, an analysis or research can be made using the affiliation information. It also provides better legitimacy for the publication as the affiliated organization is known. The information about organizational units should be made as close to the source system as possible and it should be made mandatory for that reason.

Disclaimer: At this time there is no reliable identifier in place for organizations. By providing information in both human and machine readable formats, there would still be no up to date information on e.g. organizational units as they tend to change as time passes and organization structures change.

Discipline / Field of science	CERIF representation	OpenAIRE Guidelines for CRIS Managers	European Publication Information Infrastructure
The first, so-called primary field of science is mandatory .	XML element `Subject` containing the classification identifier and having a `scheme` attribute to specify the classification scheme identifier	optional, possibly multiple (0..*)	mandatory, possibly multiple (1..*)

Justification: For analysis and research purposes, a discipline should be provided for the publication. Although many schemas are available, for data comparability all publications should use standard OECD's revised Frascati Manual classification as mandatory schema for providing fields of science (FoS). Multiple entries for FoS can be assigned for each publication.

Disclaimer: The practices on how to classify discipline for publications vary heavily and many of the source systems use different variations of FoS classification or even use a local classification for disciplines. Discipline can in addition be defined by 1) publication itself, 2) journal of the publication, 3) author of the publication or 4) organizational unit where the author comes from.

Conditional attributes

ISSN	CERIF representation	OpenAIRE Guidelines for CRIS Managers	European Publication Information Infrastructure

The ISSN number of the series publishing the journal, monograph or parent publication according to the primary printed version. If there is no printed version, the ISSN number of the electronic version will be indicated.	XML element `ISSN`	optional, possibly multiple (0..*)	mandatory (1) if publication type of following: `conference paper, journal article, review article, research article, editorial, data paper` otherwise optional, possibly multiple (0..*)

ISBN	CERIF representation	OpenAIRE Guidelines for CRIS Managers	European Publication Information Infrastructure
Publication or parent publication ISBN number.	XML element `ISBN`	optional, possibly multiple (0..*)	mandatory (1) if publication type of following: `bibliography, book, book part, report` otherwise optional, possibly multiple (0..*)

Source title	CERIF representation	OpenAIRE Guidelines for CRIS Managers	European Publication Information Infrastructure
The source (another Publication) where this publication appeared. E.g. a journal article lists here the journal where it appeared. To be used for a publishing channel.	XML element `PublishedIn` with embedded XML element `Publication`	optional (0..1)	mandatory (1) if publication type of following: `book part, editorial` otherwise optional (0..1)

Peer review	CERIF representation	OpenAIRE Guidelines for CRIS Managers	European Publication Information Infrastructure
Whether the publication is peer reviewed.	XML element using tags	Not present	mandatory (1) if publication type of following `conference paper, journal article, review article, research article, editorial, data paper` otherwise optional (0..1) Could also be enriched via journal/book information

Optional attributes

Audience	CERIF representation	OpenAIRE Guidelines for CRIS Managers	European Publication Information Infrastructure
Based on the target audience of the publication.	XML element using tags	Not present	optional (0..2) Classification: 0=Scientific 1=Professional 2=General/public Could also be enriched via journal/book information

DOI	CERIF representation	OpenAIRE Guidelines for CRIS Managers	European Publication Information Infrastructure
The Digital Object Identifier (DOI) of the publication.	XML element `DOI`	optional, possibly multiple (0..*)	optional, possibly multiple (0..*)

Volume	CERIF representation	OpenAIRE Guidelines for CRIS Managers	European Publication Information Infrastructure
Volume of the journal or series in which the article appeared.	XML element `Volume`	optional (0..1)	optional (0..1)

Number	CERIF representation	OpenAIRE Guidelines for CRIS Managers	European Publication Information Infrastructure
Issue of the journal or series in which the article appeared.	XML element `Number`	optional (0..1)	optional (0..1)

Start page	CERIF representation	OpenAIRE Guidelines for CRIS Managers	European Publication Information Infrastructure
Publication's start page number in which the article was published in the same format as in the original article or source database.	XML element `StartPage`	optional (0..1)	optional (0..1)

End page	CERIF representation	OpenAIRE Guidelines for CRIS Managers	European Publication Information Infrastructure
Publication's end page number in which the article was published in the same format as in the original article or source database.	XML element `EndPage`	optional (0..1)	optional (0..1)

Supporting documents

VIRTA European pilot - the data contents

https://openaire-guidelines-for-cris-managers.readthedocs.io/en/latest/cerif_xml_publication_entity.html