Supplementary information

Example of reduction of a relation (citation/edge) Our dump has the benefit of focusing solely on citations, and thus we do not need to indicate the citation type. We only retain the source and target IDs. These IDs are also optimized to take up as little space as possible. The following is an example of how an edge is represented in the OpenAIRE JSON dump:


```
{
  "provenance": {
    "provenance": "Inferred by OpenAIRE",
    "trust": "0.9"
  },
  "relType": {
    "name": "Cites",
    "type": "citation"
  },
  "source": "doi_________::7e8d84fc096936557defb78d22cca97c",
  "sourceType": "product",
  "target": "dedup_wf_002::27d83ddfd6e54378d88445aa793d5cb8",
  "targetType": "product",
  "validated": false
}
```


An edge in our dump is simply:


```
159486578   118392581
```


Estimates of OpenAIRE JSON files uncompressed Table 4 shows a rough estimate of the full uncompressed size of the OpenAIRE dump. The dump would require several orders of magnitude more RAM than most consumer computers.

Table 4: Estimate of OpenAIRE JSON files when uncompressed

Relation (edges) Size Estimate	
One relation JSON (uncompressed)	6.5 GB
One relation JSON (compressed)	0.5 GB
Compression ratio	$6.5/0.5 = 13$
Compressed relation folder size	10 GB
Estimated uncompressed per folder	$13 \times 10 = 130$ GB
Number of relation folders	14
Total uncompressed (relations)	$14 \times 130 =$ **1.82 TB**
Publication (nodes) Size Estimate	
One publication JSON (uncompressed)	0.5 GB
One publication JSON (compressed)	0.1 GB
Compression ratio	$0.5/0.1 = 5$
Compressed publication folder size	10 GB
Estimated uncompressed per folder	$5 \times 10 = 50$ GB
Number of publication folders	14
Total uncompressed (publications)	$14 \times 50 =$ **0.70 TB**
Combined Uncompressed Estimate	**2.5 TB**

Filtered fields We remove a lot of fields from the original OpenAIRE graph dump. Table 5 is an overview of the fields we have filtered out of the publications (nodes). Table 6 is an overview of the fields filtered out of the relations (edges).

Field	Description
Bibliographic	
`subTitle`	Alternative or explanatory title
`publisher`	Publishing entity
`version`	Version of the result
`sources`	Dublin Core dc:source
`contributors`	Contributing persons or entities
`subjects`	Keywords with scheme and provenance
`formats`	File formats
`coverages`	Coverage information
`originalIds`	Identifiers at original sources
`dateOfCollection`	When OpenAIRE last collected the record
`lastUpdateTimeStamp`	Timestamp of last update in OpenAIRE
`embargoEndDate`	Date embargo ends
Open Access	
`bestAccessRight`	Openest access right (code, label, scheme)
`isGreen`	True if green Open Access
`isInDiamondJournal`	True if published in a Diamond Journal
`openAccessColor`	gold, hybrid, or bronze
`publiclyFunded`	True if outcome of a funded project
Impact indicators	
`indicators.citationImpact.citationClass`	Citation class label
`indicators.citationImpact.impulse`	Impulse score
`indicators.citationImpact.impulseClass`	Impulse class label
`indicators.citationImpact.influence`	Influence score
`indicators.citationImpact.influenceClass`	Influence class label
`indicators.citationImpact.popularity`	Popularity score
`indicators.citationImpact.popularityClass`	Popularity class label
`indicators.usageCounts.downloads`	Download count
`indicators.usageCounts.views`	View count
Instances (per-version metadata)	
`instances`	Per-version access rights, URLs, license, refereed status
Geolocation	
`geoLocations`	Box, place, and point geolocation data
`countries`	Associated countries with ISO codes
Software/dataset specific	
`codeRepositoryUrl`	URL to source code repository
`contactGroups`	Groups responsible for the software
`contactPeople`	Persons responsible for the software
`documentationUrls`	URLs to software documentation
`programmingLanguage`	Programming language of software
`size`	Declared size of dataset
`tools`	Tools for interpretation of result
Sub-fields of included objects (partially used)	
`authors.name`	Author first name
`authors.surname`	Author surname
`authors.rank`	Author position
`authors.pid`	Author ORCID identifier
`container.edition`	Journal or proceeding edition
`container.ep`	End page
`container.iss`	Journal issue number
`container.issnLinking`	Linking ISSN
`container.issnOnline`	Online ISSN
`container.issnPrinted`	Printed ISSN
`container.sp`	Start page
`container.vol`	Volume
`container.conferenceDate`	Conference date
`container.conferencePlace`	Conference location
`language.label`	Language label in English

Table 5: Fields present in the OpenAIRE product schema that are not included in the output dataset.

Field	Description
Core relation fields	
`sourceType`	Type of the source entity (e.g. product, project, organization)
`targetType`	Type of the target entity
`relClass`	High-level category of the relation (e.g. citation, affiliation)
`relType`	Specific relation type describing the semantic meaning
Provenance and validation	
`provenance`	Information about how the relation was generated or provided
`validated`	Indicates whether the relation has been validated
`validationDate`	Date when the relation was validated
Additional metadata	
`confidence`	Confidence score for the relation
`inference`	Indicates whether the relation was inferred or explicitly provided

Table 6: Fields present in the OpenAIRE relation schema that are not included in the output dataset. The core relation field "relType" is used for filtering to only obtain "Cites" relations. And removal of dangling edges ensures that the output dataset only includes edges with sourceType and targetType "product".