Aggregation workflow

Cécile Devarenne
Operations Officer
Metadata training, Europeana Sounds project
Athens, 23rd/24th of October 2014
Content

- Europeana's aggregation team
- Europeana Publication Policy
- Aggregation workflow
- Submission deadlines
- Ingestion processes and tools
- Acceptance criteria and Europeana validation of data
- Guidance and help – Europeana pro
- Future plans for aggregation workflow
Europeana’s aggregation team: who are we?
Europeana’s aggregation team

• Partner relationships, business development, administration
  • Henning Scholz, Joris Pekel, Gina Van der Linden
• Technical support
  • Operations officers: content@europeana.eu
  • Data support, feedback and ingestion of your collections into Europeana portal and API
Europeana Publication Policy
Europeana Publication Policy

Clear criteria for acceptance or decline of metadata for publication and for take down of legacy metadata from the Europeana database

• Ingestion workflow (deadlines, timelines, prioritization)
• Content scope (what is a digital object? what content does Europeana aggregate?)
• Technical validation of metadata quality (expected values)
• Metadata licensing (CC0)
• Rights Statements for digital objects
  • All digital objects with valid edm:rights chosen from http://pro.europeana.eu/web/guest/available-rights-statements
  • Public Domain material labelled with the Public Domain mark in edm:rights
  • edm:rights & dc:rights not in contradiction
Aggregation workflow and submission deadlines: how does it work?
Submission of data: preliminary steps for your project

• (1) Data Exchange Agreement to Europeana (DEA)
  • Europeana Sounds project needs to submit the signed Data Exchange Agreements for each contributing data provider
  • The Europeana Data Exchange Agreement establishes the terms under which Europeana can make use of the previews and descriptive metadata provided by cultural institution
  • More information to be found here: http://pro.europeana.eu/ensuring-permissions-for-aggregators

• (2) Data contribution form
  • One form for the whole project
  • General information on data to be submitted to Europeana
  • Schedule of data delivery: ingestion planning

• (3) Submission of data samples and feedback taken into account
Submission of data: (4) publication cycles

- Operations officers work on a monthly cycle
- Submission of data in the form of datasets: a coherent batch of records, for the Europeana Sounds project, probably one dataset for each of your data providers
- A dataset takes on average 40 mins to process
- Around 200 datasets are processed by the Operations officers for each cycle of publication
- Datasets go through a full flow of operations before they are production ready
- Datasets need to be submitted on time in order for this production cycle to work
- Datasets are submitted by the technical/content coordinators of your project
- The earlier you submit datasets the more feedback we can give!
Submission of data: new provider timeline
Submission of data: regular ingestion cycle timeline

1st of month 1
- Delivers samples of metadata or full datasets for feedback
  - Sends feedback about:
    - structure of the metadata
    - mandatory elements
    - rights statements

Around the 21st of month 1
- Delivers full datasets to be included in the coming publication
  - Processes the datasets or sends feedback about:
    - structure of the metadata
    - mandatory elements
    - rights statements

Around the 1st of month 2
- If needed, delivers a corrected version of datasets ready to be ingested
  - Ingests the datasets to be included in the coming publication, provided that they are compliant to the publication policy

Between the 10th and the 20th of month 2
- Publication is finalized
  - Processes datasets can be retrieved on the Europeana portal
  - Communicates that the publication is live and sends final report for the processed datasets

Submission deadline: no new dataset will be accepted for month 2 publication after the deadline

Monthly ingestion complete
Ingestion processes and tools: what happens to your data when submitted to Europeana?
Europeana’s set of ingestion tools

- Unified Ingestion Manager (UIM): orchestrator of data flows triggered in various tools and plugins
- SugarCRM (Customer Relationship Management): reference entries for datasets and organisations
- REPOX: harvester to get the collections uploaded into Europeana
- Europeana’s instance of Mint (Metadata INTeroperability): mapping and editing tool for ingested datasets
- Data plugins
  - Itemization, Europeana identifiers generation
  - Dereferencing
  - Enrichment
  - Redirects
  - Extraction of hierarchies
  - Thumbnails caching
Europeana ingestion data flows

1. In browser, check raw xml
2. In SugarCRM populate with info about organisations and datasets
3. In UIM, load dataset
4. In REPOX, harvest dataset
5. Via UIM load dataset/s in MINT
6. In MINT map and edit dataset/s
7. Via MINT load transformed dataset/s in UIM
8. In UIM Save hierarchies (if needed)
9. In UIM Create Record Redirects (if needed)
10. In UIM Enrich dataset (to enrich, save and index dataset)
11. In UIM Dereference dataset (if needed)
12. Dataset/s published in Europeana.eu
Steps to get data ingested

From the moment your data was submitted:

- Checks on raw xml (Browser)
  - Prior to harvesting
  - Identification of key issues
- Creation/update of dataset information, checks on validity of the supplied harvesting information (SugarCRM - REPOX)
- Harvesting (REPOX)
- Mapping/editing of datasets (Europeana Mint)
  - Mapping tool for all datasets
  - Adapted for Europeana in order to process multiple formats (EDM, ESE, any metadata standard with provided XSLT)
  - Drag and drop appropriate elements
  - Quality checks and data cleaning if necessary
  - Transformation and validation of records according to EDM schema and schematron rules
- EDM Internal data: Europeana ready material
Steps to get data ingested

• Operations on data following transformation:
  • Itemization and creation/management of Europeana identifiers for permalinks to your records in Europeana
  • Extraction of hierarchies for datasets including EDM hierarchies
  • Thumbnails caching
• Enrichments of data:
  • From links to linked data exposed ontologies, generation of additional contextual data (dereferencing)
  • From analysis of the provided data, automated semantic enrichment (Europeana enrichment)
• If necessary (when a change of identifiers was communicated to Europeana), creation of redirections between previous and newly generated identifiers
• Data ready! monthly deploy on Europeana portal and API
Acceptance criteria: how exactly is the Publication Policy implemented?
Acceptance criteria

• Data Exchange Agreement to Europeana
• Datasets submitted via OAI-PMH protocol, FTP or file
• Metadata are accepted for publication after the feedback of the Europeana Operations Officers
  • EDM schema and guidelines
  • Rights labeling
• Datasets are prioritized for publication if the edm:rights in the majority of the metadata of the dataset is PDM, CC0, CC BY or CC BY-SA
Europeana validation

Automatic validation:

- Validation according to the EDM schema
- Validation of the mandatory properties
- Unique identifiers within a dataset

- Metadata records that don’t meet this validation are invalidated or discarded
- Providers can fix issues first and resubmit or let Europeana ingest the records that are valid, and fix the invalid records at a later stage

- Validation of urls for thumbnail creation (ImageMagick)
## Mandatory properties

<table>
<thead>
<tr>
<th>Applicable class</th>
<th>Mandatory Properties (or alternatives)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aggregation</td>
<td>edm:dataProvider</td>
</tr>
<tr>
<td>Aggregation</td>
<td>edm:isShownAt or edm:isShownBy</td>
</tr>
<tr>
<td>Aggregation</td>
<td>edm:provider</td>
</tr>
<tr>
<td>Aggregation</td>
<td>edm:rights</td>
</tr>
<tr>
<td>Aggregation</td>
<td>edm:aggregatedCHO</td>
</tr>
<tr>
<td>Aggregation</td>
<td>edm:ugc (when applicable)</td>
</tr>
<tr>
<td>ProvidedCHO</td>
<td>dc:title or dc:description</td>
</tr>
<tr>
<td>ProvidedCHO</td>
<td>dc:language for text objects</td>
</tr>
<tr>
<td>ProvidedCHO</td>
<td>dc:subject or dc:type or dc:coverage or dcterms:spatial</td>
</tr>
<tr>
<td>ProvidedCHO</td>
<td>edm:type</td>
</tr>
</tbody>
</table>
Europeana validation

Validation by the Operations officers:

- Feedback is according to the EDM schema and guidelines
- Checks on the connections between the EDM classes and the general structure of the data
- Correct use of vocabularies, recommendations to include geolocations
- Checks on the types of values: literals vs resources (e.g. a thumbnail always need to be a valid url)
- Checks on links to digital representations of the objects; if direct links to a file, check that they are of reasonable size
- Provision of thumbnails highly recommended
- Feedback on (near) duplicate records
- Feedback on rights statements in edm:rights and dc:rights
- Feedback on any other metadata quality related matters (duplication of properties, encoding in the data, wrongly mapped properties, etc.)
Happy ingestion :-)

- The data is represented according to expectations for both sides
- Users can search and retrieve rich content:
  - Developers can make the best use of the API
  - Objects are clicked through and re-used from the Europeana portal
Die Zauberflöte / Auswahl

Description:
Ach ich fühls, es ist entschwunden: aus "Zauberflöte" / (Mozart), Die Zauberflöte / Auswahl

Creator:
Mozart, Wolfgang Amadeus; Greet, Alice de

Contributor:
Mozart, Wolfgang Amadeus; Greet, Alice de

Coverage:
Dresden (Dresden); http://en.wikipedia.org/wiki/Dresden: Dresden (Dre?den, Dresde, Dre?no, Stadt Dresden, State Capital Dresden, Dr??d'any, K?nigliche Haupt- und Residenzstadt Dresden, Drezdzany, Haupt- und Residenzstadt Dresden, Landeshauptstadt Dresden); Same as:

- (DE-588b)37172-5
- (DE-588b)5556921-3
- (DE-588)37172-5
- (DE-588c)4012995-0
- 4012995-0
- http://sws.geonames.org/2935022

Geographic coverage:
Dresden (Dre?den, Dre?no, Stadt Dresden, State Capital Dresden, Dr??d'any, K?nigliche Haupt- und Residenzstadt Dresden, Drezdzany, Haupt- und Residenzstadt Dresden, Landeshauptstadt Dresden); Same as:

- (DE-588b)37172-5
- (DE-588b)5556921-3
- (DE-588)37172-5
- (DE-588c)4012995-0
- 4012995-0
- http://sws.geonames.org/2935022

Latitude: 51.05089; Longitude: 13.73832

Subject:
Schellackplatte; Oper

Identifier:
http://mediathek.sltub-dresden.de/ton70913451.html

Language:
Schellackplatte Oper

Data provider:
Sächsische Landesbibliothek - Staats- und Universitätsbibliothek Dresden

Provider:
Deutsche Digitale Bibliothek

Providing country:
Germany
Guidance and help
Guidance and help

Europeana Professional:
http://pro.europeana.eu/provide-data

Content inbox – for all ingestion & metadata related matters
content@europeana.eu
Questions?
Future plans for aggregation workflow
Future plans for aggregation workflow

- Future plans to open up part of the Europeana ingestion workflow to providers
  - Providers can log-in the Europeana ingestion suite, identify the aggregator/project they work for
  - Providers can select the datasets they want to update, or add new datasets
  - Providers can upload their data (OAI-PMH and FTP protocols)
  - Providers can map their data to EDM, or edit data that is already EDM
  - Providers can validate the data against the EDM schema and preview them prior to submission
- Other processes being considered for refactoring: semantic validation, link checking, thumbnail caching, enrichment
Future plans for aggregation workflow

• Benefits for providers:
  • Possibility to map to EDM
  • Validation according to the EDM schema (with schematron rules we implemented)
  • Preview before publication
  • Self service, less dependent on Europeana, saving time (you can do many steps yourself, and you spot errors earlier)

• Benefits for Europeana:
  • Operations scaled up – amount of projects, aggregators and therefore datasets has grown exponentially in the last years
  • More focus on EDM modeling and metadata related questions
  • Ingestion process transparent and more connected to the process at aggregators side
Thank you!

Cécile Devarenne

cecile.devarenne@europeana.eu or content@europeana.eu